

Cluster Analysis of Multimodel Ensemble Data from SAMEX

AHMAD ALHAMED AND S. LAKSHMIVARAHAN

School of Computer Science, University of Oklahoma, Norman, Oklahoma

DAVID J. STENSRUD

NOAA/National Severe Storms Laboratory, Norman, Oklahoma

(Manuscript received 2 October 2000, in final form 24 July 2001)

ABSTRACT

Short-range ensemble forecasts from the Storm and Mesoscale Ensemble Experiment (SAMEX) are examined to explore the importance of model diversity in short-range ensemble forecasting systems. Two basic techniques from multivariate data analysis are used: cluster analysis and principal component analysis. This 25-member ensemble is constructed of 36-h forecasts from four different numerical weather prediction models, including the Eta Model, the Regional Spectral Model (RSM), the Advanced Regional Prediction System (ARPS), and the Pennsylvania State University–National Center for Atmospheric Research fifth-generation Mesoscale Model (MM5). The Eta Model and RSM forecasts are initialized using the breeding of growing modes approach, the ARPS model forecasts are initialized using a scaled lagged average forecasting approach, and the MM5 forecasts are initialized using a random coherent structures approach. The MM5 forecasts also include different model physical parameterization schemes, allowing us to examine the role of intramodel physics differences in the ensemble forecasting process.

Cluster analyses of the 3-h accumulated precipitation, mean sea level pressure, convective available potential energy, 500-hPa geopotential height, and 250-hPa wind speed forecasts started at 0000 UTC 29 May 1998 indicate that the forecasts cluster largely by model, with few intermodel clusters found. This clustering occurs within the first few hours of the forecast and persists throughout the entire forecast period, even though the perturbed initial conditions from some of the models are very similar. This result further highlights the important role played by model physics in determining the resulting forecasts and the need for model diversity in short-range ensemble forecasting systems.

1. Introduction

Research in ensemble forecasting has largely focused upon finding the most appropriate methods for generating the ensemble member initial conditions to be used within a single modeling system (Mullen and Baumhefner 1988; Toth and Kalnay 1993, 1997; Buizza and Palmer 1995; Houtekamer and Derome 1995; Molteni et al. 1996). However, the role of model physics variability in ensembles is now also being recognized as important for both medium-range (Buizza et al. 1999; Harrison et al. 1999; Evans et al. 2000) and short-range forecasting (Stensrud et al. 2000). Further emphasizing this point are the results from several studies of ensemble systems that suggest model diversity, not just model physics variability, is an important component of ensemble forecasting systems (Atger 1999; Stensrud et al. 2000; Ziehmann 2000; Evans et al. 2000; Fritsch et al.

2000; Hou et al. 2001). In particular, these results show that it is difficult at present for a single model configuration with perturbed initial conditions to provide sufficient spread to capture the atmospheric variability. Yet the reasons for the improvement of ensemble forecasts when a multimodel ensemble is used remain largely unclear, as does the relative importance of model and initial condition uncertainty.

To improve our understanding of the importance of model diversity to short-range ensemble forecasting, data from the Storm and Mesoscale Ensemble Experiment (SAMEX) are analyzed using clustering methodologies. Alhamed and Lakshmivarahan (2000) demonstrated the utility of these approaches for analyzing short-term ensemble forecast output of unimodel (Eta Model) data. Here the intent is to extend this analysis to multimodel, short-range ensemble data and explore the roles of model initial condition, model diversity, and model physics uncertainty in ensemble systems.

SAMEX is a multi-institutional numerical weather prediction experiment, coordinated by the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma, that occurred during May of 1998.

Corresponding author address: S. Lakshmivarahan, School of Computer Science, University of Oklahoma, 200 Felgar St., Rm. 114, Norman, OK 73019-0631.
E-mail: varahan@ou.edu

SAMEX represents a collaboration among CAPS, the National Severe Storms Laboratory (NSSL), and the National Centers for Environmental Prediction (NCEP). One goal that SAMEX hoped to achieve is to apply short-term ensemble forecasting techniques and related statistical verification strategies to multimodel ensemble forecasts. More information on SAMEX can be found in Hou et al. (2001).

Our intent in this study is to apply several cluster analysis (CA) and principal component analysis (PCA) methods to datasets containing various meteorological fields over an entire forecast time period. These techniques allow us to objectively compare the forecast fields from the different models that participated in SAMEX throughout the forecast time window, and to explore their similarities. The SAMEX data provides an excellent opportunity to 1) further explore and compare both initial condition and model variability in the creation of a short-range ensemble forecasts, 2) examine how model diversity affects the distribution of solutions within the ensemble, and 3) compare model diversity and model physics diversity within ensembles. Using CA and PCA methods, the trajectories from the model forecasts that constitute the ensemble are analyzed to learn how the members of the ensemble cluster as they evolve with time. Subjective clustering is also used to evaluate the ability of the algorithms to produce clusters that make sense meteorologically.

To make this paper self-contained we review the clustering algorithms and the basic idea of the principal component analysis in section 2. A description of the properties of the SAMEX dataset used in the analysis is discussed in section 3. Results from several different CA and PCA methods on the precipitation dataset are examined in section 4, followed by results from several other forecast fields in section 5. A final discussion is found in section 6.

2. Clustering algorithms

In order to understand better the results presented, a short overview of the clustering methodology used is warranted. Let $\mathbf{X} = [x_{ij}]$, $1 \leq i \leq m$, $1 \leq j \leq n$, be the *data matrix*. Each column represents an *object* (e.g., ensemble member), and each row represents an *observation*. Thus, x_{ij} denotes the i th observation on the j th object. The aim of clustering is to discover similarities among objects (Romesburg 1984). In the rest of this paper, each object can be viewed as a vector of m dimension, where m is the number of observations, and each row can also be viewed as a vector of n dimension, where n is the number of objects. Hence, x_{*j} refers to the j th object, and x_{i*} refers to the i th observation across the n objects.

It is desired in some cases to normalize the data matrix before applying the clustering algorithms. Several methods exist for normalizing a data matrix (Romesburg 1984). Each method transforms the data matrix \mathbf{X} into

new data matrix \mathbf{Z} of the same size. The new matrix, \mathbf{Z} , called the normalized data matrix, can then be used in data analysis.

Let

$$\bar{x}_{*j} = \frac{1}{m} \sum_{i=1}^m x_{ij} \quad \text{and} \quad (2.1)$$

$$\sigma_{*j} = \left[\frac{1}{m-1} \sum_{i=1}^m (x_{ij} - \bar{x}_{*j})^2 \right]^{1/2}. \quad (2.2)$$

Clearly, \bar{x}_{*j} and σ_{*j} are the unbiased estimates of the sample mean and the sample standard deviation for the j th object. The simplest type of normalization translates and scales the objects so that they have zero mean and unit variance:

$$z_{ij} = \frac{x_{ij} - \bar{x}_{*j}}{\sigma_{*j}} \quad \text{for } 1 \leq i \leq m. \quad (2.3)$$

Further details are found in Romesburg (1984).

Cluster analysis is a multivariate statistical technique that is used to discover the property of a collection of objects to group together based on certain similarity measures. There are two classes of methods for performing cluster analysis: *hierarchical* and *partitional* methods. A hierarchical method operates on the similarity or dissimilarity matrix to construct a tree depicting specified similarity relationships among the objects (Anderberg 1973). It organizes the objects into nested sequence of clusters. Partitional, or nonhierarchical, methods generate a partition of the objects in an attempt to classify the objects. In these latter methods, the number of clusters is specified in advance.

In order to discover a natural clustering of a set of objects, the first step in a clustering algorithm is to compute the similarity or dissimilarity between objects. These pairwise similarity–dissimilarity measures are arranged in the form of a square, $n \times n$, matrix where the (ij) th entry denotes the similarities coefficients between objects i and j , and n is the number of objects. Since such a matrix is symmetric by definition, we only need to consider only one-half of the matrix, say the lower triangular part. Three different similarity–dissimilarity measures are used in this study.

- 1) *Euclidean distance*: The most commonly used dissimilarity measure is the Euclidean distance, e_{jk} , which measures the distance between two objects, say j and k . It can be defined as follows:

$$e_{jk} = \sqrt{\sum_{i=1}^n (x_{ij} - x_{ik})^2}, \quad (2.4)$$

where

$$0 \leq e_{jk} \leq \infty.$$

The data are normalized using (2.3) prior to the Euclidean distance calculations.

- 2) *Correlation coefficient*: The most commonly used similarity measure is the correlation coefficient, r_{jk} . The correlation coefficients, r_{jk} , are the elements of the $n \times n$ correlation matrix \mathbf{R} , which can be defined as follows:

$$\mathbf{R} = \frac{1}{m-1} \mathbf{Z}^T \mathbf{Z}, \quad (2.5)$$

where \mathbf{Z} is defined using (2.3), such that the input data have zero mean and unit variance.

- 3) *Euclidean similarity*: Another similarity measure, called Euclidean similarity, can be derived from the Euclidean distance coefficient (Elmore and Richman 2001). Let $\mathbf{E} = [e_{ij}]$, $1 \leq i, j \leq n$ be the Euclidean distance matrix, and define the Euclidean similarity matrix $\bar{\mathbf{E}} = [\bar{e}_{ij}]$, $1 \leq i, j \leq n$, where

$$\bar{e}_{ij} = 1 - \frac{e_{ij}}{\max_{1 \leq i, j \leq n} (e_{ij})}, \quad 0 \leq \bar{e}_{ij} \leq 1.0$$

for $i, j = 1, 2, \dots, n$. (2.6)

This measure allows for dissimilarity-based PCA analyses.

Since the use of different measures can produce different clusters (Alhamed and Lakshmivaran 2000), it is important that all three of these measures be used to evaluate the model forecasts. If the three measures produce similar clusters, then the conclusions drawn from this analysis are much stronger. Further information on the clustering techniques is found in the appendix.

3. Data

During May 1998, SAMEX personnel collected and stored forecast output from four different numerical weather prediction models at 3-hourly intervals out to 36 h. The models are the NCEP Eta Model (Black 1994), the NCEP Regional Spectral Model (RSM; Juang and Kanamitsu 1994), the CAPS Advanced Regional Prediction System (ARPS; Xue et al. 2000), and the Pennsylvania State University–National Center for Atmospheric Research fifth-generation Mesoscale Model (MM5; Dudhia 1993; Grell et al. 1994) run at NSSL. Forecasts were started each day at 0000 UTC. Each model used a different domain and had different horizontal and vertical resolutions, but the model output fields were interpolated to a common grid that represents the largest shared region from all the models. Note that the MM5 forecasts used a two-way nesting procedure and the MM5 output used is only from the inner domain, such that the domains from all the models extended well beyond the contiguous 48 states. The final shared domain consists of 117×81 grid points (a total of 9477 points) covering a large part of the United States at 30-km grid spacing (Fig. 1).

Three of these ensembles, the Eta, RSM, and ARPS,

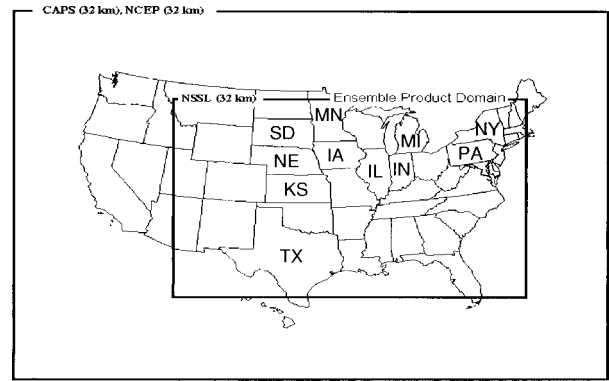


FIG. 1. Domain of the SAMEX model forecast data. The states of Iowa (IA), Illinois (IL), Indiana (IN), Kansas (KS), Michigan (MI), Minnesota (MN), Nebraska (NE), New York (NY), Pennsylvania (PA), South Dakota (SD), and Texas (TX) are mentioned in the text and indicated in the map.

used only variations in initial and boundary conditions to create the ensemble members. The Eta and RSM were started with a control and four bred mode perturbations from the global forecast system (Toth and Kalnay 1993, 1997). The Eta and RSM bred perturbations start with the same initial and boundary conditions, but the RSM also allowed for regional breeding at higher resolution than the global breeding. The ARPS was started with a control initial condition from the Eta Model and four initial and boundary conditions using the scaled lagged average forecasting approach (Hou et al. 2001). This approach is an extension of the lagged average forecasting method of Hoffman and Kalnay (1983).

The MM5 forecasts also used the Eta Model forecasts for the control initial and boundary conditions. Random coherent structures were added to nine of the initial and boundary conditions to mimic analysis uncertainty (Errico and Baumhefner 1987; Stensrud et al. 2000). This approach has slower error growth on the boundaries when compared with the other techniques, but still produces growing modes (Stensrud et al. 2000). In addition, 10 different configurations of MM5 were used, 1 for each forecast. Different physical parameterization schemes for the planetary boundary layer and deep convection were used, and the values of moisture availability were altered. The three convective schemes used were the Betts–Miller (Betts and Miller 1986), Kain–Fritsch (Kain and Fritsch 1990), and Grell (Grell 1993) schemes, and the two planetary boundary layer schemes used were the Blackadar nonlocal closure scheme (Zhang and Anthes 1982) and the Burk–Thompson 1.5-order closure scheme (Burk and Thompson 1989). Moisture availability was calculated using an antecedent precipitation index from daily rainfall totals over the previous three months and interpolated to the MM5 grid. Values of moisture availability were varied by $\pm 10\%$ while maintaining the mean value over the domain as in Stensrud et al. (2000).

The dataset for each forecast field examined in this

TABLE 1. List of institutions and models for the SAMEX ensemble members.

Institution	Model	Member no.
CAPS	ARPS 4.3.5	1–5
NCEP	Eta	6, 8, 10, 12, 14
NCEP	RSM	7, 9, 11, 13, 15
NSSL	MM5 v2	16–25

study represents the output of four forecast models starting at 0000 UTC 29 May 1998. Therefore, the ensemble consists of the 25 members as listed in Table 1. Output fields that are examined include mean sea level pressure, 500-hPa geopotential height, 3-hourly accumulated precipitation, 250-hPa wind speed, and convective available potential energy (CAPE).

While the data analysis is not focused on verification of the ensemble data, as in Hou et al. (2001), a brief overview of the evolution of the atmosphere on this day is helpful. At 0000 UTC 29 May a low pressure center was located in northeastern Canada with a northeast-southwest-oriented quasi-stationary frontal boundary draped across the northern plains states. Deep convection was present to the north of this frontal boundary in southern Minnesota (locations of states shown in Fig. 1) and was also present over a majority of the southeastern states. The convection over Minnesota organized into a derecho, a long-lived damaging wind storm, that moved quickly eastward across the northern states, reaching western Michigan by 0500 UTC 29 May and New York by 1400 UTC the same day. The number of damaging wind reports exceeded 170, the number of severe hail reports exceeded 100, and 16 severe weather watches were issued during the 36-h period over which the model forecasts are valid. Deep convection persisted throughout most of the southeastern states throughout this entire time period, while the convection in the northern United States shifted eastward with time. At the 36-h time, lee cyclogenesis was occurring with a low center in western South Dakota. Thus, even though no distinct midlatitude cyclone passed through the model domain, the weather events on this day were numerous and challenging to forecast.

4. Data analysis

We begin the analysis with the 3-hourly precipitation forecasts, because these are arguably the most important forecast field for short-range forecasting. As discussed in Alhamed and Lakshmivarahan (2000), there are several approaches and measures that can be used to analyze these data into clusters. However, these different approaches can lead to different clustering trees, or dendrograms (Wilks 1995). Therefore, two different hierarchical cluster analysis algorithms are used and the results compared. The unweighted pair-group method arithmetic average (UPGMA) is used with two different similarity measures and Ward's method is used with the

one dissimilarity measure. The similarity measures chosen are the correlation coefficient [(2.5)] and Euclidean similarity [(2.6)], and the dissimilarity measure chosen is Euclidean distance [(2.4)]. Also, two nonhierarchical cluster analysis algorithms, *k*-mean and nucleated agglomerative (NA), are used. Clustering using the rotated principal component analysis is done based on both similarity measures.

It is important to examine the results from several methods and measures, since the selection of these methods and measures can influence the results (Alhamed and Lakshmivarahan 2000). The only way to gain confidence that the results are independent of the techniques used is to show consistency across many techniques. Further details on the clustering methodology are found in the appendix.

For each forecast field 13 different data matrices are constructed, 1 for each output time, with a CA and PCA conducted separately for each output time. The data matrix, $\mathbf{X}_{m \times n}$, $m = 9477$, $n = 25$, represents the quantities of the field at each grid point from the 25 ensemble members. Thus, each member is represented by a column with 9477 elements with 1 element per grid point. Recall that the objective of these analyses is to study the behavior of a multimodel ensemble and see how members from different models cluster. Results from the similarity measures are discussed first, followed by the results from the dissimilarity measure.

a. Precipitation: CA and PCA based on correlation

The correlation matrix (2.5), $\mathbf{R}_{25 \times 25}$, is computed for each time instance and then used as a similarity matrix for the CA method in order to determine the hierarchical clustering structures of the 25 members of the ensemble. The clustering dendrograms resulting from applying the UPGMA method on the 3-hourly accumulated precipitation data indicate that the members are diverging rapidly with time (Fig. 2). Also note that the members from one model tend to build their own cluster before merging with members from other models. If the clustering trees at 24, 27, and 30 h are cut to form four clusters, then the resultant clusters represent the four different models. At the other output times, when members from different models join each other to form a cluster, this merge occurs at a very low level of correlation, indicating the lack of similarity between the model fields. This tendency points out the coherence between the members from one model and the isolation between the members from different models.

