Development of an Automated Classification Procedure for Rainfall Systems

MICHAEL E. BALDWIN* AND JOHN S. KAIN*

Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma

S. LAKSHMIVARAHAN

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

(Manuscript submitted 17 November 2003, in final form 27 September 2004)

ABSTRACT

An automated procedure for classifying rainfall systems (meso-α scale and larger) was developed using an operational analysis of hourly precipitation estimates from radar and rain gauge data. The development process followed two main phases: a training phase and a testing phase. First, 48 hand-selected cases were used to create a training dataset, from which a set of attributes related to morphological aspects of rainfall systems were extracted. A hierarchy of classes for rainfall systems, in which the systems are separated into general *convective* (heavy rain) and *nonconvective* (light rain) classes, was envisioned. At the next level of classification hierarchy, convective events are divided into *linear* and *cellular* subclasses, and nonconvective events belong to the *stratiform* subclass. Essential attributes of precipitating systems, related to the rainfall intensity and degree of linear organization, were determined during the training phase. The attributes related to the rainfall intensity were chosen to be the parameters of the gamma probability distribution fit to observed rainfall amount frequency distributions using the generalized method of moments. Attributes related to the degree of spatial continuity of each rainfall system were acquired from correlogram analysis. Rainfall systems were categorized using hierarchical cluster analysis experiments with various combinations of these attributes. The combination of attributes that resulted in the best match between cluster analysis results and an expert classification were used as the basis for an automated classification procedure.

The development process shifted into the testing phase, where automated procedures for identifying and classifying rainfall systems were used to analyze every rainfall system in the contiguous 48 states during 2002. To allow for a feasible validation, a testing dataset was extracted from the 2002 data. The testing dataset consisted of 100 randomly selected rainfall systems larger than 40 000 km^2 as identified by an automated identification system. This subset was shown to be representative of the full 2002 dataset. Finally, the automated classification procedure classified the testing dataset into stratiform, linear, and cellular classes with 85% accuracy, as compared to an expert classification.

1. Introduction

Precipitation-producing weather systems appear in many forms. For example, some are clearly linear in shape, whereas others are more circular; some contain scattered regions of intense rainfall, while others produce widespread light precipitation. Many other distinctions could be applied.

It can be very useful to describe rainfall systems using these morphological characteristics. For example, a classification based upon attributes of this kind could be used as the basis for climatological studies of precipitation systems (e.g., Houze et al. 1990). For this study, classification procedures were developed with forecast verification interests in mind. In particular, this study determined ways of describing specific aspects of precipitating systems using statistical techniques. These characteristics were successfully used as the foundation of an automated classification procedure. We anticipate that characteristics that demonstrate the ability to *classify* rainfall systems will also be useful in *comparing* observed and predicted rainfall patterns. Such a comparison is motivated by the desire to validate the "realism" of forecasts, a fundamental aspect that appears to be missing in traditional meteorological verification strategies (Anthes 1983).

Classification is the process of systematically placing objects into categories or classes, based upon the *similarity* of an object to other members of a group. An *object* is the general term representing an individual entity that one wishes to classify. Characteristics that describe the objects are often referred to as attributes.

* Additional affiliation: NOAA/National Severe Storms Laboratory, Norman, Oklahoma.

Corresponding author address: Michael E. Baldwin, OU/CIMMS, 1313 Halley Circle, Norman, OK 73069.
E-mail: mbaldwin@ou.edu

Similarity is the degree of sameness between objects, which is typically determined by some measure of distance or correlation. In particular, one can represent the set of characteristics that describe each object as an attribute vector. An intuitive measure of distance commonly used in classification is the Euclidean distance, which is the well-known L2 or vector norm between attribute vectors. In classification, the process of generating attributes that characterize objects in a useful way is known as feature extraction (Duda et al. 2000). Useful features are those that help the classification procedure place objects into their proper class. The general problem in classification is to determine the class of an object based upon its similarity to the characteristics of known classes or other objects in a set of data. Members of the same class should be more similar to each other than they are to objects belonging to other classes. This is the guiding principle for establishing a set of *classification rules* that assign objects to classes. The classification rules can be developed first from a training dataset and then can be used for assigning classes to other objects previously unseen. Cluster analysis, a general name for a variety of mathematical methods that can be used to determine which objects in a set are similar, is often used as a classification tool (Romesburg 1984).

Several classes of rainfall systems have previously been defined; some based upon the underlying physical processes that produced the rainfall, such as the general classes of *convective* and *stratiform* (Houghton 1968). Other general classes were based upon the space and time scales associated with each system, such as *synoptic* and *mesoscale* (Austin and Houze 1972; Orlanski 1975). Previous automated rainfall classification techniques have focused on segmenting rainfall systems and can be characterized as "microclassification" approaches. For example, several methods of identifying and tracking individual thunderstorms have been developed for short-term forecasting purposes or for use in weather-related decision support systems (e.g., Kessler 1966; Wilson et al. 1998; Lakshmanan 2001). Other researchers have developed automated rainfall classification procedures in order to estimate vertical latent heating profiles or improve rainfall estimation (e.g., Steiner et al. 1995; Yuter and Houze 1997; Biggerstaff and Listemaa 2000). These classification schemes subdivide a rainfall system into convective and stratiform segments on a pixel-by-pixel basis.

In contrast, this work takes a "macroclassification" approach to classify rainfall systems in their entirety, by using statistically based measures of the morphological characteristics of each system. In a typical mesoscale convective system (MCS), the convective and stratiform regions are interrelated and interdependent parts of a system. For example, the stratiform region would not exist if the convection had not transported ice crystals away from the convective updrafts (Gamache and Houze 1983). In some cases, evaporation of rainfall within the stratiform region helps to enhance the mesoscale circulation that allows the convection to propagate (e.g., Zhang and Gao 1989). Therefore, in this work an MCS was considered a "convective" entity and was not subdivided into convective/stratiform regions.

Previous comprehensive studies of MCSs were also performed using a "macroclassification" approach (e.g., Bluestein and Jain 1985; Bluestein et al. 1987; Blanchard 1990; Houze et al. 1990; Geerts 1998; Parker and Johnson 2000; Jirak et al. 2003). These studies examined MCSs primarily using visual analysis of radar images. Rainfall systems were classified based upon how they developed over time and how closely they matched archetypical examples. The common characteristic among these studies was the use of visual inspection of the radar images as the primary analysis tool. A primary goal of this work is to develop an automated classification procedure that can be applied in place of the subjective analysis techniques previously used. A previous example of such an automated classification procedure that utilized infrared satellite imagery (Evans and Shemo 1996) was recently used to document the diurnal cycle of organized convection over the global Tropics and midlatitudes (Tsakraklides and Evans 2003). The current work is unique since rainfall estimates from radar and rain gauge data are utilized, rather than the satellite data used in the Evans and Shemo (1996) procedure.

Ideally, one would perform a classification of rainfall systems based upon observations of the physical processes occurring within each system. However, given the current capabilities of observational platforms, one can only routinely observe the shape and structure of an object and cannot observe the details that allow for understanding of the function of every system. Although in some situations, it may be possible to determine relationships between statistical measures of structure and the physical processes operating within a system (e.g., Perica and Foufoula-Georgiou 1996). In this case, such determination is left for future work.

The following general hierarchy of classes for rainfall systems was used in this work. The first branch of the classification separates rainfall systems into *convective* and *nonconvective* classes. Precipitation systems belonging to the convective class are those that contain relatively high rainfall rates. Presumably, the high precipitation rates are associated with circulations containing upward vertical motions larger than the fall speed of precipitation particles (e.g., Houghton 1968; Houze 1997). The large upward vertical motions result in rapid growth of raindrops by collection of cloud water in the updraft (Houghton 1968). On the other hand, precipitation systems belonging to the nonconvective class are those that produce relatively low precipitation rates. The low precipitation rates are generated by much weaker upward vertical motions, such as the widespread lifting associated with large-scale warm advection. As described by Yuter and Houze (1997), "the fall

speeds of the precipitating ice particles (~1–2 m s^{-1}; Locatelli and Hobbs 1974) far exceed the magnitude of the vertical air motion" for members of this class. The main precipitation particle growth process in this case is diffusion (Houghton 1968), a slow process. Since routine observations of vertical motion are not available, measurements of precipitation accumulation were used in this work to distinguish between convective and nonconvective rainfall systems. Specifically, systems producing higher precipitation rates were classified as convective and others were considered stratiform.

The next branch of the classification hierarchy separates convective systems based upon the structure of the rainfall pattern, creating *linear* and *cellular* subclasses. In the linear class, the convective precipitation exhibits a degree of elongation or linear organization, similar to the *elongated* classes of Anderson and Arritt (1998) and Jirak et al. (2003). A rainfall system belonging to the cellular class is characterized by convective precipitation elements somewhat randomly positioned throughout the system, akin to the *unclassifiable* class of Houze et al. (1990) and the *chaotic* class of Blanchard (1990). In this work, all nonconvective systems belong to the *stratiform* subclass. Admittedly, the proposed classification terminology is somewhat ambiguous. For instance, many systems classified as linear will contain convective cells and a region of stratiform rain. However, this general classification is designed to be consistent with further refinement in the future. For example, subclasses of cellular systems that indicate the degree of discreteness of the convective precipitation could be created, such as *isolated cells* and *multicell clusters*. Since *cell* is a common term among these related subclasses, cellular was chosen as the parent class designation.

The process used in this work to develop an automated classification procedure included collection of a dataset for training and testing, creation of methods of identifying rainfall systems or objects, "training" the classification procedure by determining a set of features that describe the objects in a useful manner, creation of a classification procedure by establishing rules for partitioning objects into the classification hierarchy, and testing the classification procedure. In this work, a national mosaic of high-resolution rainfall data, known as the "stage IV" analysis (Baldwin and Mitchell 1998), was used to create a dataset for the classification development process. Experiments were performed on a training dataset to find a set of characteristics that produced results from a classification using a cluster analysis algorithm that was in good agreement with a classification performed by a human. These characteristics were used as the basis of an automated classification, which was validated using an independent testing dataset. This evaluation showed that the classification procedure correctly placed 85% of the objects into their linear, cellular, and stratiform classes. It should be noted that this procedure was developed and tested using relatively large-scale rainfall systems (meso-α and larger). While this work represents the initial steps toward the ultimate goal of creating a well-refined, completely automated classification system, these results show that the procedures demonstrated here can be used as a foundation for continued development and refinement.

A description of the training phase of the process of developing an automated classification procedure is presented in section 2. This includes descriptions of precipitation datasets, objective criteria used in expert classification, feature extraction methods, and results of cluster analysis experiments. The testing phase of this development process is documented in section 3. This includes discussion of the automated procedures for object identification and classification, summary statistics obtained from analysis of an entire year of rainfall data, and an independent evaluation of the classification procedure. Concluding remarks and a discussion of future work are provided in the final section.

2. Training phase

The objective of this work is to develop an automated classification procedure using statistically based attributes that characterize various morphological aspects of rainfall systems. Such characteristics may eventually be used to compare predicted and observed rainfall systems in a forecast verification system. However, it is not clear a priori which statistical characteristics of rainfall systems will result in an automated classification that agrees with a meteorologist's expert opinion. This section describes the training phase in the process of developing the automated classification procedure. To discover a set of essential features that allow proper classification, a training dataset was used. Each rainfall analysis is treated as a snapshot, and the rainfall systems are analyzed without regard to their temporal evolution. A classification was first performed manually, by an expert in meteorological analysis. Trial attributes that characterize several morphological aspects of rainfall systems were then obtained. Experimental classifications were performed using cluster analysis with a variety of combinations of the trial attributes. The combination of trial attributes in the cluster analysis experiments that resulted in the best agreement with an expert classification became the basis for an automated classification procedure. Once the training phase was complete, the development process moved to the testing phase, which will be described in section 3.

a. Precipitation data

To begin this work, a dataset used in the training and testing phases was established. The so-called stage IV rainfall analysis (Fulton et al. 1998; Seo 1998; Baldwin and Mitchell 1998) produced at the National Centers for Environmental Prediction (NCEP) was obtained. The stage IV analysis is a national mosaic of optimal