To investigate the meteorological significance of these results, plots of the 3-hourly accumulated rainfall valid at 30 h are shown (Fig. 3). A quick examination of these data indicates that the rainfall patterns cluster by model first. Even though the Eta Model and RSM forecasts are initialized using bred modes, the precipitation forecasts are amazingly similar in location of the precipitation regions from one forecast to another. For

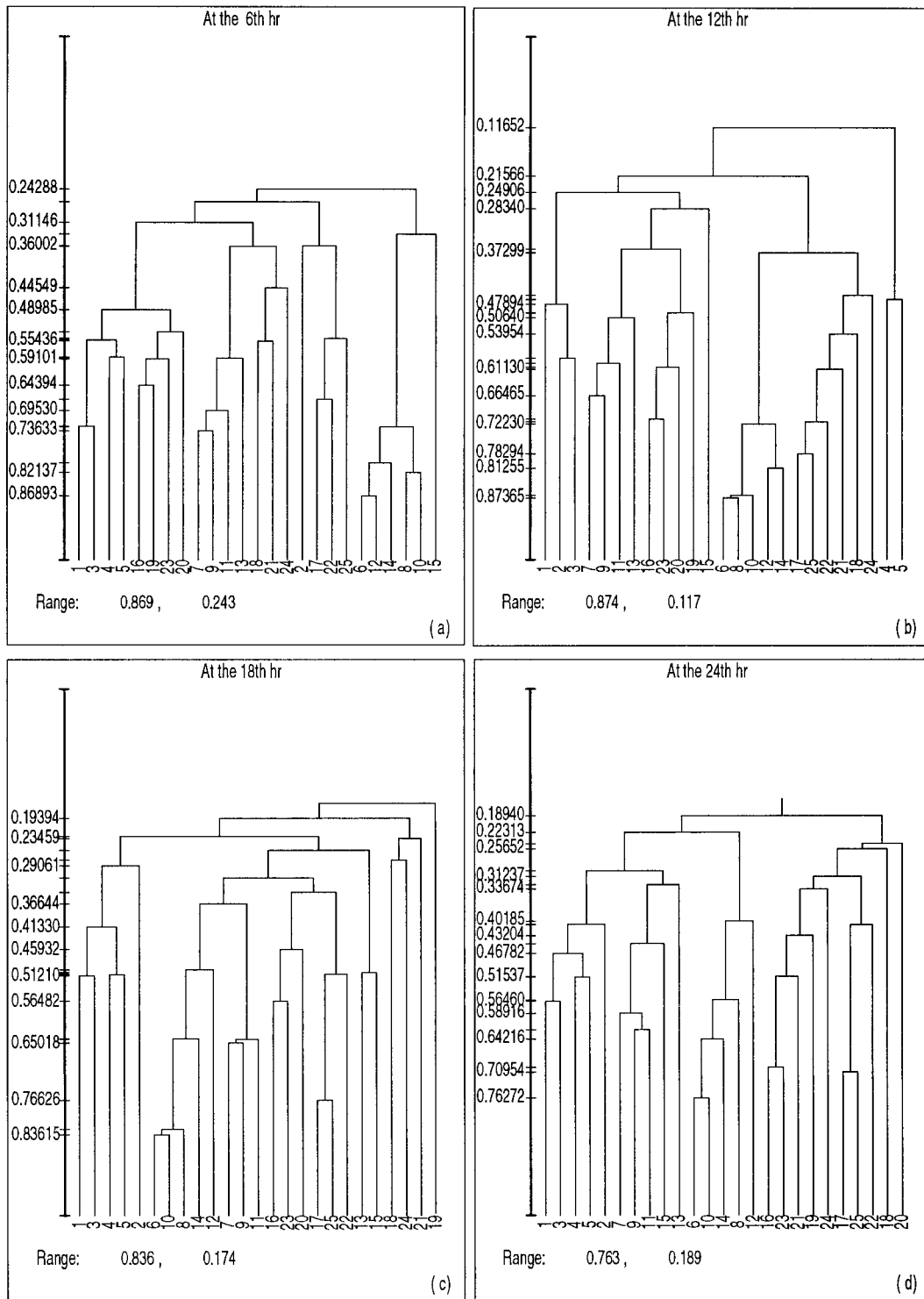


FIG. 2. UPGMA dendrograms for the accumulated precipitation based on the correlation measure for forecast times of 6, 12, 18, 24, 30, and 36 h. Dendrograms, or “tree” diagrams, illustrate the order in which the model forecasts cluster. The two forecasts that are most alike cluster first on the lowest branch of the diagram. The higher the merging point, the lower the level of similarity between the two forecast clusters.

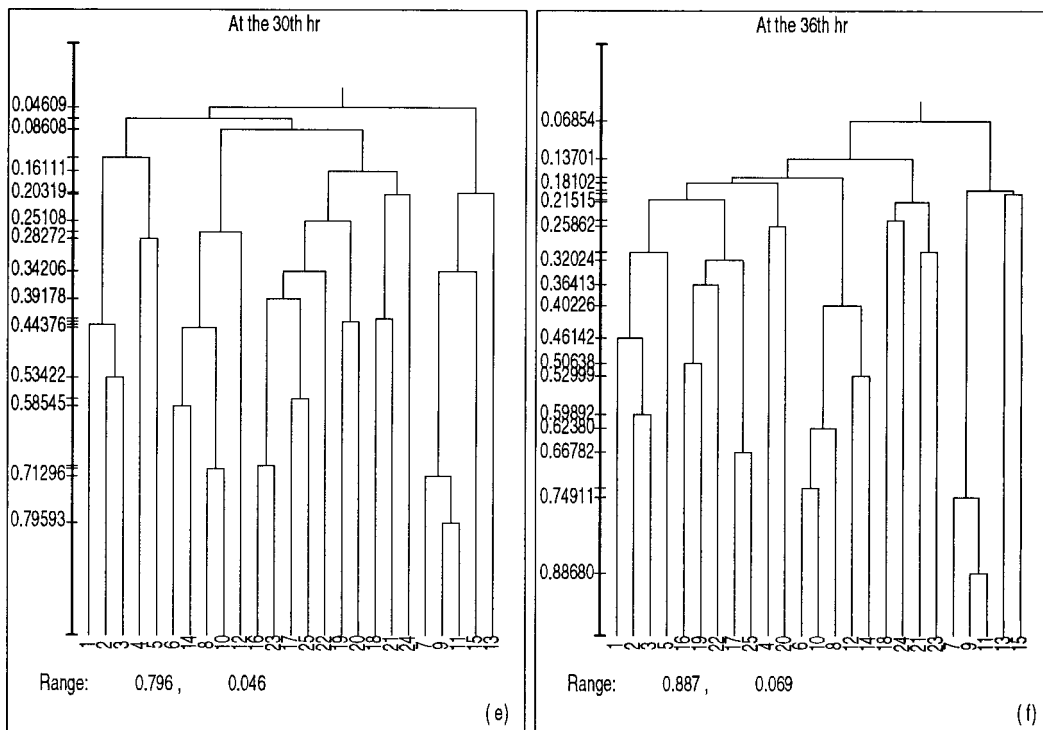


FIG. 2. (Continued)

example, all the Eta Model forecasts have a local precipitation region in western Illinois and along the Gulf coast. Similarly, all the RSM forecasts have a local precipitation region along the Minnesota–South Dakota border with varying amounts of precipitation along a bowing line from Iowa to New York. Looking at the other forecasts, note that all the ARPS forecasts produce a precipitation region in Kansas and produce much less precipitation over the Gulf of Mexico when compared with the Eta Model and RSM forecasts. Finally, all the MM5 forecasts produce precipitation from southern Indiana eastward into Pennsylvania, but have a greater variety of precipitation patterns and amounts elsewhere in the domain. Examination of other time periods (not shown) produces similar conclusions. The subjective analyses are largely in agreement with the CA results in that the forecasts cluster first by model, indicating a strong coherence in forecasts from the same numerical weather prediction model.

This result suggests that the model framework and physical parameterization schemes dominate the evolution of the precipitation forecast. Recall that both the Eta Model and the RSM use the breeding of growing modes technique, based upon the global model output, to produce the different initial and boundary conditions for their five-member ensembles. Yet these models rarely cluster together first, and when they do cluster together the level of correlation is low (Fig. 2). These results agree with those of Stensrud et al. (2000) who show that model physics is a very important component

of short-range ensemble systems and can dominate over initial condition uncertainty when the large-scale forcing for upward motion is weak. However, the large-scale forcing for this case is not weak, as a relatively strong short wave was passing through the upper Midwest. Yet model physics still appears to play a very important role in the evolution of the model forecasts of precipitation.

To further explore this ensemble behavior, the eigenvalues of the correlation matrices for the 12 output times also are computed. The variance of the full 25-member ensemble explained by the first eigenvalues of the correlation matrices, which is the fraction of the total variance contained in the first mode of the analysis, indicates that the data are not concentrated on one dimension (Fig. 4). Hence, the difference between the variance explained by the first eigenvalue at 3 h, 39.2%, and at 36 h, 20.9%, points out the slight divergence of the members of the ensemble as the time passes. However, the behavior of the individual models is not the same, as is seen using the multimodel dataset. When the correlation matrices for the individual models are constructed and their eigenvalues are computed, a large spread in the variance explained by the first eigenvalue is found from the four models (Fig. 4). The precipitation forecasts from the Eta Model are most alike (i.e., have the largest variance explained by the first eigenvalue), and the forecasts from MM5 are most different, with the RSM and ARPS in between. Note that three of the models have percentages of variance explained above 50% throughout most of the forecast time period. This

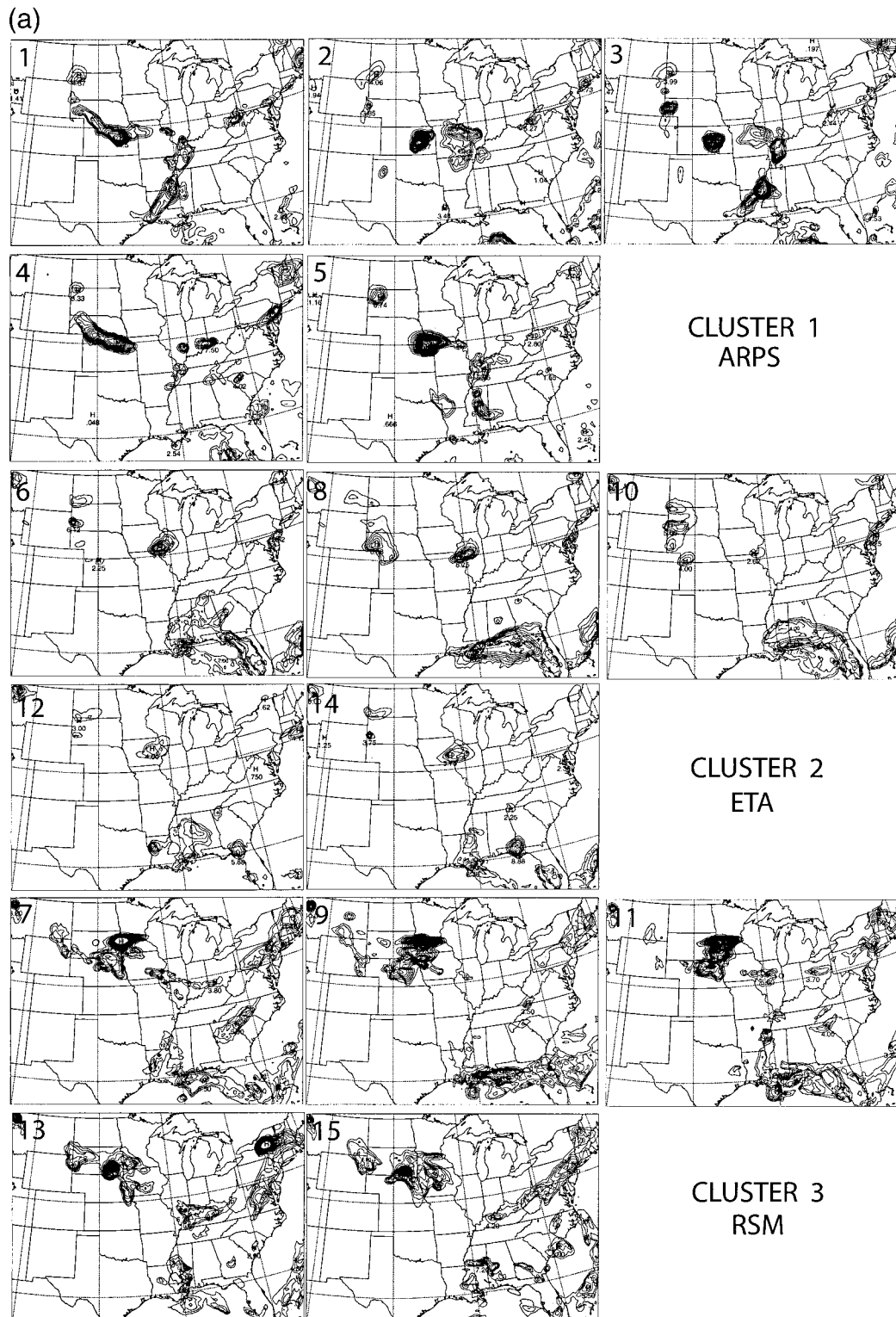


FIG. 3. The 3-h accumulated precipitation (mm) valid at 30 h from all 25 ensemble members grouped subjectively into four clusters. Numbers in the upper-left-hand corner indicate the ensemble member as defined in Table 1. Isolines every 1 mm.

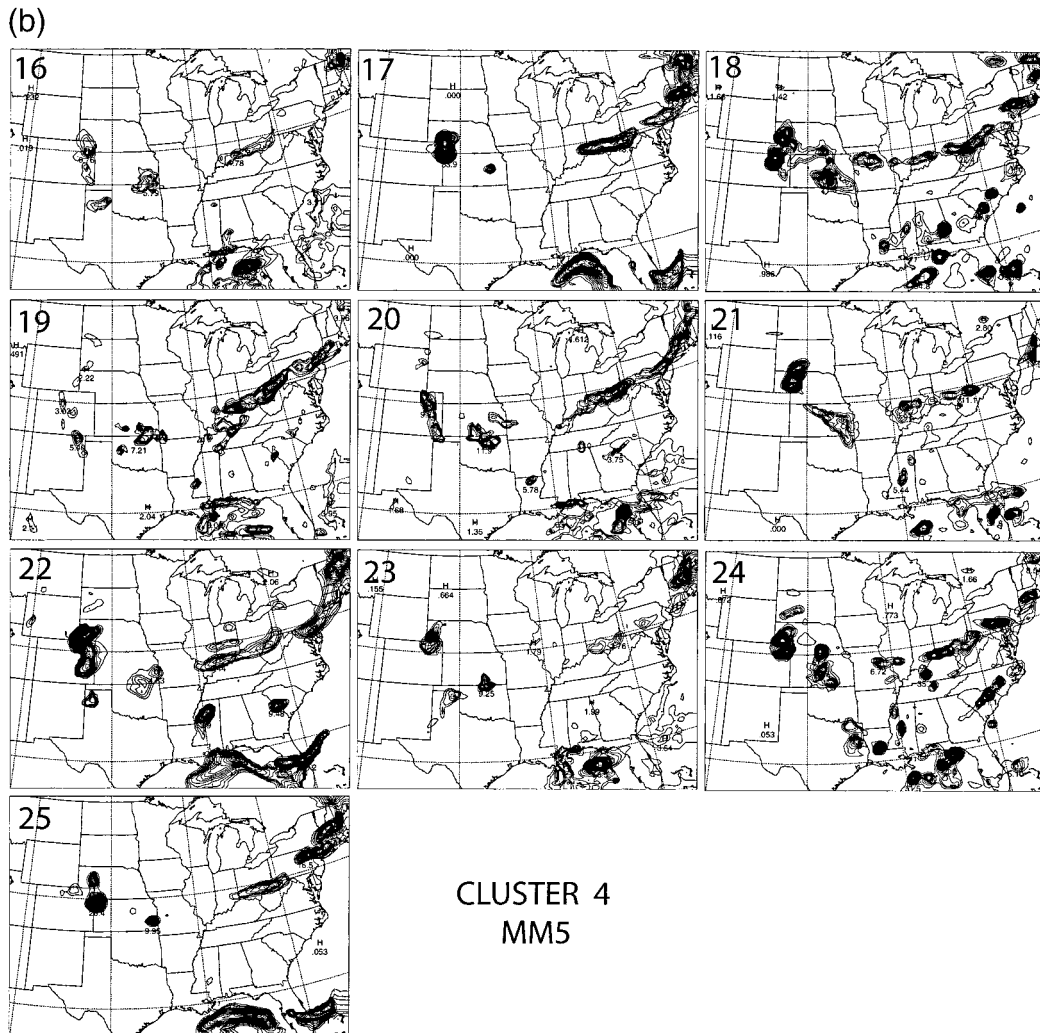


FIG. 3. (Continued)

again highlights the important role played by model physics in the generation of precipitation forecasts, since the one individual model with variations in model physics has the lowest variance explained by the first eigenvalue.