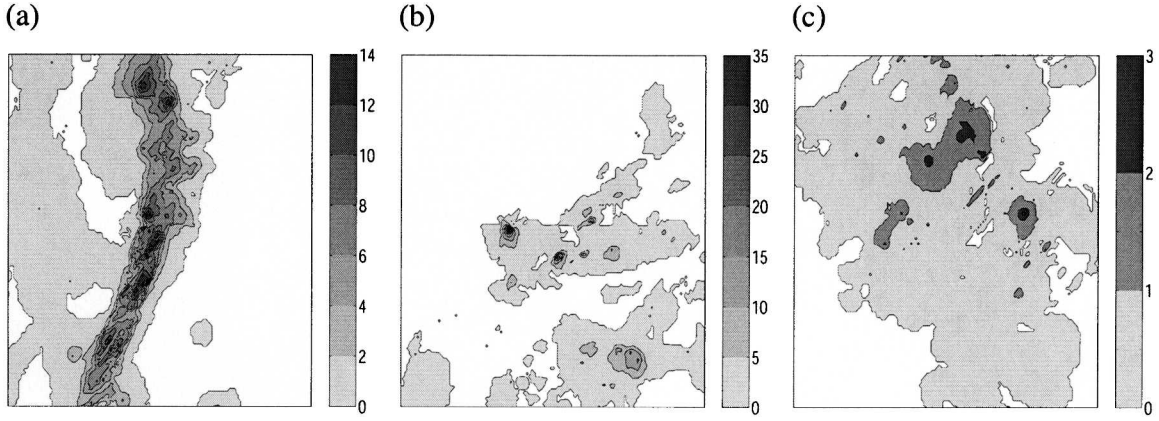

FIG. 1. Cases (a) 6, (b) 30, and (c) 41 of the training dataset. Grayscale on the side of each image indicates rainfall amounts in mm. Contour intervals are (a) 2, (b) 5, and (c) 1 mm.

estimates of 1-h accumulated rainfall using radar and rain gauge data, which are available at the top of every hour on a 4 km × 4 km mesh covering the lower 48 states. The analysis includes a mean radar bias correction, separate radar-only and gauge-only analysis mosaics, and a "multisensor" analysis combining the radar and gauge estimates using an optimal estimation technique (Seo 1998). The multisensor mosaic was used in this work.

One hundred forty-eight separate precipitation events occurring at different times and locations across the United States were selected for inclusion in the overall precipitation dataset. This dataset was divided into a training dataset, consisting of 48 cases, and a testing dataset containing 100 events. Briefly, the testing dataset consisted of randomly selected rainfall systems observed during 2002 that were larger than 40 000 km^2 as identified by an automated identification system. Details of the testing dataset are left to section 3. However, for the training dataset, events were selected by hand, and the event selection criteria were based upon the occurrence of typical rainfall patterns that are often found across the United States during the year. The late summer–fall 2000 time period was selected because of data availability and because this period typically represents a transition from warm-season convective to cool-season stratiform precipitation regimes. For the training dataset, the size of the domain was chosen to be fixed at 128 × 128 4-km grid boxes, which is approximately 500 km by 500 km. For each case in the training set, the domain was centered visually near the event of interest. This method of selecting rainfall systems for the training dataset obviously cannot be used as an automated identification procedure. The automated rainfall object identification procedure that was developed to analyze rainfall events beyond the training dataset in this work is described in section 3a. The purpose of the training dataset was to test the usefulness of various trial attributes for eventual use as the basis for an automated classification procedure. The number of cases contained within the training dataset (48) may seem small, given the wide spectrum of rainfall systems occurring in nature. However, the positive results obtained in the testing phase (section 3) demonstrate that the training dataset was large enough to sufficiently sample the natural variability of rainfall systems, at least for the general classification hierarchy used in this work. Certainly, more cases will be required to further refine the classification procedure as the development process is revisited in future work.

b. Expert classification

Expert classification of the rainfall systems in the precipitation dataset was based entirely upon the 1-h accumulated rainfall pattern; no other information (such as meteorological conditions, temporal continuity, location, time of year, time of day) associated with the events was provided. Objective criteria were used in the expert classification. Convective events were defined as those where a substantial fraction ($> \sim 5\%$) of the rainfall system received rain rates of 5 mm hr^{-1} or higher. Otherwise, the system was classified as stratiform. The 5 mm hr^{-1} threshold is similar to other rain rate and radar reflectivity thresholds that have been used in previous research to delineate convective and stratiform regions. For example, Johnson and Hamilton (1988) used a 6 mm hr^{-1} rain rate threshold, while Geerts (1998) used the 20-dBZ threshold to delineate the convective region as long as there was a maximum reflectivity of at least 40 dBZ embedded within it. Examples of convective events from the training dataset are given in Figs. 1a and 1b, which show that a significant fraction of each system received rainfall greater than 5 mm in an hour. An example of a stratiform (nonconvective) event is provided in Fig. 1c, where widespread light precipitation was observed. Within convective events, the region of heavier rainfall was surrounded by a rectangular bounding box, rotated to be parallel to the primary axis of the convective rainfall. If the aspect ratio of such a rectangle was 3 or greater,

TABLE 1. Location, time, date, expert classification, gamma parameters, area, and eccentricity of 0.6 correlogram contour for the 48 cases of the training dataset.

Case	Central location (°N, °W)	Date, time (UTC)	Expert classification	Gamma-shape parameter	Gamma-scale parameter	Eccentricity 0.6 contour	Area 0.6 contour
1	42.8, 94.9	17 Aug 2000, 0500	Linear	0.48	6.46	5.00	65.0
2	39.0, 90.0	05 Oct 2000, 1100	Linear	0.62	2.11	2.58	74.8
3	42.9, 123.7	28 Oct 2000, 1100	Linear	0.63	1.47	4.04	202.2
4	36.6, 99.4	25 Oct 2000, 0200	Linear	0.42	4.52	2.48	32.2
5	39.2, 97.3	29 Oct 2000, 0700	Linear	0.48	1.28	4.14	165.5
6	39.5, 82.5	21 Sep 2000, 0200	Linear	0.30	3.16	5.06	126.5
7	39.5, 82.5	21 Sep 2000, 0300	Linear	0.26	2.97	5.16	103.2
8	37.0, 102.0	01 Nov 2000, 0100	Linear	0.44	3.57	5.06	40.5
9	38.4, 97.2	22 Sep 2000, 2300	Linear	0.37	7.54	2.76	22.1
10	40.0, 83.7	21 Sep 2000, 0100	Linear	0.49	1.92	2.83	56.6
11	35.5, 78.5	25 Sep 2000, 2300	Linear	0.44	3.60	2.49	72.2
12	39.9, 86.3	20 Sep 2000, 2100	Linear	0.39	3.14	2.77	47.0
13	40.0, 85.7	20 Sep 2000, 2200	Linear	0.42	2.73	3.10	77.6
14	40.1, 85.1	20 Sep 2000, 2300	Linear	0.52	2.75	2.62	68.0
15	34.8, 97.2	01 Nov 2000, 1800	Linear	0.24	6.63	8.68	43.4
16	35.3, 87.6	09 Nov 2000, 0400	Linear	0.52	5.03	3.04	109.5
17	40.2, 84.5	21 Sep 2000, 0000	Linear	0.52	2.17	2.30	39.1
18	35.5, 85.2	25 Sep 2000, 0900	Linear	0.16	8.13	3.40	61.2
19	38.8, 90.8	25 Sep 2000, 1000	Cellular	0.59	2.44	2.62	89.2
20	40.0, 86.0	04 Oct 2000, 2200	Cellular	0.33	4.32	3.07	27.7
21	36.9, 97.7	25 Oct 2000, 1600	Cellular	0.67	3.13	2.70	54.0
22	39.1, 104.2	17 Aug 2000, 2200	Cellular	0.19	4.31	2.12	4.2
23	41.4, 92.8	04 Oct 2000, 0200	Cellular	0.40	2.72	2.09	35.5
24	31.2, 101.6	17 Oct 2000, 1300	Cellular	0.38	7.37	1.86	24.2
25	40.0, 86.0	04 Oct 2000, 2300	Cellular	0.52	3.45	2.16	69.1
26	40.0, 86.0	05 Oct 2000, 0000	Cellular	0.46	3.49	1.50	60.0
27	40.0, 86.0	05 Oct 2000, 0100	Cellular	0.52	3.06	1.30	52.2
28	40.0, 86.0	05 Oct 2000, 0200	Cellular	0.51	2.65	1.49	59.7
29	38.8, 90.8	25 Sep 2000, 1100	Cellular	0.54	2.33	1.84	222.5
30	32.4, 93.0	05 Oct 2000, 2300	Cellular	0.17	5.40	1.58	6.3
31	32.3, 110.0	17 Aug 2000, 2300	Cellular	0.14	11.94	1.00	1.0
32	31.8, 85.8	22 Sep 2000, 1100	Cellular	0.31	5.05	2.69	10.8
33	39.3, 88.9	25 Sep 2000, 12000	Cellular	0.57	2.01	1.62	110.0
34	30.0, 99.6	02 Nov 2000, 2100	Cellular	0.22	6.46	3.61	3.6
35	35.0, 95.5	16 Oct 2000, 0300	Cellular	0.49	3.18	1.56	62.3
36	41.89, 86.1	17 Aug 2000, 1400	Cellular	0.52	2.61	1.72	112.0
37	38.5, 83.2	17 Aug 2000, 1800	Cellular	0.30	5.14	1.98	49.5
38	44.6, 123.3	10 Oct 2000, 0200	Stratiform	0.50	1.16	1.26	25.3
39	45.1, 123.1	01 Oct 2000, 0300	Stratiform	0.52	0.92	1.63	32.6
40	36.1, 118.7	10 Oct 2000, 0600	Stratiform	0.51	0.62	1.63	55.3
41	38.7, 94.5	25 Sep 2000, 0000	Stratiform	0.76	0.46	1.33	54.3
42	41.2, 76.3	26 Sep 2000, 1500	Stratiform	0.59	0.57	1.52	38.1
43	44.6, 123.3	10 Oct 2000, 0000	Stratiform	0.62	0.62	2.11	27.5
44	34.6, 92.1	04 Nov 2000, 0900	Stratiform	0.73	0.71	2.24	8.9
45	42.3, 93.6	06 Nov 2000, 1900	Stratiform	1.65	0.70	1.42	103.9
46	34.2, 97.4	08 Nov 2000, 1000	Stratiform	0.59	0.94	1.70	34.1
47	44.6, 123.3	10 Oct 2000, 0100	Stratiform	0.67	0.62	1.52	38.1
48	30.3, 97.7	18 Nov 2000, 1600	Stratiform	1.11	0.94	1.30	65.2

the system was classified as linear (e.g., Fig. 1a), otherwise it was considered cellular (e.g., Fig. 1b). The 3 to 1 ratio was relaxed from the 5 to 1 ratio used by Bluestein and Jain (1985) to allow for the motion of rainfall systems since the current work used precipitation accumulated over 1 h. Table 1 provides detailed information for the training dataset along with the results of the expert classification and other attributes that will be described in section 2d. Note that the events were not uniformly distributed, as the majority of the events in the training dataset belonged to the convective class. A larger sample of such events was necessary, since the characteristics of convective events vary much more than those associated with stratiform systems. Results of expert classification of the testing dataset are discussed in section 3.

c. Object identification

Rainfall systems are defined as contiguous areas of precipitation, similar to Ebert and McBride (2000). For each of the 48 events found within the training dataset, the pattern of nonzero rainfall from the entire 500 km by 500 km domain plus a surrounding "trace region" of approximately 15% of the area of measurable rainfall

was considered an object for classification. Details of the automated object identification procedure used for the testing dataset are left to section 3. An estimation of the number of rainfall values below the instrument detection limit ("trace") was necessary for the estimation of various statistical parameters. Hosking and Stow (1987) studied high-resolution measurements of rainfall and found that 30%–40% of nonzero rain periods produced rainfall accumulations less than the resolvable limit by conventional recording rain gauges. Spatially, it is reasonable to expect that the size of the area receiving "trace" amounts of precipitation to be some fraction of the total area receiving detectable precipitation. The fraction used in this work was determined by experiment. A gamma probability density function was fit to the frequency distribution of rainfall for each of the 48 cases in the training dataset for a variety of "trace fraction" values using the maximum likelihood estimation method of Wilks (1990). The median of the 48 trace fractions that resulted in the best fit (in terms of mean absolute error) of the gamma distribution to the observed histograms in the dataset was 18%. As an approximation to this, the size of each object was increased by roughly 15% by extending the edge of the measurable precipitation region by a constant number of analysis grid boxes in each direction, figured by assuming a circular area of precipitation.

d. Attribute selection

Numerous attributes were investigated for their ability to allow the classification system to arrange rainfall systems into their proper classes. Several of the attributes that were found to be useful are described below. Given the general nature of the classification hierarchy used in this work, many of these attributes are invariant to translation, rotation, locations of maxima, and spatial scale. If one wishes to design a more refined classification procedure, additional attributes that allow for proper discrimination of objects into the desired classification hierarchy must be included.