Clustering the 25 members of the ensemble using PCA yields a similar result as found with CA when varimax, an orthogonal rotation algorithm, is used. A strong simple structure is obtained by rotating the first four PCs (Table 2). Examination of the PCA clustering structure of the 25 members at 36 h shows that the members of each cluster all belong to a single model; there is no intermodel clustering found. This result is also found using the precipitation output at 24, 27, and 30 h.

The results from both the correlation-based CA and PCA agree on the tendency of the members from one model to be in one cluster. Therefore, the correlation between the forecast members from a single model is

usually higher than the correlation between forecast members from different models in this dataset.

b. Precipitation: CA and PCA based on Euclidean similarity

The purpose of this analysis is to provide another similarity-based CA and PCA for comparison with results found using the correlation similarity measure. For each output time, the data matrix is first centered and then the Euclidean similarity matrix (2.6), $\mathbf{E}_{25 \times 25}$, is computed. Based on the \mathbf{E} matrix, CA using the UPGMA method is applied. The results are then compared to the correlation-based CA.

The clustering dendrograms resulting from the Euclidean similarity measure are different from its correlation-based counterpart (cf. Figs. 2 and 5). Although the tendency of one model to form one cluster still exists to some extent, intermodel clusters are found at several

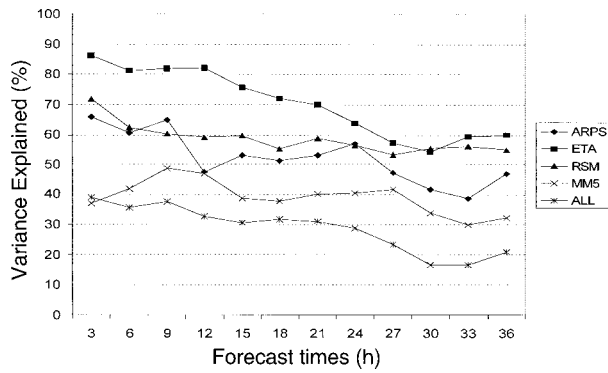


FIG. 4. Comparison of the variance explained (%) by the first eigenvalue across the four models for the 3-h accumulated precipitation dataset vs forecast time (h). Key indicates the model clusters shown.

output times. At 36 h, one large cluster containing members from different models is formed at a relatively early stage of the hierarchy, which points out the similarity between these members.

When the Euclidean similarity clusters are compared with the model forecast fields at 30 h (Fig. 3), it is more challenging to explain the cluster groupings. Euclidean similarity is influenced by the precipitation amplitudes in addition to location (Elmore and Richman 2001), such that the clusters may in part reflect the differences in the precipitation maximum values. For this output time the Eta Model and ARPS typically have lower maximum values of precipitation than the RSM and MM5. Note that the Eta Model and ARPS are grouped together before the other two models (except for ARPS member 5, which has the highest precipitation maximum of all the ARPS forecasts). Therefore, which cluster measure one prefers may depend upon how important are precipitation amplitudes in comparison with precipitation patterns.

A PCA is conducted using the same Euclidean similarity matrices. A strong simple structure is obtained by rotating six principal components (PCs) using the varimax algorithm. By examining the varimax simple structure (Table 3), it is seen that the clustering structure of the 25 members at 36 h also has members from different models. Yet there remain instances where the forecasts from a single model tend to cluster together, even though they are part of a larger cluster. The model forecasts that form the largest number of clusters are from the MM5. These results are qualitatively similar to those from the CA-based analysis using the correlation measure.

c. Precipitation: CA based on the Euclidean distance

The final measure used in evaluating the ensemble data is Euclidean distance (2.5), which is a dissimilarity measure. For each time instance, a distance matrix, $E_{25 \times 25}$, based on Euclidean distance is computed for the centered data matrix using (2.3) and used as the dissim-

TABLE 2. Rotated loadings of the first four PCs using varimax for the 3-h accumulated precipitation data at the 36th hour based on the Euclidean correlation. Italics indicate the largest loading for a given model forecast.

n	PC1	PC2	PC3	PC4
1	0.41	0.07	0.09	<i>0.65</i>
2	0.05	0.03	0.02	<i>0.81</i>
3	0.25	0.05	0.04	<i>0.64</i>
4	0.22	0.07	0.13	<i>0.31</i>
5	0.10	0.02	0.12	<i>0.62</i>
6	0.13	0.07	<i>0.83</i>	0.23
7	0.03	<i>0.89</i>	-0.02	0.04
8	0.43	0.15	<i>0.71</i>	-0.10
9	0.07	<i>0.94</i>	-0.04	0.03
10	0.28	0.18	<i>0.83</i>	-0.04
11	0.07	<i>0.91</i>	0.02	0.03
12	-0.20	0.03	<i>0.67</i>	0.32
13	-0.03	<i>0.31</i>	0.17	0.16
14	-0.09	0.03	<i>0.60</i>	0.55
15	0.03	<i>0.33</i>	0.15	-0.01
16	<i>0.67</i>	0.09	0.06	0.20
17	<i>0.61</i>	0.00	0.07	0.13
18	<i>0.45</i>	-0.03	-0.10	0.09
19	<i>0.59</i>	0.09	0.11	0.30
20	<i>0.44</i>	0.04	0.00	0.16
21	<i>0.39</i>	0.00	0.03	-0.02
22	<i>0.53</i>	0.09	0.15	-0.03
23	<i>0.56</i>	0.03	0.02	0.17
24	<i>0.47</i>	-0.02	0.14	-0.05
25	<i>0.65</i>	0.01	0.03	0.15

ilarity matrix for the CA. Ward's method is used to determine the clusters. It is expected that the resultant clustering hierarchies determined by Ward's method will be different from their UPGMA counterparts, due to the sensitivity of cluster analysis to both the changes of the clustering methods and the resemblance measures.

The clustering dendrograms resulting from the Euclidean distance measure indicate that it is not possible to obtain a clustering structure that corresponds only to the four models at any of the forecast times (Fig. 6). Although the four-model clustering structure does not exist, the tendencies of the members of one model to build their own cluster before merging with members from other models, as observed in the previous analyses, is strong. However, the intermodel clustering is seen in several cases.

The nature of the Euclidean distance measure makes it difficult to assess the degree of similarity between the members of a given cluster. But the intermodel clusters are formed at earlier stages in the hierarchies when compared to the correlation-based hierarchies. These results, and those from the similarity measures, suggest that the Euclidean-based CA and PCA view the members of the ensemble differently and reveal the intersimilarity embedded in the four models that, collectively, construct the ensemble.

d. Precipitation: CA based on the k-mean and nucleated agglomerative (NA) algorithms

Our results thus far are determined only using hierarchical clustering methods. To further explore the en-

TABLE 3. Rotated loadings of the first six PCs using varimax for the 3-h accumulated precipitation data at the 36th hour based on the Euclidean similarity measure. Italics indicate the largest loading for the given model forecast.

n	PC1	PC2	PC3	PC4	PC5	PC6
1	<i>0.83</i>	0.20	0.03	-0.23	0.15	0.09
2	<i>0.83</i>	0.19	0.00	-0.17	0.14	0.08
3	<i>0.79</i>	0.19	0.03	-0.16	0.15	0.07
4	<i>0.78</i>	0.20	0.02	-0.14	0.18	0.07
5	<i>0.81</i>	0.19	0.00	-0.20	0.14	0.08
6	<i>0.83</i>	0.19	0.02	-0.18	0.07	0.10
7	0.20	<i>0.80</i>	-0.05	-0.05	0.00	0.08
8	<i>0.77</i>	0.22	0.12	-0.17	0.13	0.13
9	0.31	<i>0.84</i>	0.00	-0.09	0.07	0.04
10	<i>0.80</i>	0.22	0.08	-0.18	0.09	0.15
11	0.32	<i>0.81</i>	-0.01	-0.11	0.05	0.04
12	<i>0.85</i>	0.19	0.00	-0.12	0.08	0.09
13	<i>0.55</i>	0.15	-0.32	-0.13	-0.24	-0.05
14	<i>0.86</i>	0.19	-0.01	-0.15	0.08	0.08
15	0.28	0.12	-0.05	-0.06	0.00	<i>0.94</i>
16	<i>0.80</i>	0.21	0.05	-0.25	0.18	0.10
17	0.41	0.12	0.05	-0.75	0.06	0.04
18	0.23	0.03	-0.06	-0.26	<i>0.78</i>	0.01
19	<i>0.81</i>	0.21	0.03	-0.26	0.15	0.09
20	<i>0.77</i>	0.20	0.03	-0.22	0.17	0.08
21	0.47	0.10	0.11	0.09	<i>0.54</i>	-0.01
22	<i>0.63</i>	0.20	0.02	-0.31	0.15	0.04
23	<i>0.78</i>	0.20	0.08	-0.20	0.20	0.09
24	0.24	-0.02	<i>0.92</i>	-0.09	-0.02	-0.05
25	0.39	0.12	0.03	-0.77	0.14	0.05

semble data, we now turn to two nonhierarchical algorithms, *k*-mean and NA. Nonhierarchical, or partitional, methods require the number of clusters and the seed point for each cluster to be specified in advance (appendix). For each forecast output time of the precipitation dataset, the *k*-mean algorithm is applied. The number of clusters chosen is four (*k* = 4), and the seed points are specified subjectively by choosing one seed from each model. The clustering structures resulting from applying *k*-mean show at least one intermodel cluster for each forecast time (Table 4).

To test whether the intermodel clusters break into small clusters, each having members from one model, the *k*-mean algorithm is run with the number of clusters successively increasing by one. The seed points again are specified subjectively. With *k* values of 5–19, *k*-mean continues to produce intermodel clusters (not shown). Typically, *k*-mean produces small clusters with three members, each of which is from a different model. With *k* = 20, no intermodel cluster is found using the *k*-mean algorithm.

The NA algorithm is applied on the 36-h forecast precipitation data. The range of the number of clusters is chosen to be $k_{max} = 7$, $k_{min} = 2$, and the seed points are specified subjectively to span the four models. Results indicate that the resultant partitions from applying the NA algorithm are consistent with the results from *k*-mean when comparing the four-cluster partition. An intermodel cluster is found by NA that combines members from ARPS, the Eta Model, and MM5.

These results indicate that both the resemblance measures and algorithms chosen determine the clusters that are found.

5. Analysis of other fields

While the 3-h accumulated precipitation data are likely the most variable of all the datasets saved during SAMEX, the discontinuous nature of the precipitation field makes interpreting the clustering results more difficult. Therefore, other standard output fields, which are more smoothly varying, also are examined. The fields selected are mean sea level pressure, 250-hPa wind speed, 500-hPa geopotential height, and CAPE. As before, results from the similarity measures are discussed first, followed by the results from the dissimilarity measure.

a. Sea level pressure: CA based on correlation and Euclidean distance

The clustering dendrograms resulting from applying the UPGMA method on the mean sea level pressure fields, based on the correlation measure, show that the dendrogram heights increase with time (Fig. 7). This suggests that the model fields are diverging with time. However, starting from 6 h, the members of the ensemble follow the same pattern, forming three clusters at a high degree of correlation (more than 0.8). The first cluster contains the 5 members of the ARPS model, the second cluster contains the 10 members of both the Eta Model and RSM, and the third cluster contains the 10 members of MM5. Therefore, although the height of the trees is increasing, the divergence is occurring between the three clusters themselves and not between the members of the individual cluster. As for the individual clusters, they continue to be formed at approximately the same degree of correlation. This indicates the compactness of the individual clusters and the growing isolation of the three clusters as time increases in view of the correlation of the 25 members. Therefore, the members are grouped into three clusters corresponding to the three institutions that collectively construct the ensemble.

Since mean sea level pressure is a derived field and not explicitly forecast by the models, the differences in the sea level pressure derivations could explain the clustering found. A close examination of the fields, however, indicates that this conclusion is not warranted. Sea level pressure fields at 36 h from the model forecasts show significant differences in the locations of important, large-scale meteorological features (Fig. 8). The ARPS forecasts all indicate a pressure trough to the east of the Rocky Mountains with a distinct low center located somewhere in Kansas. A strong region of high pressure is located over the Great Lakes region. In contrast, both the Eta Model and RSM have a single low pressure center in the northern plains region and a much weaker

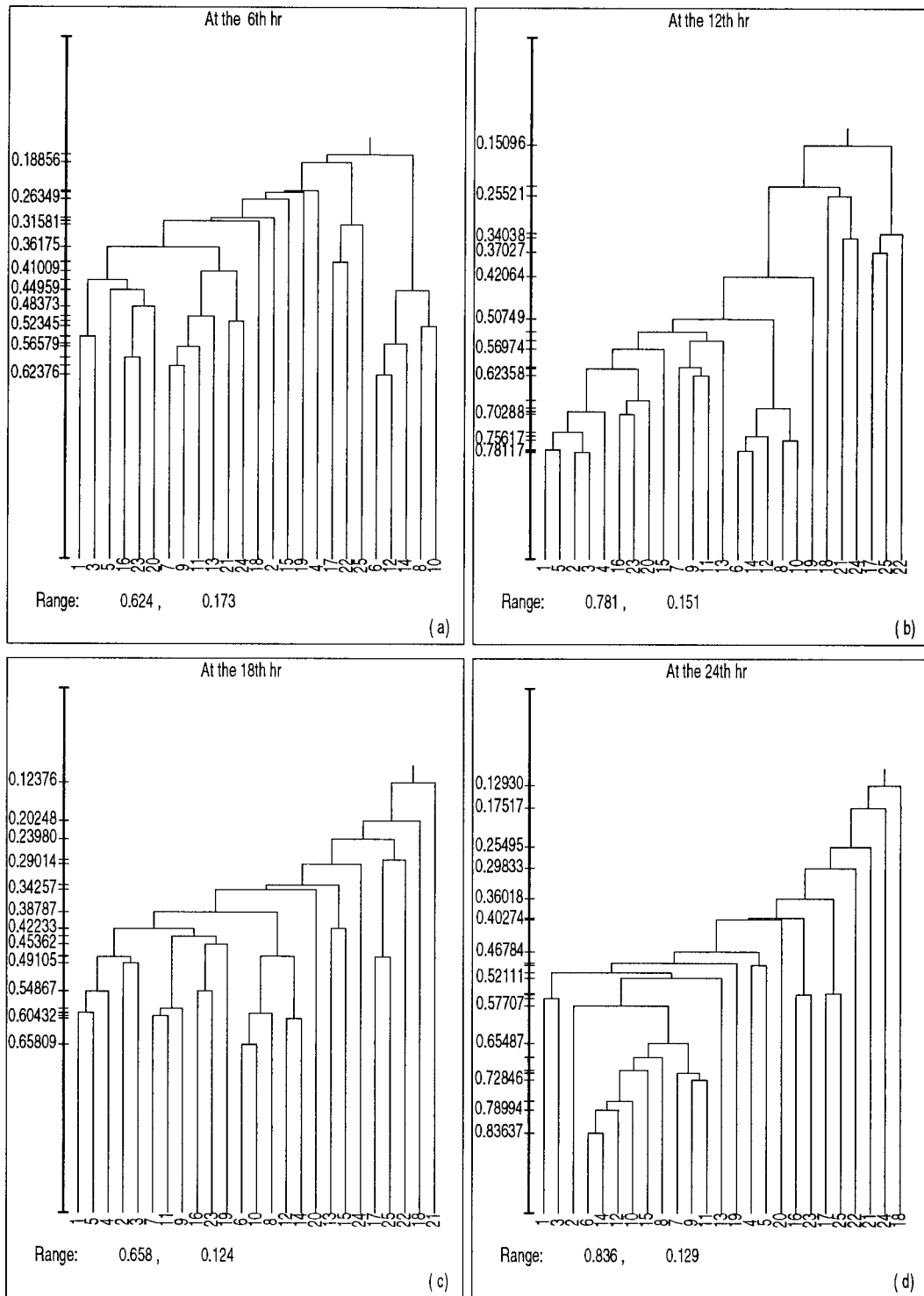


FIG. 5. UPGMA dendrograms for the accumulated precipitation based on the Euclidean similarity measure for forecast times of 6, 12, 18, 24, 30, and 36 h. Data are centered using (2.3).

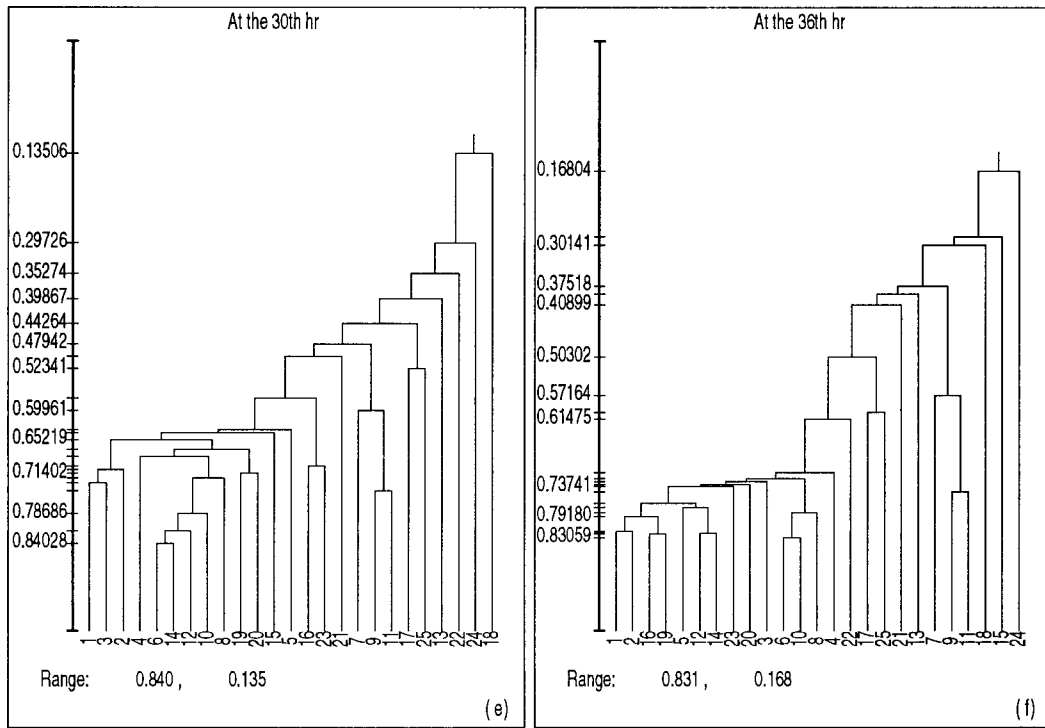


FIG. 5. (Continued)

region of higher pressure over the Great Lakes. The MM5 forecasts provide yet a third general solution, with various weaker low pressure centers in the plains states and a more distinct region of high pressure over the Great Lakes region. The greater structure in the sea level pressure fields over the plains states seen in the MM5 forecasts is due in part to the presence of deep convection in the model forecasts in this region, which influences the surface pressure through the inclusion of downdrafts in two of the convective parameterization schemes used and the explicit microphysical parameterization. The patterns of sea level pressure from the three clusters are so different that it is difficult to ascribe them to the derivation used; the structures are more indicative of differences in the timing and evolution of large-scale features in the model forecasts.

The clustering dendrograms resulting from applying Ward's method on the sea level pressure fields, based on the Euclidean distance, again shows the three clusters of the ARPS, Eta Model and RSM, and MM5 (Fig. 9). Owing to the nature of the Euclidean distance measure, it is not straightforward to assess the degree of compactness and isolation by examining the height of the trees. However, because all the trees are drawn using the same vertical scale, it is possible to use the levels of mergence to intuitively evaluate the degree of similarity among the various clusters. Such an analysis supports our earlier observation that the three clusters are compact and isolated from each other.

b. The 250-hPa wind speed: CA using correlation and Euclidean distance

Results from applying the UPGMA method on the 250-hPa wind speed data (Fig. 10), based on the correlation measure, support the conclusions found using the mean sea level pressure data. Although there is more variation in the clusters over the forecast interval than found with the sea level pressure data, the members of this ensemble are largely correlated. They are all placed in one large cluster with a correlation of 0.82 by 36 h. From 0 to 18 h, the clustering structure changes from time to time. However, from 21 h onward the members follow the same three-cluster pattern seen using the mean sea level pressure data (i.e., ARPS, Eta and RSM, MM5). Inside the Eta and RSM cluster, the Eta cluster is built first before building the large cluster; hence, no intermodel cluster is found.

Examination of the 250-hPa wind fields from the model forecasts again shows differences in the large-scale features predicted (not shown). The ridge over the Rocky Mountains in the Eta Model and RSM is farther east than in the ARPS or MM5, and the winds are stronger in the Eta Model and RSM than in the other two models. In contrast, the trough over the eastern United States is deeper in the ARPS and MM5 forecasts than in the Eta Model and RSM forecasts, which have a more zonal flow pattern. Thus, the three clusters found using the CA are supported by a subjective analysis of these data and illustrate that the models are producing dif-

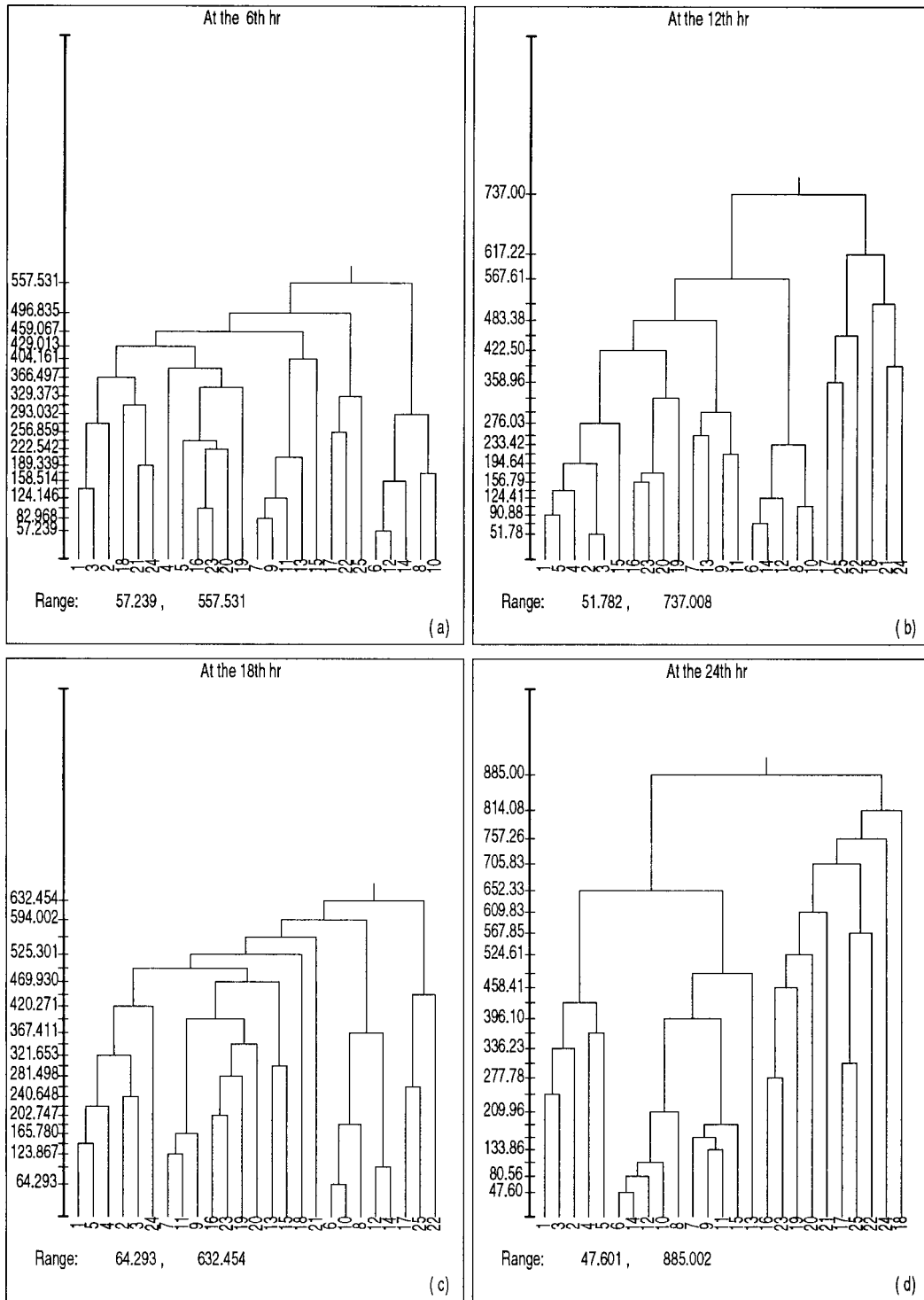


FIG. 6. Ward dendrograms for the accumulated precipitation based on the Euclidean distance measure for forecast times of 6, 12, 18, 24, 30, and 36 h. Data are centered using (2.3).

ferent evolutions of large-scale atmospheric features even on timescales as short as 36 h or less. It is also interesting to note that even with bred mode perturbations in the Eta Model and RSM, the forecast 250-hPa

wind speeds from these two models are most alike; the model variability is exceeding the initial condition variability for this day.

The clustering dendrograms resulting from applying

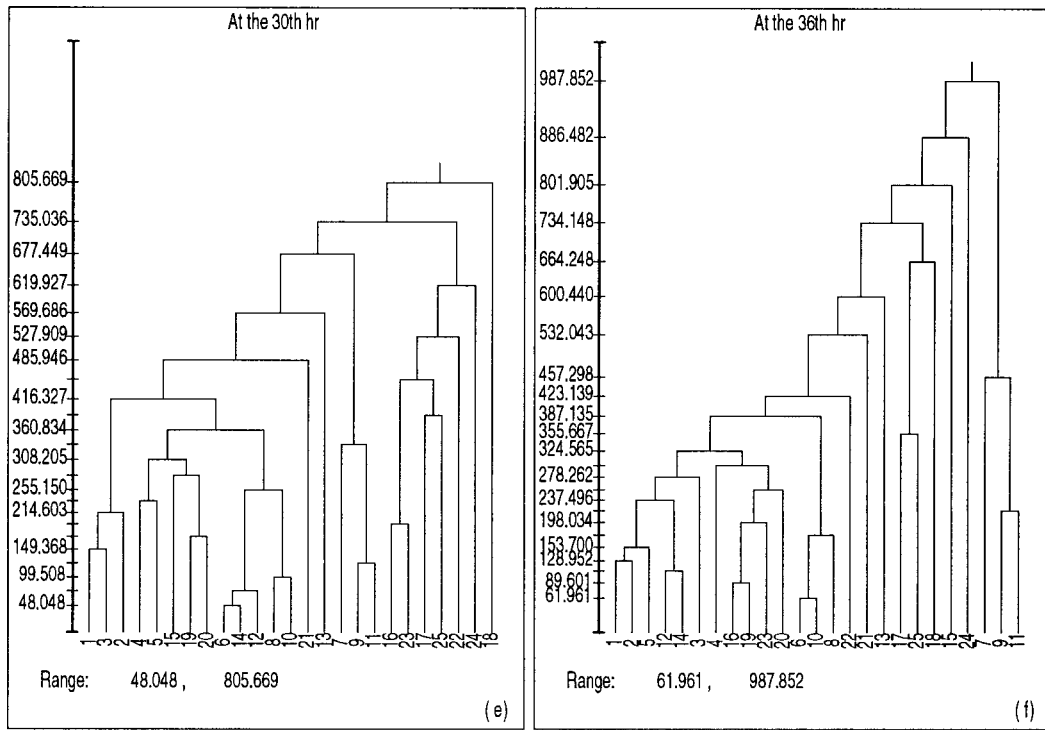


FIG. 6. (Continued)

Ward’s method on the wind speed data, based on the Euclidean distance (Fig. 11), show that the members follow a four-model clustering structure (i.e., ARPS, Eta, RSM, MM5). At 36 h, members form three clusters of ARPS, Eta and RSM, and MM5. Inside the Eta and RSM cluster, the Eta cluster again is built first before building the large cluster, hence, no intermodel cluster is found, supporting the conclusion found from the correlation measure.

c. The 500-hPa height field: CA based on correlation and Euclidean distance

The clustering dendrograms resulting from applying UPGMA method on the 500-hPa geopotential height data, based on the correlation measure, indicate that CA fails to recover the clustering structures because members are extremely correlated (Fig. 12). In essence, all the trees are so short that no useful information on the clusters can be obtained. By 36 h the forecast fields are joined into one large cluster at correlation value of 0.96. However, even under this situation the members are seen to form three clusters—ARPS, Eta and RSM, and MM5—during the last three output times.

The clustering dendrograms resulting from applying Ward’s method on the geopotential height data, based on the Euclidean distance measure, indicate that the resultant clustering structures change over time with respect to cluster membership (Fig. 13). The members from ARPS and MM5 tend to form their own clusters

at a relatively early stage in the hierarchy, before merging with members from other models. At a few output times, Eta Model members build their own cluster. Starting from 30 h onward, the members of the ensemble follow the cluster pattern seen previously (i.e., ARPS, Eta and RSM, MM5). Inside the Eta and RSM three-cluster, the Eta cluster again is built first before the large cluster is formed; hence, no intermodel cluster is found. According to the various levels at which the hierarchies are built, the three clusters are reasonably isolated, with the MM5 cluster being the most compact one for all output times. Although Stensrud et al. (2000) argue that geopotential height may not be the most useful measure for evaluating ensemble forecast output, the conclusions drawn from these data continue to support the notion that the differences in model physics and numerics are dominating the clustering of the ensemble members.

d. CAPE: CA based on correlation and Euclidean distance

The clustering dendrograms resulting from applying the UPGMA method on the CAPE data, based on the correlation measure, indicate that between 3 and 33 h the members of the ensemble follow a four-model clustering structure of ARPS, Eta, RSM, and MM5 (Fig. 14). Members of each model form their own cluster before merging with members from other models. The Eta Model cluster is the most compact cluster, while the MM5 cluster is the least compact one. However, at 36 h the clustering

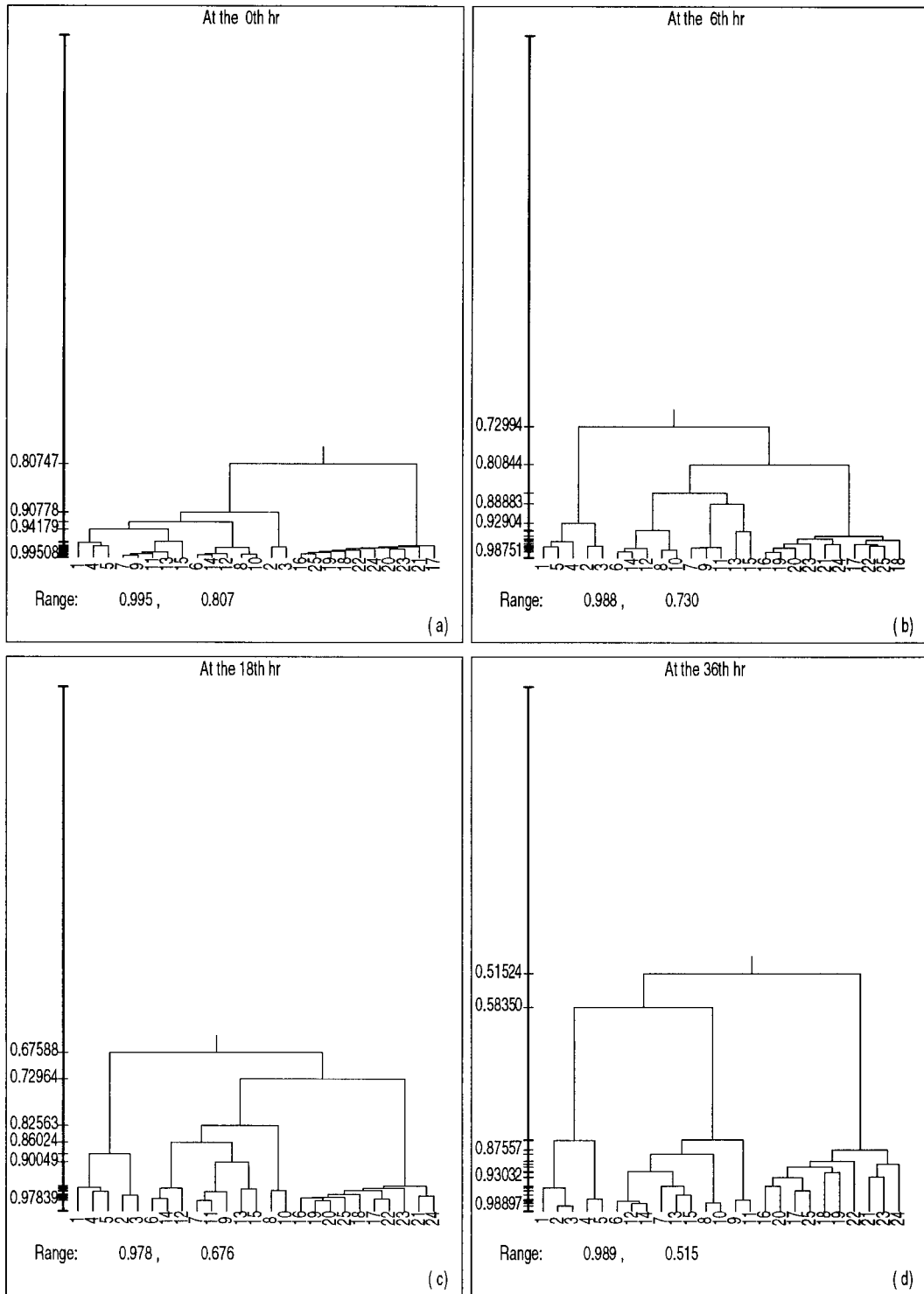


FIG. 7. UPGMA dendrograms for the mean sea level pressure field based on the correlation measure for forecast times of 0, 6, 18, and 36 h.