1) INTENSITY-RELATED ATTRIBUTES

One might expect that parameters that generally describe the distribution of rainfall amounts within a precipitation system would be useful attributes for classification. Typically, the histogram, or frequency distribution, of precipitation amounts within a rainfall system is positively skewed. For example, in heavy rain events, the majority of the precipitation area receives light to moderate amounts of precipitation, but a substantial fraction receives heavier amounts, resulting in a distribution that possesses a long "tail." On the other hand, widespread light rainfall produces a distribution that is "humped" near a light amount of rainfall with little or no "tail." One way to compactly describe the observed histogram is to fit a theoretical probability distribution to it and use the distribution parameters as attributes. For this work, the gamma distribution was selected since it is positively skewed and nonnegative, provides a reasonable representation with only two parameters (α, β), and has often been used for the analysis of precipitation data (e.g., Wilks 1990). The gamma probability density function is (Wilks 1995)

$$f(x; \alpha, \beta) = (x/\beta)^{\alpha-1}[\exp(-x/\beta)][\beta\Gamma(\alpha)]^{-1}, \quad x \geq 0, \alpha, \beta > 0, \tag{1}$$

where $\Gamma(\alpha)$ is the standard gamma function.

The α parameter is commonly referred to as the *shape* parameter since it mainly affects the shape of the distribution function. Figure 2a shows two example gamma probability density function curves for varying shape parameter values. For small values of the shape parameter ($\alpha < 1$), the distribution is skewed strongly to the right with $f(x) \rightarrow \infty$ as x approaches zero. For values of $\alpha > 1$ the distribution function begins at the origin and reaches a maximum value at $x = \beta\,(\alpha - 1)$. For very large values of the shape parameter, the gamma distribution is similar to the Gaussian distribution. The role of the parameter β, known as the *scale* parameter (Fig. 2b), is mainly to affect the tail of the distribution. For larger scale parameter values, the distribution is "pulled" to the right, representing an increased frequency of larger values of x and creating a thicker tail. For smaller values of the scale parameter, the frequency of smaller values of x is increased, creating a thinner tail and "pushing" the distribution toward the left. Because the shape and scale parameters of the gamma distribution briefly describe the array of rainfall amounts within a system, they were selected as trial attributes in the training phase in the development of the automated classification procedure. Since the frequency distribution does not contain information on the location of rainfall amounts, these attributes are invariant to rainfall system translation and rotation.

Rainfall data, like most meteorological variables, are spatially correlated. For this reason, a robust method of parameter estimation for the theoretical distribution that does not rely upon an assumption of independence is desired. One such parameter estimation technique is known as the generalized method of moments (GMM; Hansen 1982; Hamilton 1994). GMM can be formulated to allow correlation in the data to affect the parameter estimation. GMM could be considered an extension to the more familiar method of moments for parameter estimation. In the method of moments, a set of equations is developed to cover the number of unknown parameters found in the theoretical distribution. In the case of the gamma distribution, the shape and scale parameters are the two unknowns; therefore two equations relating these to known quantities are needed. For example, these two equations could be determined by equating the first two sample moments to the population moments. Given this equation set, parameters obtained via the method of moments tech-

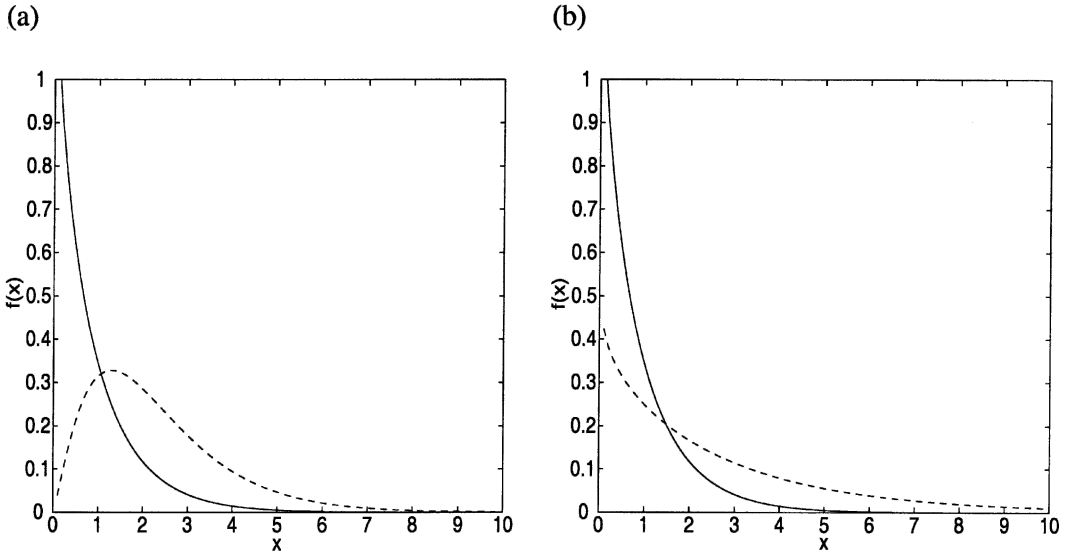

FIG. 2. Plots of the gamma probability density function for (a) $\alpha = 0.9$, $\beta = 1.0$ (solid) and $\alpha = 2.3$, $\beta = 1.0$ (dashed), and (b) $\alpha = 0.9$, $\beta = 1.0$ (solid) and $\alpha = 0.9$, $\beta = 3.0$ (dashed).

nique would fit the observed mean and variance exactly, but higher-order moments would not be taken into account. In some cases, it may be desirable for the parameters to provide a better fit to higher-order moments, such as the observed skewness or kurtosis. The GMM technique allows for this by adding higher-order moments to the equation set, resulting in a nonlinear system of equations that can then be solved by least squares methods. A detailed description of GMM is provided in the appendix.

Selection of the number of moments and degree of serial correlation in the data to use in the GMM estimates was based upon previous work (Baldwin and Lakshmivarahan 2002). The selection criteria was based upon the quality of classification experiments using the GMM estimated shape and scale parameters as attributes. Several different combinations of number of moments (2 to 4) and values of lag correlation in the data ($q = 0$ to 5) were tested. Baldwin and Lakshmivarahan (2002) showed that the three-moment (first, second, and third moments) GMM estimates produced the best separation of rainfall events into general convective and stratiform classes. In general, these results found that convective events were typically associated with relatively low values of the shape parameter and high values of the scale parameter, while stratiform events had relatively high values of the shape parameter and low values of scale. The results were not sensitive to the choice of lag-correlation value; therefore $q = 1$ was chosen to account for serial correlation in the data in this work. The results of the GMM estimates of the gamma-shape and -scale parameters for the training dataset are summarized in Table 1. For example, the shape parameter for the linear convective event shown in Fig. 1a was 0.30 and the scale parameter was 3.16. For the event classified as stratiform shown in Fig. 1c, the gamma-shape parameter was 0.76 and the scale parameter was 0.46.

2) SPATIAL-CONTINUITY-RELATED ATTRIBUTES

The intensity-related attributes cannot provide information on the spatial continuity and variability of the rainfall within an object. For instance, identical histograms can be obtained from events that are randomly unorganized or spatially continuous, since the frequency distribution ignores information on the *location* of rainfall amounts. To provide information on aspects of the spatial continuity and variability within rainfall objects, additional attributes related to the shape and structure of the spatial patterns are required. There is a long history of research using geostatistical tools to examine the characteristics of spatial radar/rainfall data (e.g., Kessler 1966; Zawadzki 1973). For example, Kessler and Russo (1963) noted how the ellipticity of the autocorrelation was an objective measure of the "systematic bandedness in the pattern" and how the orientation of the major axis reflected the orientation of the reflectivity bands. There are several measures of spatial variability and continuity to choose from (Isaaks and Srivastava 1989; Deutsch and Journel 1998); in this work two-dimensional plots of the autocorrelation, known as the correlogram, are used. The correlogram displays the correlation between all possible pairs of data values separated by a given lag vector, plotted as a function of lag in both the x and y directions.

For each event in the precipitation dataset, a correlogram is computed using the Geostatistical Software Library (GSLIB), a freely available package of geostatistical algorithms (Deutsch and Journel 1998). Inspection of these plots in the training dataset revealed

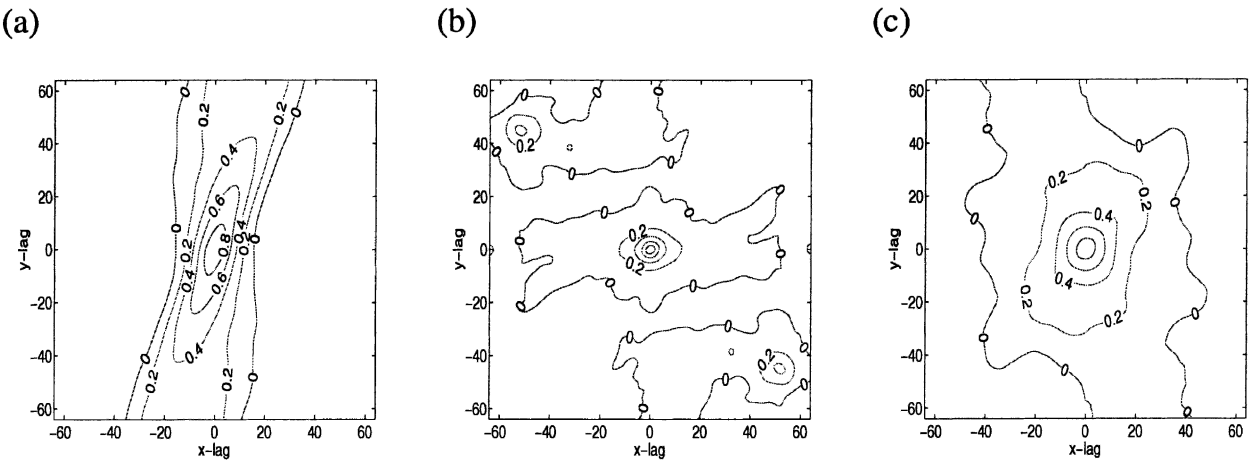

FIG. 3. Correlogram plots corresponding to rainfall case (a) 6, (b) 30, and (c) 41 of the training dataset. Contour interval is 0.2. Lags indicated are numbers of 4-km grid boxes from the original analysis (see Fig. 1).

that linear rainfall patterns were associated with elliptical correlation contours (Fig. 3a), while cellular and stratiform precipitation patterns produced circular correlation contours (Figs. 3b and 3c). Therefore, one might expect that summary measures of the correlation contour ellipticity would provide useful attributes for the automated classification system.

The *area* and *eccentricity* of various contour values in each correlogram from the training dataset were approximated and used as trial attributes in the training phase of the classification procedure development. Using image processing techniques, a correlation contour was considered an object. All contiguous grid points with correlation greater than the contour value that also include the origin were given the same object label using a connected component labeling algorithm (Klette and Zamperoni 1996). Next, the edge of this connected region was found using a binary edge detection algorithm (Davies 1997). These processes equated to locating the specified contour surrounding the origin on the correlogram. Contours related to secondary maxima (centered away from the origin) were not analyzed by this procedure. Once this was established, the largest distance from the origin to this edge was found, and this distance was assumed to be the length of the semimajor axis (a) of the contour object. The shortest distance from the origin to the edge was found next, and this was assumed to be the length of the semiminor axis (b). Note that these will not necessarily be orthogonal. The ratio of the semimajor and semiminor axes (a/b) was used as an approximate measure of the eccentricity of the contour. For a circular contour, this ratio will be equal to 1.0; the ratio will increase as the contour becomes more elongated. The product of the two axis lengths (ab) was also used as an approximate measure of the area covered by the contour. These attributes are also invariant to rainfall system translation and rotation. The results of this analysis of the training dataset are summarized in Table 1 for the 0.6 correlation contour. For example, the eccentricity parameter for the 0.6 correlation contour for the linear convective event shown in Fig. 1a was 5.06 and the area parameter was 126.5. For the event classified as cellular shown in Fig. 1b, the eccentricity parameter for the same contour was 1.58 and the area parameter was 6.3. These results show that the cases classified as linear tended to have higher eccentricity parameters than those classified as cellular or stratiform.

e. Cluster analysis

In this work, experimental classification was performed using a hierarchical cluster analysis algorithm, specifically Ward's (1963) method. Ward's method is an *agglomerative* clustering technique, where clusters are grouped together to include increasing numbers of objects in a stepwise fashion. At the lowest level of the clustering hierarchy, each cluster contains a single object. Larger clusters are formed by merging the two clusters that produce the smallest increase in the within-cluster variance. This merger step is repeated until a single cluster containing all objects is created. Ward's method has been found to produce satisfactory results for meteorological data in previous research (Alhamed et al. 2002).