TABLE 4. Cluster partitions found using *k*-mean for the 3-h accumulated precipitation dataset at all the 3-h forecast times (T).

T	Clusters			
3	1, 4, 5, 16, 19, 20, 23	6, 8, 10, 12, 14	7, 9, 11, 13, 15, 17, 18, 21, 22, 24, 25	2, 3
6	1, 2, 3, 4, 5, 16, 19, 20, 23	6, 8, 10, 12, 14	7, 9, 11, 13, 15, 18, 21, 24	17, 22, 25
9	1, 2, 3, 4, 5, 15, 16, 19, 20, 23	6, 8, 10, 12, 14, 21, 24	7, 9, 11, 13, 18	17, 22, 25
12	1, 2, 3, 4, 5, 6, 8, 10, 12, 14, 15	16, 19, 20, 23	7, 9, 11, 13	17, 18, 21, 22, 24, 25
15	1, 2, 3, 4, 5, 16, 19, 20, 23	6, 8, 10, 12, 14	7, 9, 11, 13, 15, 22	17, 18, 21, 24, 25
18	1, 2, 3, 4, 5, 12, 14, 19, 24	6, 8, 10, 17, 22, 25	7, 9, 11, 13, 15	16, 18, 20, 21, 23
21	1, 2, 3, 4, 5	21	6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 19, 20, 23	17, 18, 22, 24, 25
24	1, 2, 3, 4, 5	16, 21, 23, 24	6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 19, 20	17, 18, 22, 25
27	1, 2, 3, 4, 5	18	6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 19, 20, 21	16, 17, 22, 23, 24, 25
30	1, 2, 3, 4, 5, 6, 8, 10, 12, 13, 14, 15, 19, 20, 21	18	7, 9, 11	16, 17, 22, 23, 24, 25
33	1, 2, 3, 4, 5, 6, 8, 10, 12, 14, 16, 19, 20, 21, 22, 23	13, 15	7, 9, 11	17, 18, 24, 25
36	1, 2, 3, 4, 5, 6, 8, 10, 12, 13, 14, 15, 16, 19, 20, 21, 22, 23	24	7, 9, 11	17, 18, 25

structure is different. Members from ARPS and Eta form their own clusters at a high degree of correlation, which emphasizes the closeness of the members of these two models. The MM5 members break into two clusters, with one group containing four members forming a separate cluster before joining the ARPS cluster. The other MM5 cluster group remains separate from all the other clusters. On the other hand, two members from the RSM join the Eta Model cluster before building one large cluster containing all the members of both the Eta Model and RSM.

Examination of the CAPE fields from the individual model runs again indicates that the clustering is not from using different methods to calculate CAPE, although the methods used are different. The CAPE fields indicate obvious differences in the large-scale structure of the fields that are impossible to explain from different calculation methods. The more northern placement of the low center in the Eta Model and RSM forecasts at 36 h results in a more northerly extension of the CAPE field as compared with the other models (Fig. 15). The ARPS forecasts all have values of CAPE exceeding 1000 J kg⁻¹ stretching from the Gulf of Mexico across eastern Texas and northward into Nebraska. In contrast, all the other models have a local minima of CAPE in eastern Texas, producing a vastly different pattern in the fields. Thus, the differences are mainly due to the use of the different models and not the model initialization or CAPE derivation procedures.

The clustering dendrograms resulting from applying Ward's method on the CAPE data, based on the Euclidean distance measure, show that the members follow the four-model clustering structure of ARPS, Eta, RSM, and MM5 at all output times (not shown). This agrees with the results from the correlation measure from 3 to 33 h. The Eta cluster again is shown to be the most compact cluster, followed by the ARPS cluster. Although the overall results of both analyses differ at 36

h, no intermodel cluster is found using either clustering algorithm/measure.

e. PCA based on the correlation and Euclidean similarity

In addition to being a tool for clustering, the correlation-based PCA helps to form a preliminary analysis of the divergence/convergence among the variables. This is done through the analysis of the eigenvalues and the variances they explain. A cross comparison of the variance explained by the first two eigenvalues of the correlation matrices for all the fields examined agrees with some of the CA results (Table 5). This is especially true for the 500-hPa geopotential height and 250-hPa wind fields. For these two fields, the first eigenvalue at every output time accounts for nearly all of the variance in the data. This explains why the correlation-based CA fails to recover the clustering structures for this field due to the fact that members are concentrated on one dimension. Therefore, all members are placed into one cluster with a very high correlation value.

The correlation-based PCA also is applied to the mean sea level pressure data at 36 h. A strong simple structure is obtained by rotating the first three PCs using promax (Table 6). Retaining only three PCs explains 92.9% of the total variance. By examining the simple structure shown in Table 6, the members are grouped into three clusters corresponding to the three institutions that collectively construct the ensemble, namely ARPS, Eta and RSM, and MM5. This structure is very consistent with the CA analyses of this field presented earlier. Results from using the Euclidean similarity measure are identical in that three clusters are formed that correspond to the three institutions that provided the model forecasts.

(a)

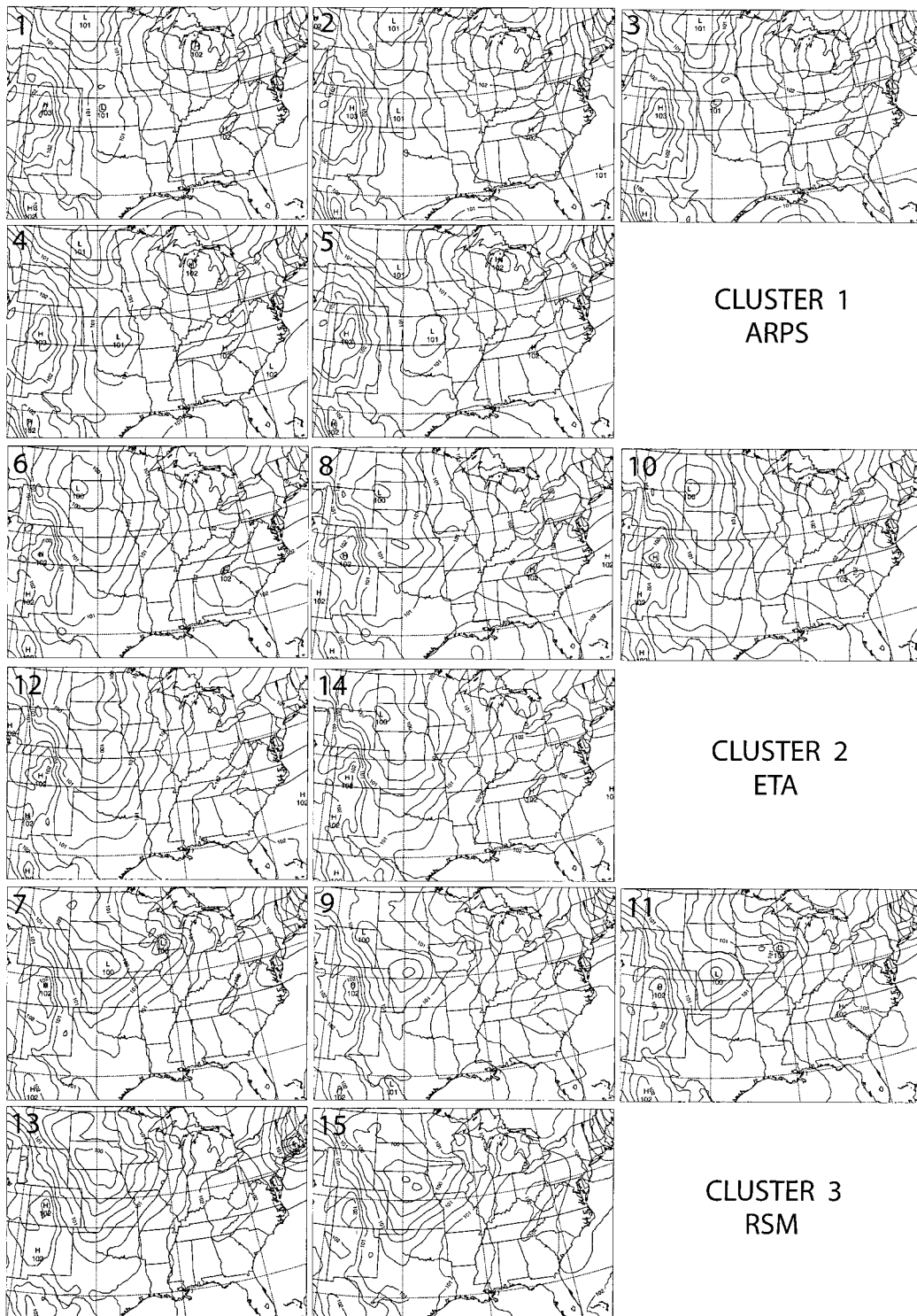


FIG. 8. Mean sea level pressure (hPa) valid at 36 h from all 25 ensemble members grouped subjectively into four clusters. Numbers in the upper-left-hand corner indicate the ensemble member as defined in Table 1. Isolines every 2 hPa. Relative high and low pressure regions are indicated.

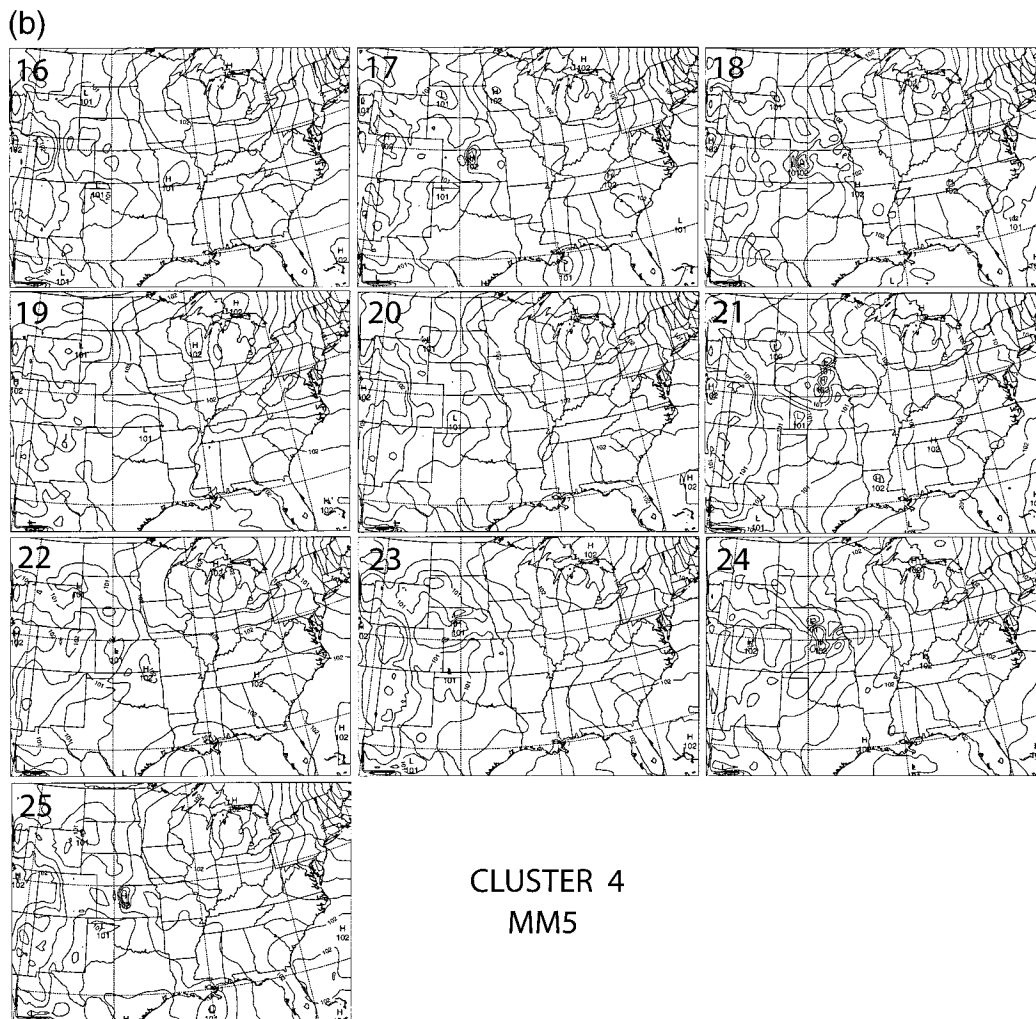


FIG. 8. (Continued)

6. Discussion

The goal of this work is to explore the importance of model diversity in short-range ensembles using an algorithmic clustering of multimodel, short-range ensemble forecasts. The multimodel ensemble data are from SAMEX, an experiment coordinated by CAPS at the University of Oklahoma and involving scientists at NCEP and NSSL. The full ensemble consists of 25 forecasts made from four different numerical weather prediction models: 5 forecasts from the ARPS, 5 forecasts from the NCEP Eta Model, 5 forecasts from the NCEP RSM, and 10 forecasts from MM5 run at NSSL. Each model forecast extends out to 36 h beginning from 0000 UTC 29 May 1998. Since two of the models (Eta, RSM) use the breeding of growing modes technique to initialize their forecasts, and one of the models (MM5) is configured with varying model physical parameterization schemes, the cluster analyses allow us to investigate the roles played by both model initial condition and model physics uncertainties. We are uncertain how the

results from a single case will change when more cases are analyzed, but note that the results presented are consistent with other studies as discussed below.

Two statistical techniques are used for these analyses, CA and PCA, and three different measures are used to define the similarity between the forecast fields, namely correlation, Euclidean similarity, and Euclidean distance. Based on the analyses of the 3-hourly accumulated precipitation fields, available every 3 h, it is observed that the members of each model tend to build their own cluster. The correlation-based analyses support this finding very strongly, while the other measures also show this tendency, but to a lesser extent. According to these analyses, the members of one model are shown to be coherent, while members from different models are isolated. Admittedly, the distance-based cluster analysis is able to reveal intermodels clusters, indicating similarities between members generated by different models. The existence of similarity across models is also found when using a nonhierarchical clus-

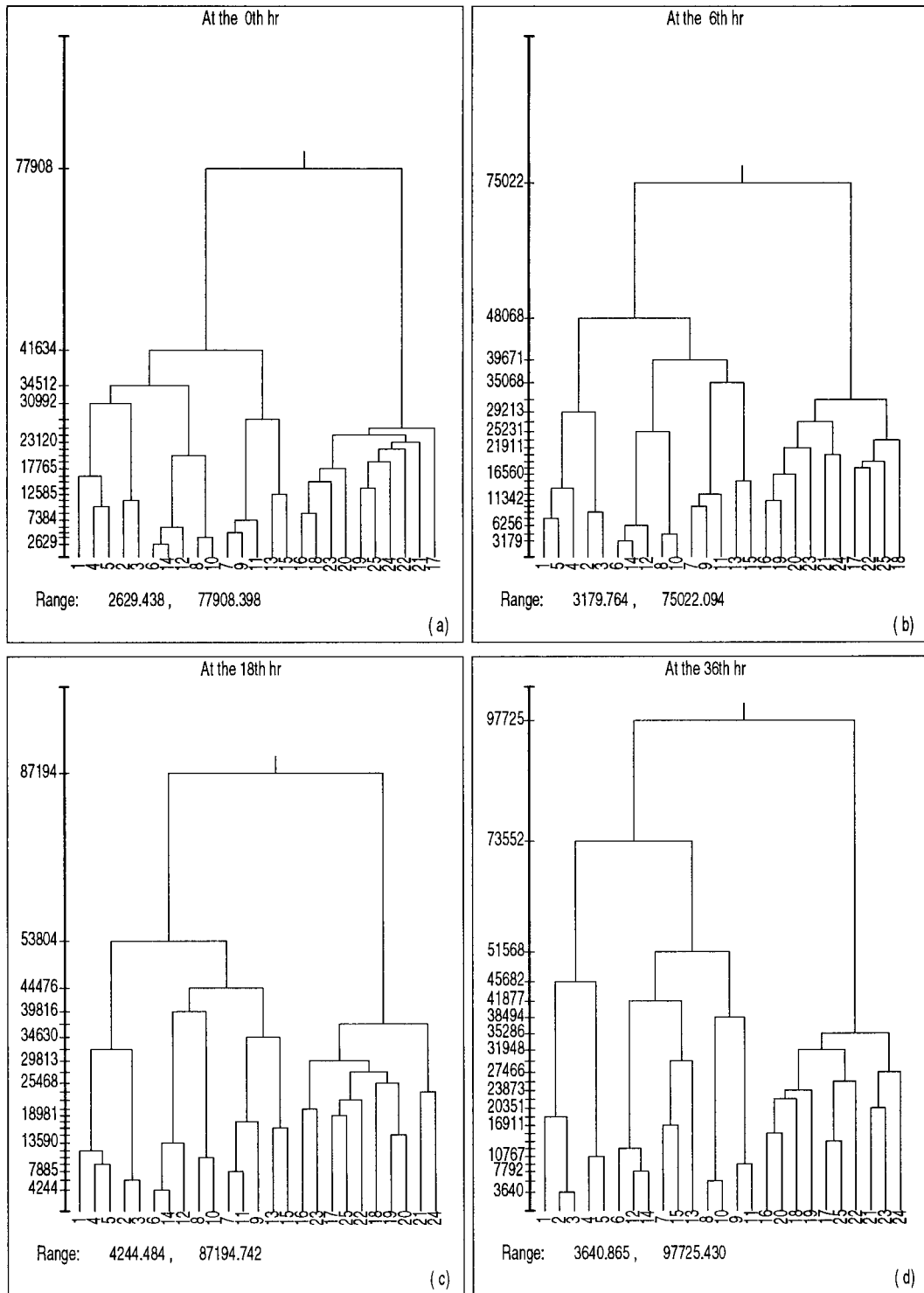


FIG. 9. Ward dendrograms for the mean sea level pressure field based on the Euclidean distance measure for forecast times of 0, 6, 18, and 36 h. Data are centered using (2.3).

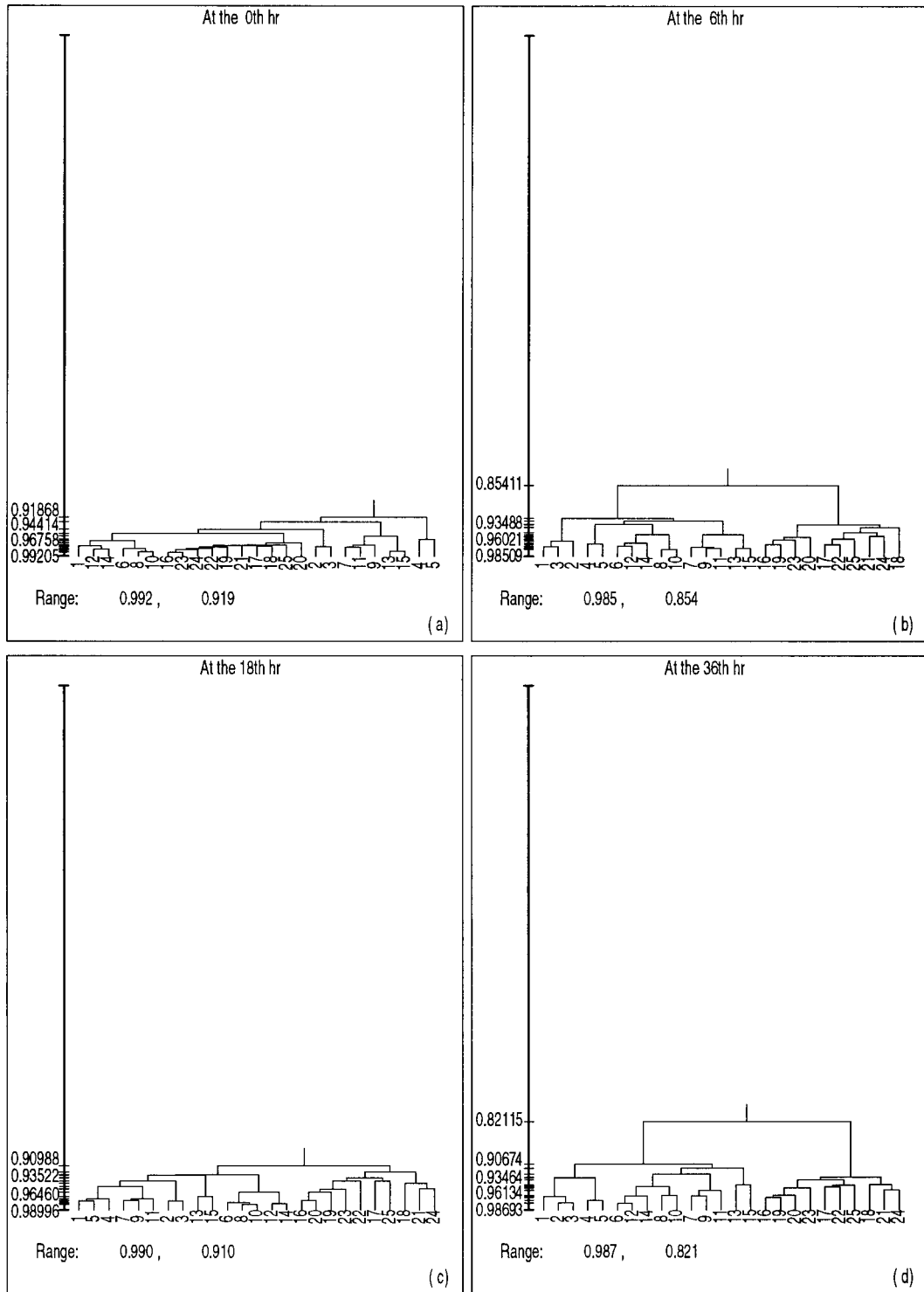


FIG. 10. UPGMA dendrograms for the 250-hPa wind velocity field based on the correlation measure for forecast times of 0, 6, 18, and 36 h.

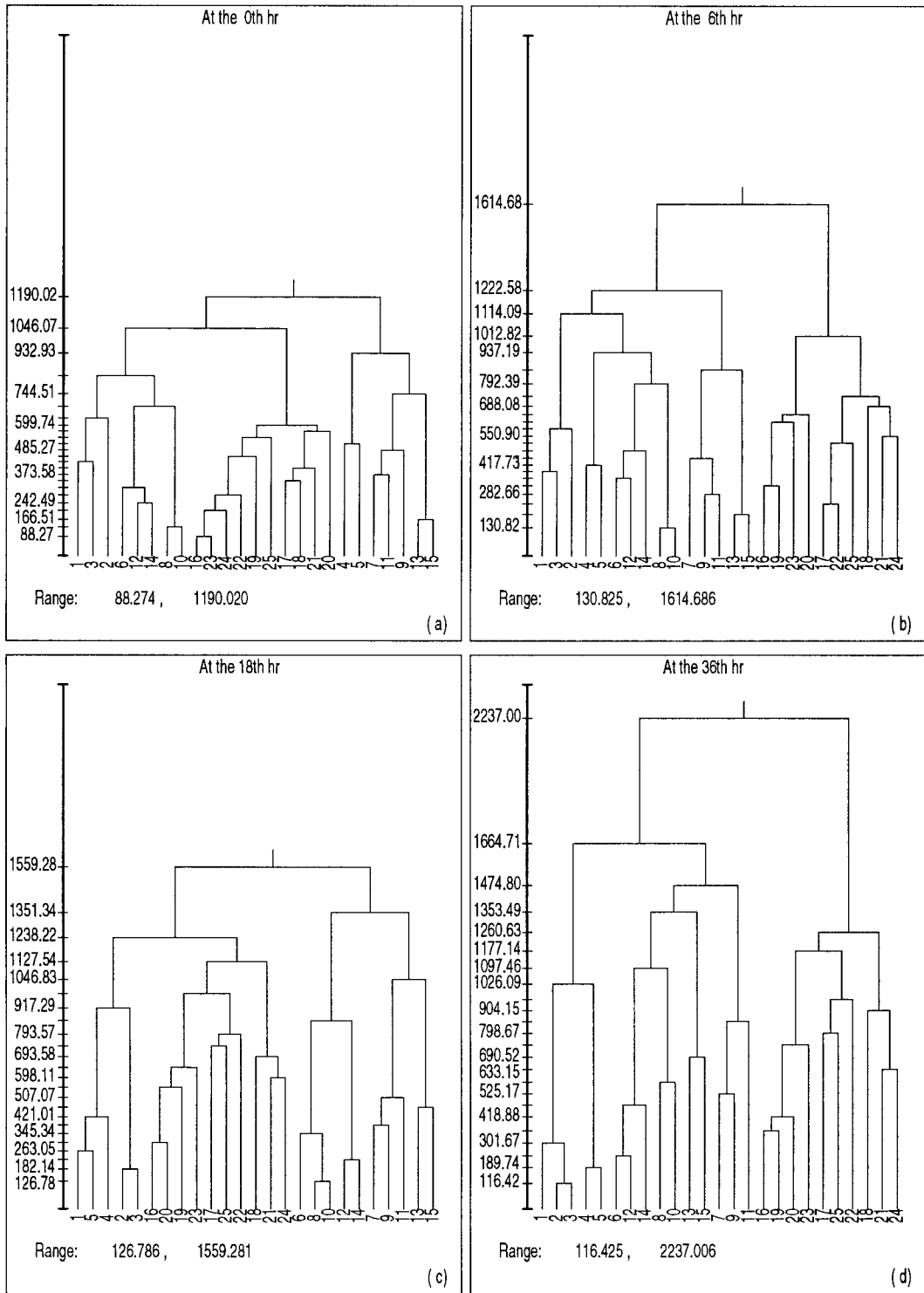


FIG. 11. Ward dendrograms for the 250-hPa wind velocity field based on the Euclidean distance measure for forecast times of 0, 6, 18, and 36 h. Data are centered using (2.3).

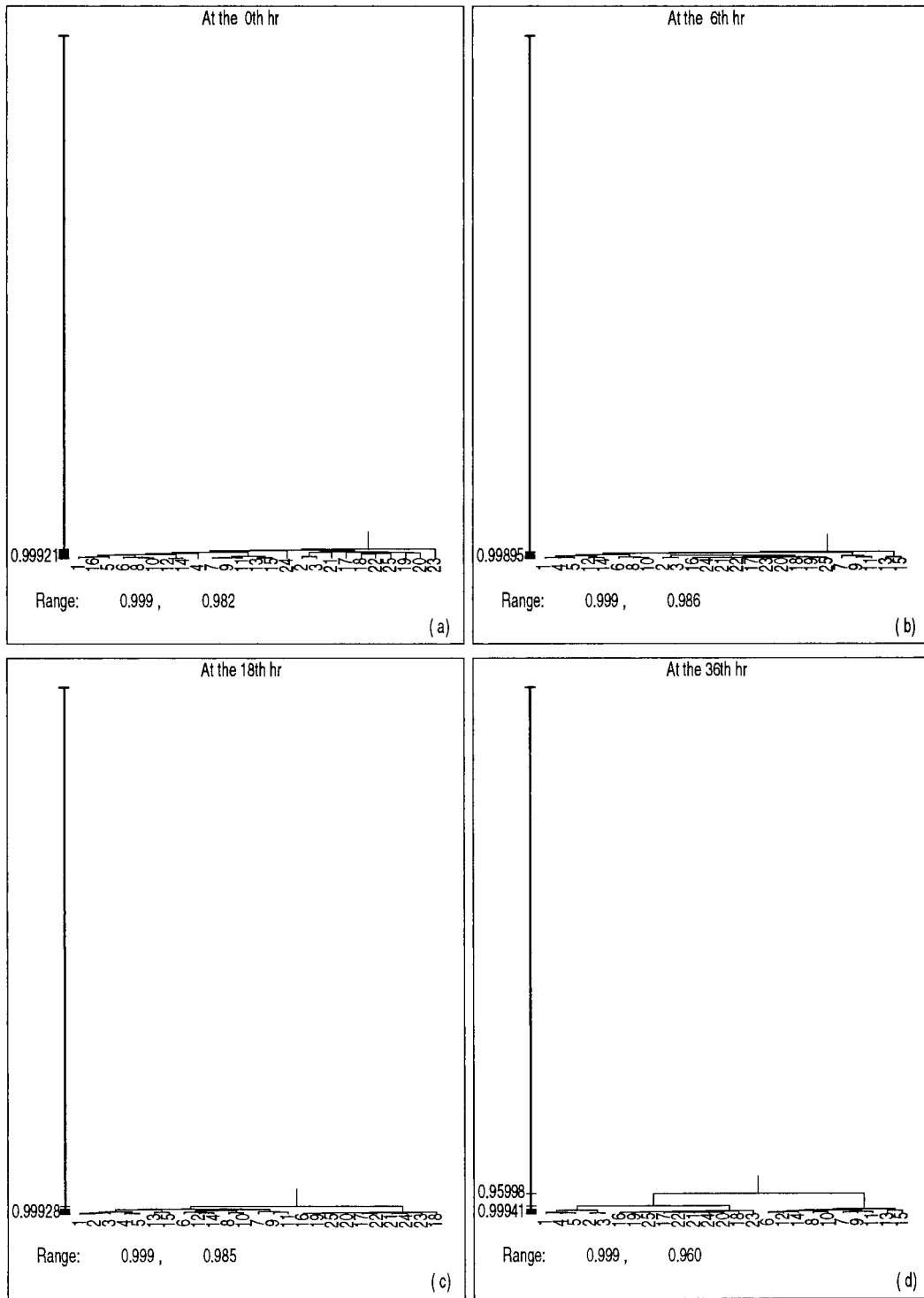


FIG. 12. UPGMA dendrograms for the 500-hPa geopotential height field based on the correlation measure for forecast times of 0, 6, 18, and 36 h.

tering algorithm such as the *k*-mean algorithm on these precipitation data.

When other forecast fields are examined, such as 250-hPa wind speed, 500 hPa geopotential height, CAPE,

and mean sea level pressure, the tendency for the forecast members to cluster by model still is found. Neither similarity-based nor dissimilarity-based CA or PCA find any intermodel cluster to exist over more than a few

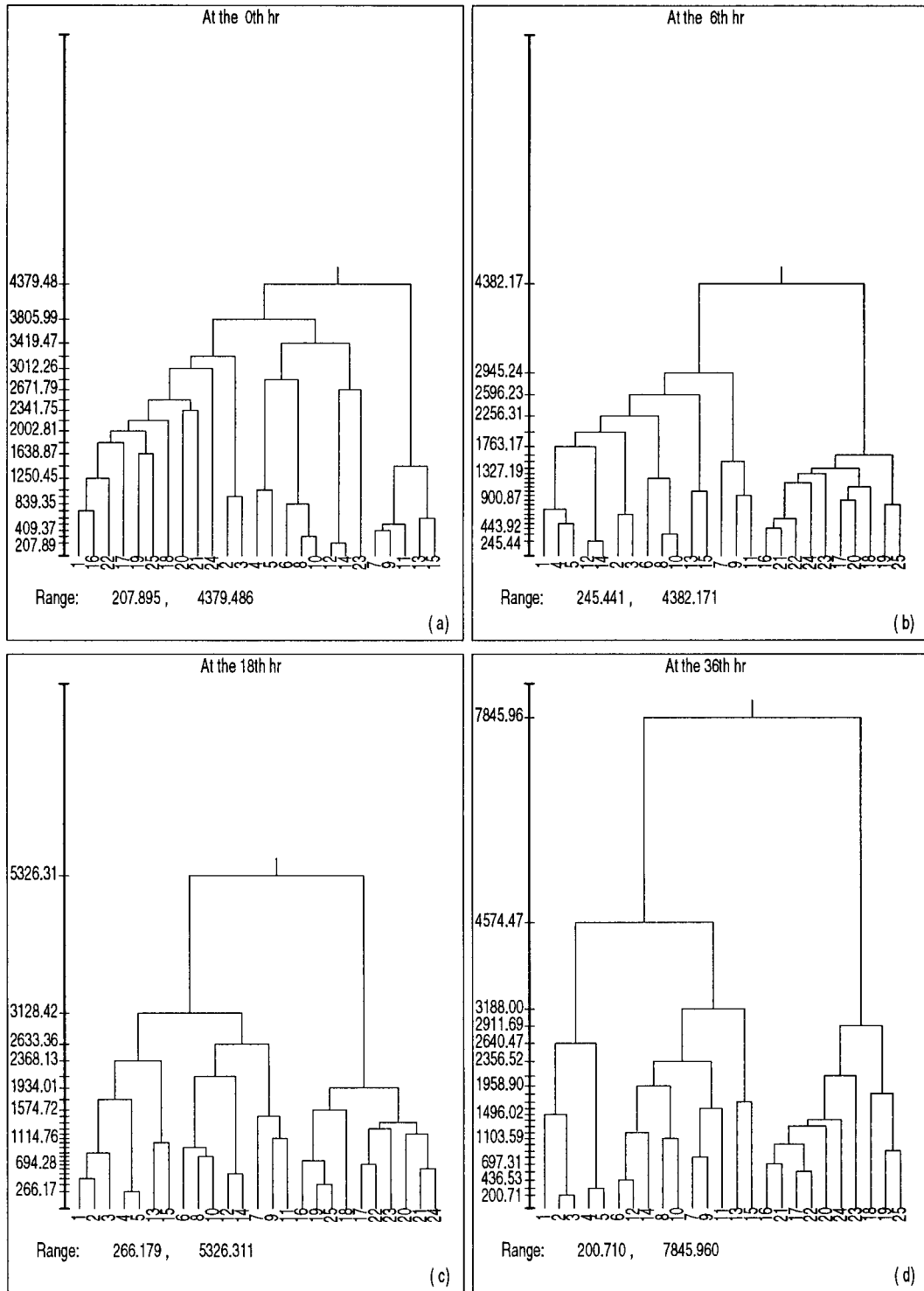


FIG. 13. Ward dendrograms for the 500-hPa geopotential height field based on the Euclidean distance measure for forecast times of 0, 6, 18, and 36 h. Data are centered using (2.3).