Hierarchical cluster analysis provides information on the relationship between objects and clusters in the dataset, which is typically represented by a tree diagram or dendrogram. A hypothetical example is provided in Fig. 4. In this case, the y axis indicates decreasing levels of similarity, and "branches" on the dendrogram indicate the level of similarity where objects are grouped into clusters. Using cluster analysis for classification requires several subjective decisions: choice of cluster analysis algorithm, number of clusters, etc. Ideally, objects will be grouped into clusters at a high level of similarity, and a relatively small number of clusters

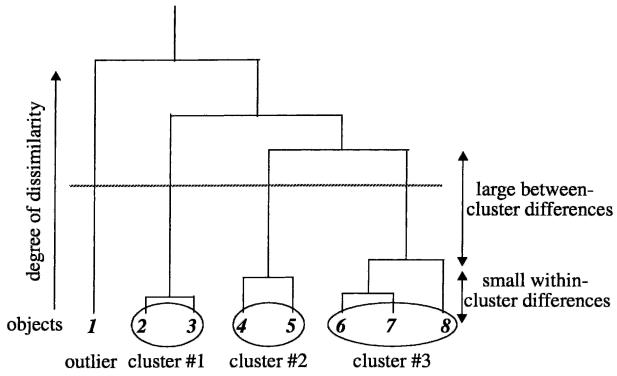

FIG. 4. Hypothetical hierarchical clustering dendrogram, indicating ideal clustering. Ideally, the dissimilarity between objects within each cluster will be relatively small, while the dissimilarities between each cluster will be relatively large. An ideal cut level, indicated by the dashed line, can be made in the gap separating the major within-cluster and between-cluster variances. This results in three main clusters, indicated by ovals surrounding the objects within each cluster, and one outlier (object 1).

will ultimately be grouped at a low level of similarity. On a dendrogram, this ideal clustering tree might look something like the hypothetical one found in Fig. 4. An experienced analyst can examine the dendrogram and determine a "cut level," or a degree of similarity where the tree can be cut, forming a discrete number of clusters, each containing a number of objects. Returning to the hypothetical example (see Fig. 4), the ideal cut level is somewhere along the "long branches" of the tree, where the number of clusters remains constant across a relatively wide range of similarity values. This produces clusters that contain objects that appear similar to other objects within a cluster, but different from objects found in other clusters. In this example, there is an object that does not appear to belong to any of the other groups. This object is less similar to any of the other clusters than the clusters are to each other. This type of object is typically called an *outlier* and is often considered unclassifiable (Doswell 1991). Of course, this is an idealized example and results using real data can differ greatly from the ideal.

For every cluster analysis classification experiment in this work, the choice of the number of clusters was made using the following criteria. The dendrograms were cut to form four to five clusters containing the majority of objects in the training dataset; no more than six outlier events were allowed. The criteria were chosen based upon examination of initial results that showed that more than three clusters were required in order to maintain the stratiform events in a unique cluster. Since outliers could not be classified, the number of outliers was limited to maintain a relatively high number of classifiable events. Every attempt was made to cut each dendrogram at a point where there was substantial separation between the intracluster and intercluster variation. To use the cluster analysis results as a classification tool, each cluster was considered a separate class of objects. The definition of each class was determined by the highest percentage of cases detected from the expert classification for that particular cluster. Since there were more than three clusters selected on each dendrogram, more than one cluster could have the same class definition (linear, cellular, stratiform). The percentage of objects that were correctly classified by their membership in the dominant expert class was used as the metric for determining the performance of the experimental classification.

f. Experimental classification results

Using the four attributes developed in section 2d (shape and scale parameters of the gamma distribution and eccentricity and area of the 0.6 correlogram contour), classification experiments using hierarchical cluster analysis were performed to determine what combination of attributes produced the best classification results. The question of whether or not to standardize the attributes prior to cluster analysis was investigated by testing the raw attributes, normalizing each attribute vector to produce zero mean and unit variance, and standardizing each attribute by dividing by its maximum value. For each of these transformations, all possible combinations of two, three, and four attributes were tested. The results of these 33 experiments involving different combinations of attributes and standardization (Fig. 5) demonstrate that the 0.6 correlogram contour eccentricity (a/b) and the gamma-scale parameter (β) were the attributes with the most discriminating power. The best results were obtained when these attributes were used in combination. When additional attributes were added, results degrade slightly. When only one of these were used in combination with other attributes, results were also degraded. Standardization of these attributes had little impact upon the classification results, since the ranges of the attribute values were quite similar (typically 0–10). Therefore, the gamma-scale parameter and correlogram eccentricity measures from *four* contours (correlation = 0.2, 0.4, 0.6, and 0.8) were tested for their effectiveness.

Figure 6 shows cluster analysis results using the gamma-scale parameter and correlogram eccentricity values for the 0.2, 0.4, 0.6, and 0.8 contours as attributes. A cut was made on the dendrogram separating the objects into five main clusters with six outliers. While the cut level is subjective, this figure clearly shows that the attributes used to describe these objects allow a cluster analysis algorithm to produce a group of stratiform objects (cluster 1), two groups of mainly cellular objects (clusters 2 and 3) and two groups of mainly linear objects (clusters 4 and 5). The first cluster was unanimously populated with all 11 stratiform events. The distinguishing feature of this stratiform cluster mean was a low value of the gamma-scale parameter. The second cluster was dominated by nine cellular cases, with one linear case included. The third cluster was also cellular dominant, with six cases. These cellu-