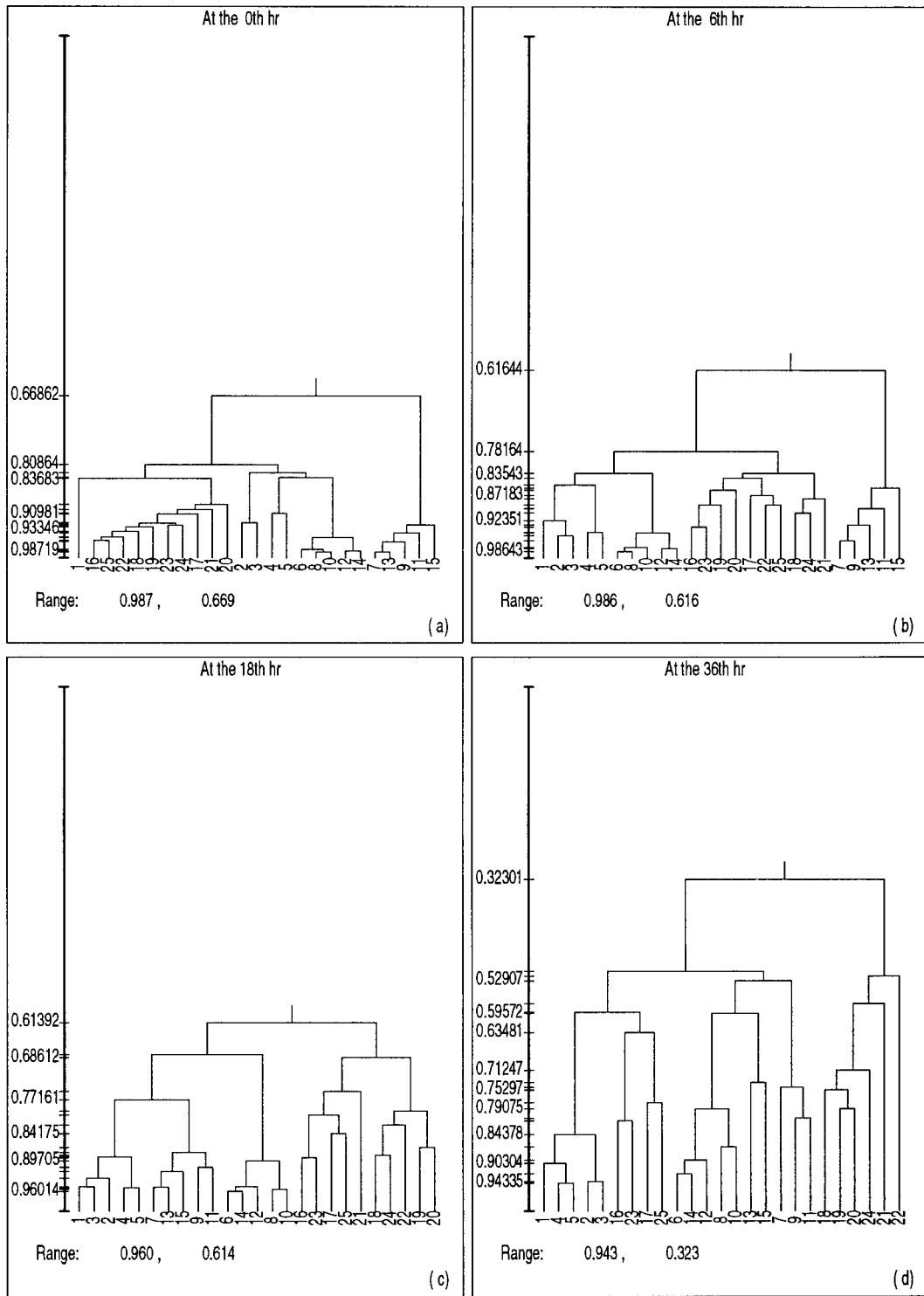


FIG. 14. UPGMA dendrograms for the CAPE field based on the correlation measure for forecast times of 0, 6, 18, and 36 h.

(a)

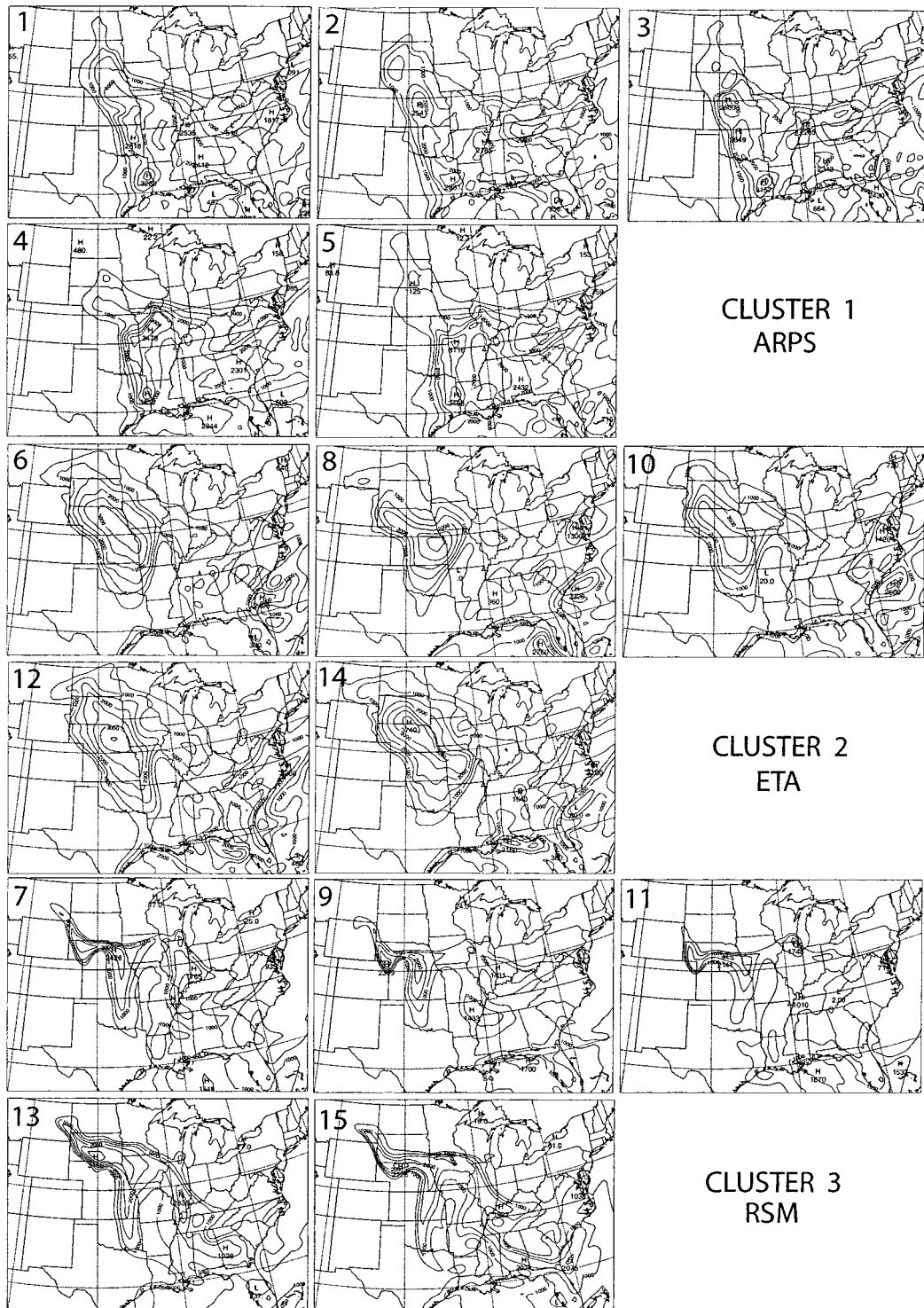


FIG. 15. Convective available potential energy (J kg^{-1}) valid at 36 h from all 25 ensemble members grouped subjectively into four clusters. Numbers in the upper-left-hand corner indicate the ensemble member as defined in Table 1. Isolines every 500 J kg^{-1} .

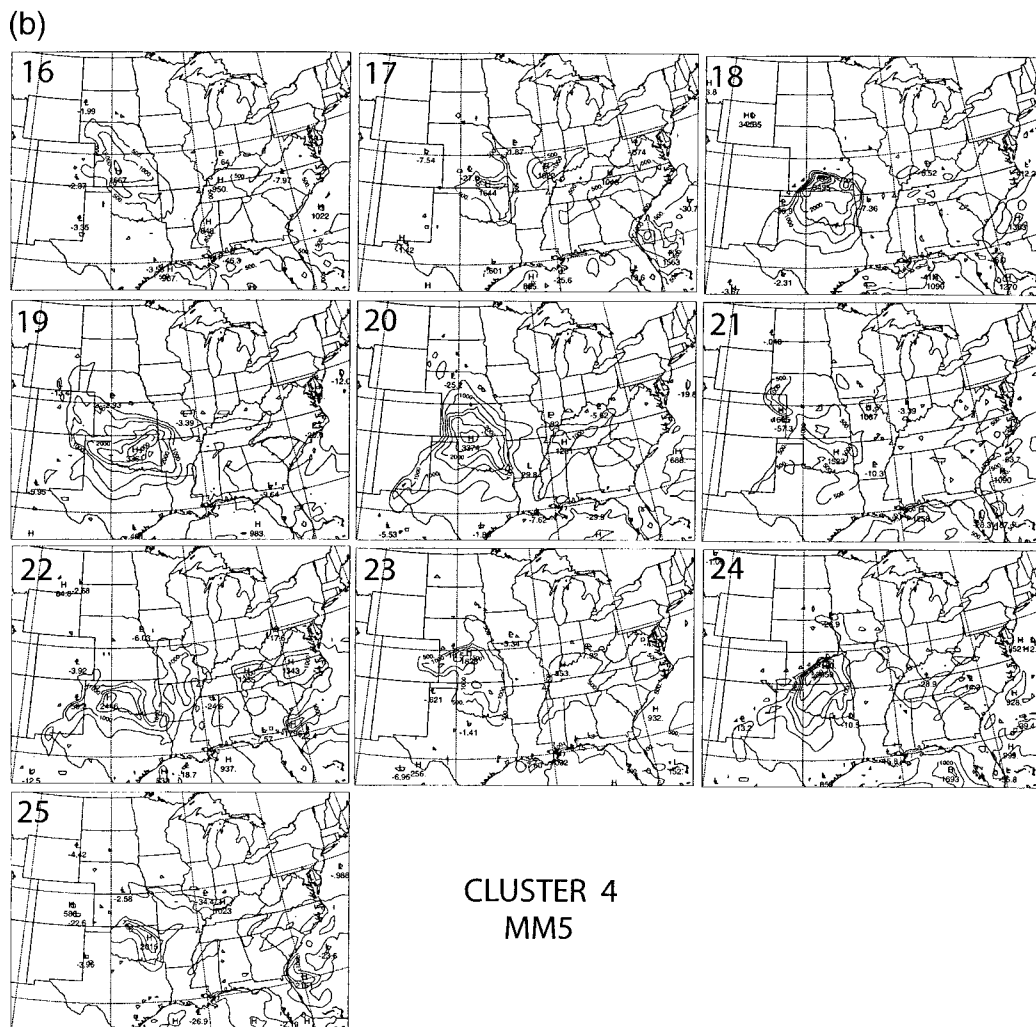


FIG. 15. (Continued)

model output times. The one strong tendency seen across all forecast times and fields is for the forecasts to cluster by the model used to produce the forecasts. This indicates that the model forecasts from a given model are more similar to themselves than to the initialization methods used.

It is observed that by 36 h the model results cluster with the model first, and then with the institution (i.e., Eta and RSM cluster next with each other, before clustering with ARPS and MM5). Even though the Eta and RSM are using the breeding of growing modes technique to generate the perturbed initial and boundary

TABLE 5. Comparison of the variance explained by the first two eigenvalues of the correlation matrices for various fields at the initial and finishing time [$\text{var}(\lambda_i)$ stands for the variance explained by λ_i].

Field	Time	$\text{var}(\lambda_1)$	$\text{var}(\lambda_2)$	$\text{var}(\lambda_1) + \text{var}(\lambda_2)$
Precipitation (mm)	Initial time	39.26	11.23	50.49
	Finishing time	20.94	10.41	31.35
Pressure (hPa)	Initial time	88.39	7.91	96.30
	Finishing time	67.14	16.26	83.40
CAPE (J kg^{-1})	Initial time	80.29	8.24	88.53
	Finishing time	51.47	13.73	65.20
500-hPa geopotential height (gpm)	Initial time	98.67	0.43	99.10
	Finishing time	97.70	1.78	99.48
250-hPa wind (m s^{-1})	Initial time	97.43	1.52	98.95
	Finishing time	89.64	7.13	96.77

TABLE 6. Rotated loadings of the first three PCs for the mean sea-level pressure data at the 36th hour based on the Euclidean correlation measure. The three clusters of italicized entries correspond to the three institutions (ARPS, Eta and RSM, and MM5) that collectively construct the ensemble.

n	PC1	PC2	PC3
1	0.04	0.03	<i>0.96</i>
2	-0.09	0.21	<i>0.89</i>
3	-0.10	0.17	<i>0.94</i>
4	0.08	-0.17	<i>0.99</i>
5	0.23	-0.17	<i>0.89</i>
6	<i>0.97</i>	0.03	-0.01
7	<i>0.98</i>	-0.06	0.05
8	<i>0.77</i>	0.16	0.13
9	<i>0.70</i>	0.10	0.25
10	<i>0.79</i>	0.14	0.14
11	<i>0.79</i>	0.10	0.13
12	<i>1.04</i>	-0.06	-0.07
13	<i>0.97</i>	-0.04	0.01
14	<i>1.07</i>	-0.03	-0.15
15	<i>1.03</i>	-0.13	0.02
16	-0.01	<i>0.98</i>	0.01
17	0.03	<i>0.96</i>	-0.03
18	-0.09	<i>0.99</i>	0.03
19	-0.32	<i>1.07</i>	0.08
20	0.02	<i>0.91</i>	0.11
21	0.28	<i>0.89</i>	-0.23
22	-0.14	<i>0.96</i>	0.13
23	0.32	<i>0.82</i>	-0.15
24	0.21	<i>0.91</i>	-0.16
25	-0.02	<i>0.91</i>	0.13

conditions, very little indication of the bred modes clustering together is seen (mode 1 from Eta and RSM clustering together, then mode 2 from Eta and RSM clustering together, etc.). The models all cluster together first by model, showing the very different model climates that appear quite quickly in the forecasts.

The height of the dendrograms can be used to evaluate how much the forecasts are alike, with taller connection points on the dendrograms indicating that the forecasts are less alike. A subjective analysis of all the clustering dendrograms indicates that the MM5 forecasts, which include model physics differences and use less rapidly growing initial conditions than the bred modes, are the most dissimilar of all the forecasts. This is particularly true of the precipitation forecasts, but also is evident in the other fields. Model physics differences, even when used within the context of a single model, provide for a substantial divergence in the solutions. However, more important is the use of different models in order to produce divergence in the forecast solutions. Our results strongly suggest that model differences provide much of the divergence in the solutions from this case and that the techniques used to vary the initial conditions are much less important. Even when using bred modes, both the Eta Model and RSM cluster with themselves first, and then they cluster together before clustering with either the ARPS or MM5. The bred mode forecasts are not outliers in the sense that they do not span the

range of the forecasts on this day. Indeed, the Eta and RSM often are the first models to cluster together. These results are supported further by our subjective analyses of the model forecast fields from selected time periods, which also indicates that the models are very similar to themselves and that it is the fields from different models that have very different structures.

While perturbing a single model does improve the forecast divergence, as shown here and by Buizza et al. (1999) and Stensrud et al. (2000), it is clear that using totally different models is likely a more fruitful approach. The addition of model diversity to the ensemble system yields forecasts with greater spread, comprising more possible solutions. However, the clustering results show that these solutions, when viewed using any of the fields selected, do not produce a smooth distribution. The solutions instead diverge rapidly into what appear to be distinctly separate envelopes of solutions with limited overlap. This behavior is suggestive of the ensemble members not being equally probable, but the addition of more cases may find that this is not the case. Regardless, Hou et al. (2001) find that each of the different model ensembles used here have their own strengths and weaknesses, but the verification measures are consistently better when all the model ensembles are combined. Wandishin et al. (2001) also show that the inclusion of a clearly inferior model still leads to improvements in model skill. As we begin to explore more ways to perturb the model physics within a single model framework, it will be interesting to see if we can capture the same degree of variability seen in multimodel ensembles.

Our results strongly suggest that model diversity is an important component of short-range ensemble forecasting systems, since the different models can contribute extra, useful information to the ensemble that would not otherwise be available. Verification of the SAMEX data by Hou et al. (2001) support this conclusion, as do the results of Atger (1999), Ziehmann (2000), and Evans et al. (2000) for medium-range ensembles. We hope that the use of completely different models in short-range ensemble forecasting will continue to be explored vigorously as it enhances the ensemble variability, which is one of the remaining concerns with present operational ensemble forecast systems.

Acknowledgments. This work would not have been possible without the participation by Joseph Schaefer, Russell Schneider, and Mike Baldwin of Storm Prediction Center (SPC), and Harold Brooks and Chuck Doswell of NSSL, each of whom spent many hours debating on several key issues that lie at the core of this project. Our special thanks are due to Michael Richman for his continued help and counsel, and to Ken Mylne and an anonymous reviewer for providing numerous, constructive comments on this manuscript. S. Lakshmivaran's effort has been supported by the senior summer visiting faculty through the UCAR visiting scientist program

offered by NCEP and hosted by SPC, Norman, Oklahoma. A. Alhamed is supported by a scholarship from the General Organization for the Vocational Training and Technical Education, Saudi Arabia. All the computations related to this project were performed on the ECAS at the University of Oklahoma, Norman.

APPENDIX

Clustering Methodology

a. Hierarchical clustering algorithms

Let s_{ij} be the similarity between objects i and j as defined by one of the similarity measures. Since the similarity is symmetric ($s_{ij} \equiv s_{ji}$), the complete schedule of similarities for all $\binom{n}{2} = 1/2n(n - 1)$ possible pairwise combinations of objects may be arrayed in a lower triangular similarity matrix. Once the similarity matrix is defined, the process of clustering is straightforward. A general procedure for agglomerative clustering based on similarity matrix is as follows (Anderberg 1973):

- Step 1: Place each object in a separate cluster to construct n clusters each of which contains only one object. Use the integer numbers 1 to n to label the clusters.
- Step 2: Find the most similar pair of clusters in the similarity matrix. Let C_i and C_j be the most similar pair with s_{ij} as the similarity between i and j where $i < j$.
- Step 3: Merge the two clusters C_i and C_j and reduce the number of clusters by 1. Label the new cluster C_i and update the similarity matrix to reflect the revised similarities between the new cluster C_i and other existing clusters other than C_j . Delete the row and column of S that corresponds to the cluster C_j .
- Step 4: Repeat step 2 and step 3 a total of $(n - 1)$ times. At each stage record the elements of each cluster and keep track of all similarity measures at each stage to have a complete record.

From this framework, it can be readily noticed that different clustering methods can be implemented by varying the procedure used for defining the most similar pair at step 2 and for updating the revised similarity matrix at step 3. Defining the most similar pair depends on whether the similarity measure is a similarity or dissimilarity coefficient. If a similarity coefficient is used, then the largest value in the similarity matrix indicates the most similar pair. On the contrary, if a dissimilarity coefficient is used, then the smallest value in the dissimilarity matrix indicates the most similar pair.

Updating the similarity matrix depends on the specific clustering method used. Let C_i and C_j be the most similar clusters, so they will be merged into a new cluster,

say C_t ; that is, $C_t = C_i \cup C_j$. One of the most well-known hierarchical clustering methods that can be used is the average linkage, which for the Euclidean distance measure is defined as

$$\frac{1}{t} \frac{1}{r} \sum_{i=1}^t \sum_{j=1}^r e_{ij}, \tag{A.1}$$

where t and r are the number of points in cluster C_t and C_r , respectively. The average linkage method is also called the unweighted pair-group method using arithmetic average (UPGMA). Another popular hierarchical clustering method, called Ward's method, is defined based on the notion of square error. In Ward's method, at each clustering step, the value of the sum-of-square error is computed for every possible merger of two clusters. The merger that produces the smallest increase in the value of sum-of-square error is taken to be the clustering of this step. This process is repeated in each step until one large cluster, which contains all objects, is obtained. Initially, each cluster contains only one object; hence, the value of the sum-of-square error at the beginning is zero.

b. Nonhierarchical or partitional cluster analysis (nHCA)

The problem in nonhierarchical cluster analysis is to partition \mathbf{X} into k clusters, such that $\mathbf{X} = C_1 \cup C_2 \cup \dots \cup C_k$ and $C_i \cap C_j = \phi$, the null set, where $1 \leq i, j \leq k$, and $i \neq j$. Let $|C_i| = n_i$, and $\sum_{i=1}^k n_i = n$, where n_i is the number of object in the i th cluster. Let m_i be the mean vector or center of cluster C_i that is defined as the centroid of cluster C_i , such that

$$m_i = \frac{1}{n_i} \sum_{x_{*j} \in C_i} x_{*j}, \tag{A.2}$$

where x_{*j} is the j th object belonging to cluster C_i .

The basic idea of partitional clustering method is to choose some initial partition of the set of objects and a criterion. The criterion is then evaluated for all possible partitions containing k clusters, where k is specified a priori. The memberships of clusters are altered if needed. Finally, pick the partition that optimizes the criterion.

The first difficulty encountered is the selection of the initial partition of the objects into clusters. The second difficulty is the selection of a criterion. Criteria are highly dependent on the problem parameters and must be simple for computational reasons, but also must be complex enough to reflect the various data structures (Jain and Dubes 1988). An initial partition can be formed by first specifying a set of k seed points. These seed points are used as cluster nuclei around which the set of n objects are grouped (Anderberg 1973). The k seed points can be the first k objects in the dataset, or they can be chosen subjectively or even randomly. Another approach takes any desired partition of the objects into k mutually exclusive clusters and computes the cluster

centroids as seed points. Two commonly used nonhierarchical algorithms are k -mean and nucleated agglomerative (NA).

k-mean algorithm: The most used partitional clustering algorithm is known as the k -mean algorithm, which is described following Anderberg (1973).

- Step 1: Begin with an initial partition of the objects into k clusters for a chosen value of k .
- Step 2: Take each object in sequence and compute the distances to all cluster centroids. If the nearest centroid is not that of an object's parent cluster, then reassign the object and update the centroids of the losing and gaining clusters.
- Step 3: Repeat step 2 until convergence is achieved; that is, continue until a full cycle through the objects fails to cause any change in cluster membership.

The idea of this algorithm is to find a clustering of the objects so as to minimize the total within-cluster sums of squares. The algorithm sequentially processes each object and reassigns it to another cluster if doing so results in minimizing the total within-cluster sums of the squares.

Nucleated agglomerative algorithm: This algorithm provides a convenient way to examine a range of partitions, instead of only one partition as in k -mean. It is implemented in agglomerative fashion, which generates a range of partitions starting from a given maximum number of initial clusters to a given minimum number of clusters. The envisioned final number of clusters should lie in this range (Mather 1976). This algorithm operates into two phases. In the first phase, it generates the clustering structure corresponding to the given maximum number of clusters. In the second phase, it iteratively merges a pair of clusters until a clustering structure corresponding to the given minimum number of clusters is obtained (Mather 1976).

● Phase I

Specify the maximum (k_{max}) and the minimum (k_{min}) number of clusters. Set $k = k_{max}$, and obtain k clusters by applying the k -mean algorithm described above.

● Phase II

Step 1: Reduce k by 1, initially $k = k_{max}$

Step 2: Merge the pair of clusters that minimizes the increase in the square deviation of the objects from their cluster centroids, which is defined as

$$\frac{n_p n_q}{n_p + n_q} \sum (m_p - m_q)^2,$$

$$p = 1, \dots, k - 1;$$

$$q = p + 1, \dots, k. \tag{A.3}$$

This formula is computed for all pair of clusters and the pair (p, q) of the minimum are

merged into one cluster. The centroid of the new cluster is computed.

Step 3: Repeat step 1 and step 2 until $k = k_{min}$.

c. PCA-based clustering

The basic equation for the PCA can be defined from the $m \times n$ data matrix $\mathbf{X} = [x_{ij}]$, where $1 \leq i \leq m, 1 \leq j \leq n, i$ is the observation and j is the variable (e.g., ensemble member). Define

$$\mathbf{X} = \mathbf{FA}^T, \tag{A.4}$$

where $\mathbf{A} = [a_{ij}; i = 1, \dots, n; j = 1, \dots, r]$ is the principal component loading, which denotes the new basis for representation, $\mathbf{F} = [f_{ij}; i = 1, \dots, m; j = 1, \dots, r]$ is the uncorrelated representation of \mathbf{X} called principal component scores, and r is the number of principal components retained where $1 \leq r \leq n$. Each PC loading, a_{*j} , is obtained by scaling the j th eigenvector, v_{*j} , of the correlation matrix by the square root of the j th eigenvalue, λ_j . This yields (Jolliffe 1986)

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}^{1/2}. \tag{A.5}$$

The data matrix \mathbf{X} in (A.4) is usually normalized using (2.3) so that the data have zero mean and unit variance.

Clustering using PCA is achieved by taking advantage of the wealth of information embedded in the principal component loading, perhaps after rotating these loadings using an analytical rotation algorithm. The rotated solution is called a simple structure. Gong and Richman (1995, p. 904) consider the rotated PCA (rPCA) to be a cluster analysis. They consider simple structure rotation as the stage that distinguishes the PCA from a purely statistical technique; its aim is to maximize variance by finding greatly related groups of correlation vectors in m space (Gong and Richman 1995). The simple structure aims to find a new frame of reference in which the PCs speared clusters of highly related variables, and to maximize the number of near-zero loadings (in the hyperplane) (Gong and Richman 1995). The rotational algorithm transforms the unrotated principal component loadings matrix \mathbf{A} to rotated matrix (simple structure) \mathbf{B} using a rotation matrix \mathbf{T} :

$$\mathbf{B} = \mathbf{AT}. \tag{A.6}$$

The rotation process is called *orthogonal rotation* if the rotation matrix \mathbf{T} is orthogonal, that is, if \mathbf{T} transforms the unrotated principal component loadings orthogonally into a simple structure. The rotation is called *oblique rotation* if \mathbf{T} does not transform the unrotated loadings under the orthogonally constraint (i.e., \mathbf{T} is not orthogonal).

The degree of resultant simple structure \mathbf{B} is determined using the notion proposed by Richman (1986). He qualifies the simple structure as strong, moderate, weak, and random. In the case where there are many variables in the hyperplane with distinct clusters of var-

ables and few complex variables, the simple structure is strong. A variable is called complex when it loads moderately on all PCs (i.e., is not in the hyperplane of any PC). The simple structure is moderate where there are slightly less variables in the hyperplane than the first case and more complex variables. In the case where there are a good number of complex variables and an indistinct cluster, the simple structure is characterized as weak. The random simple structure is the one with many complex variables and no clusters. These four types that represent the amount of simple structure present in the analysis are schematically illustrated in Richman (1986).

Our clustering algorithm using rotated principal component analysis is as follows:

- Step 1: Calculate correlation matrix **R** using (2.5).
- Step 2: Calculate principal component loading matrix **A** using (A.5).
- Step 3: Rotate r columns of **A** to simple structure **B** using (A.6) under the constraint that **T** is orthogonal (varimax) and without the orthogonality constraint on **T** (promax).
- Step 4: Examine the loadings in the simple structure (matrix **B**). Assign each variable to the PC (or cluster) on which it loads highly and low on the other PC.

Our goal of using rPCA is to obtain no overlapping (hard) clustering. In step 3 of the above algorithm, we rotate a different number of PCs using varimax first (Kaiser 1958; Richman 1986). If strong simple structure results, it is taken as the clustering solution. Otherwise, promax is used (Hendrickson and White 1964; Richman 1986). We pick the simple structure that has the least amount of overlapping.

REFERENCES

- Alhamed, A., and S. Lakshminarayanan, 2000: Clustering methodologies applied to short-term ensemble forecasting. Preprints, *Second Conf. on Artificial Intelligence*, Long Beach, CA, Amer. Meteor. Soc., 49–55.
- Anderberg, M. R., 1973: *Cluster Analysis for Applications*. Academic Press, 359 pp.
- Atger, F., 1999: The skill of ensemble prediction systems. *Mon. Wea. Rev.*, **127**, 1941–1953.
- Betts, A. K., and M. J. Miller, 1986: A new convective adjustment scheme. Part II: Single column tests using GATE wave, BOMEX, and arctic air-mass data sets. *Quart. J. Roy. Meteor. Soc.*, **112**, 693–709.
- Black, T. L., 1994: The new NMC mesoscale Eta Model: Description and forecast examples. *Wea. Forecasting*, **9**, 265–78.
- Buizza, R., and T. N. Palmer, 1995: The singular-vector structure of the atmospheric general circulation. *J. Atmos. Sci.*, **52**, 1434–1456.
- , M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887–2908.
- Burk, S. D., and W. T. Thompson 1989: A vertically nested regional numerical weather prediction model with second-order closure physics. *Mon. Wea. Rev.*, **117**, 2305–2324.
- Dudhia, J., 1993: A nonhydrostatic version of the Penn State–NCAR mesoscale model: Validation tests and simulation of an Atlantic cyclone and cold front. *Mon. Wea. Rev.*, **121**, 1493–1513.
- Elmore, K. L., and M. B. Richman, 2001: Euclidean distance as a similarity metric for principal component analysis. *Mon. Wea. Rev.*, **129**, 540–549.
- Erico, R., and D. P. Baumhefner, 1987: Predictability experiments using a high-resolution limited-area model. *Mon. Wea. Rev.*, **115**, 488–504.
- Evans, R. E., M. S. J. Harrison, R. J. Graham, and K. R. Mylne, 2000: Joint medium-range ensembles from The Met. Office and ECMWF systems. *Mon. Wea. Rev.*, **128**, 3104–3127.
- Fritsch, J. M., J. Hilliker, J. Ross, and R. L. Vislocky, 2000: Model consensus. *Wea. Forecasting*, **15**, 571–582.
- Gong, X., and M. B. Richman, 1995: On the application of cluster analysis to growing season precipitation data in North America east of the Rockies. *J. Climate*, **8**, 897–931.
- Grell, G. A., 1993: Prognostic evaluation of assumptions used by cumulus parameterizations. *Mon. Wea. Rev.*, **121**, 764–787.
- , J. Dudhia, and D. R. Stauffer, 1994: A description of the fifth-generation Penn State/NCAR Mesoscale Model (MM5). NCAR/TN-398+STR, 121 pp. [Available from MMM Division, NCAR, P.O. Box 3000, Boulder, CO 80307.]
- Harrison, M. S. J., T. N. Palmer, D. S. Richardson, and R. Buizza, 1999: Analysis and model dependencies in medium-range ensembles: Two transplant case studies. *Quart. J. Roy. Meteor. Soc.*, **125**, 2487–2516.
- Hendrickson, A. E., and P. O. White, 1964: Promax: A quick method for rotation to oblique simple structure. *Br. J. Stat. Psychol.*, **17**, 65–70.
- Hoffman, R. N., and E. Kalnay, 1983: Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus*, **35A**, 100–118.
- Houtekamer, P. L., and J. Derome, 1995: Methods for ensemble prediction. *Mon. Wea. Rev.*, **123**, 2181–2196.
- Hou, D., E. Kalnay, and K. Drogemeier, 2001: Objective verification of the SAMEX98 ensemble forecasts. *Mon. Wea. Rev.*, **129**, 73–91.
- Jain, A. J., and R. C. Dubes, 1988: *Algorithms for Clustering Data*. Prentice-Hall, 320 pp.
- Jolliffe, I. T., 1986: *Principal Component Analysis*. Springer-Verlag, 271 pp.
- Juang, H.-M., and M. Kanamitsu, 1994: The NMC nested regional spectral model. *Mon. Wea. Rev.*, **122**, 3–26.
- Kain, J. S., and J. M. Fritsch, 1990: A one-dimensional entraining/detraining plume model and its application in convective parameterization. *J. Atmos. Sci.*, **47**, 2784–2802.
- Kaiser, H. F., 1958: The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, **23**, 187–200.
- Mather, P. M., 1976: *Computational Methods of Multivariate Analysis in Physical Geography*. John Wiley and Sons, 532 pp.
- Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliaigis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119.
- Mullen, S. L., and D. P. Baumhefner, 1988: Sensitivity to numerical simulations of explosive oceanic cyclogenesis to changes in physical parameterizations. *Mon. Wea. Rev.*, **116**, 2289–2329.
- Richman, M. B., 1986: Rotation of principal components. *J. Climatol.*, **6**, 293–335.
- Romesburg, C. H., 1984: *Cluster Analysis for Researchers*. Life Time Learning, 334 pp.
- Stensrud, D. J., J.-W. Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensembles simulations of mesoscale convective systems. *Mon. Wea. Rev.*, **128**, 2077–2107.
- Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330.
- , and —, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297–3319.
- Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev.*, **129**, 729–747.

- Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction*. Academic Press, 467 pp.
- Xue, M., K. Droegemeier, V. Wong, and A. Shapiro, 2000: The Advanced Regional Prediction System (ARPS)—A multiscale nonhydrostatic atmospheric simulation and prediction tool. Part I: Model dynamics and verification. *Meteor. Atmos. Phys.*, **75**, 161–193.
- Zhang, D.-L., and R. A. Anthes, 1982: A high-resolution model of the planetary boundary layer—Sensitivity tests and comparisons with SESAME-79 data. *J. Appl. Meteor.*, **21**, 1594–1609.
- Ziehmann, C., 2000: Comparison of single-model EPS with a multi-model ensemble consisting of a few operational models. *Tellus*, **52A**, 280–299.