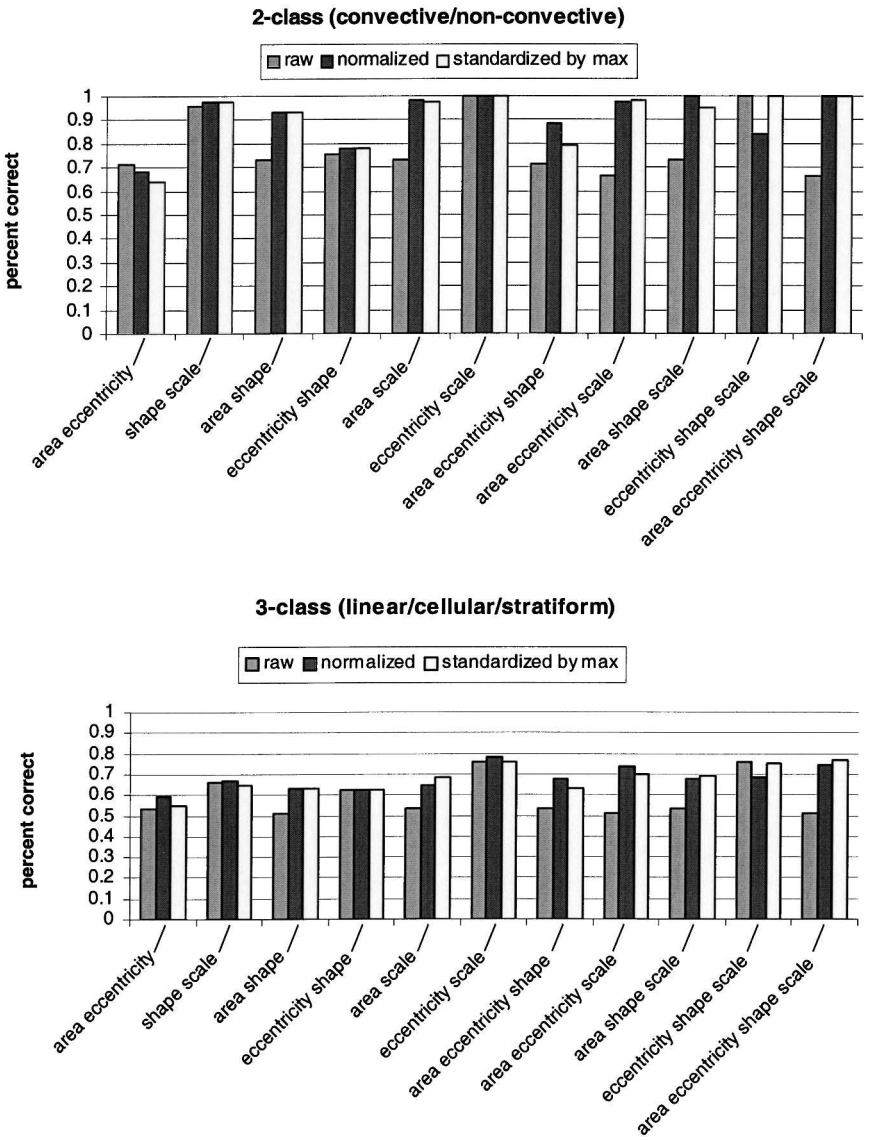

FIG. 5. Percent correct results for 33 experiments in the (a) two-class and (b) three-class cases. Results using raw attributes are in gray, attributes normalized to zero mean and unit variance are dark gray, and attributes standardized by their maximum are in white. The combination of attributes used in each experiment is indicated below each bar on the *x* axis.

lar clusters had relatively low mean values of correlogram contour eccentricity, indicating a lower degree of linear organization. The fourth cluster contains five linear cases, and the fifth cluster was split between seven linear cases and three cellular events. Clusters 4 and 5 were considered linear and had relatively high mean values of correlogram contour eccentricity, indicating a high degree of elongation. Validating these clusters, at the convective/nonconvective level of classification hierarchy, there were no incorrectly classified events, resulting in 100% correct classification. At the more refined linear/cellular/stratiform level of classification the clusters correctly placed 38 of 42 cases into the dominant class, or 90.5% correct. The classification using these five attributes successfully separated the cellular, linear, and stratiform events with over a 90% accuracy rate. This level of success demonstrates that useful attributes for an automated rainfall pattern classification system were discovered. This result was used as the basis for the automated classification procedure, which will be described in the next section.

3. Testing phase

The training phase in the process of developing an automated classification procedure discovered a set of morphological characteristics that were used by a cluster analysis algorithm to classify rainfall systems in a

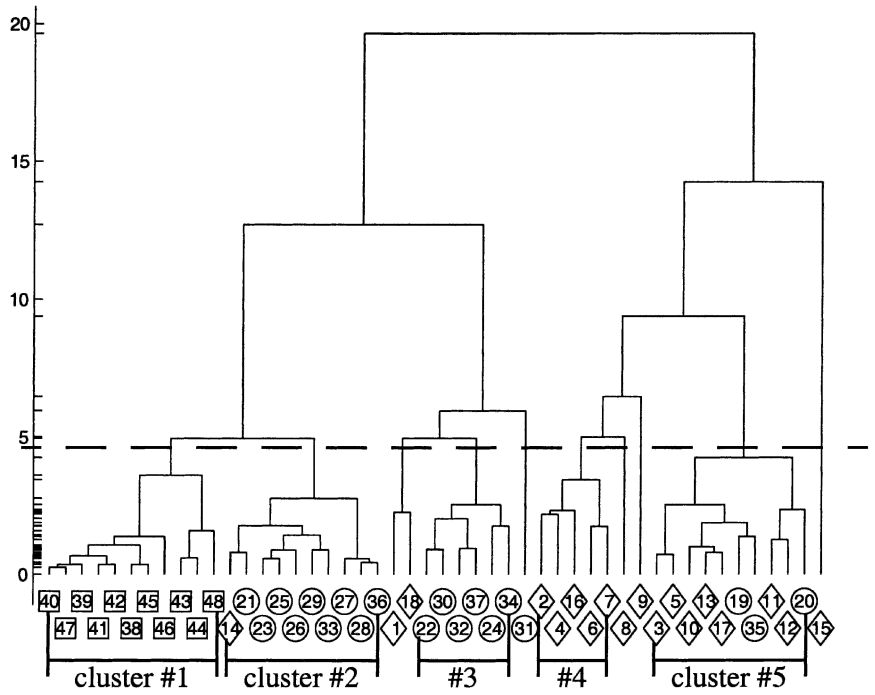

FIG. 6. Dendrogram produced by Ward's method with training dataset using the scale parameter from the gamma distribution and eccentricity of the 0.2, 0.4, 0.6, and 0.8 correlogram contours as attributes. The expert classification of each object is indicated by symbols, linear events are diamonds, cellular events are circles, and stratiform events are squares. Dashed line indicates cut level for this classification experiment.

manner consistent with an expert classification. The next phase of the development process involves establishing the procedures for identifying and rules for classifying rainfall systems in an automated fashion. Rainfall systems from 2002 were identified and analyzed so that statistical characteristics could be extracted. The parameters of the gamma distribution were fit to the histogram of rainfall amounts for each object using the generalized method of moments. A correlogram was computed for each object as well. Only the rainfall values contained within each unique object were included in the calculation. In other words, the rainfall from a neighboring rainfall system did not affect the correlogram for a given system. Correlation contours surrounding the origin on the correlogram were analyzed using several image processing routines described previously, and characteristics of those contours were extracted. Rainfall systems were classified by their similarity to classes previously defined by the cluster analysis classification experiments. To validate this automated classification procedure, results were compared to an expert classification of an independent testing dataset.

a. Automated object identification procedure

As discussed in section 2a, the rainfall objects in the training dataset were selected by hand. This method cannot be used in a completely automated classification procedure. Therefore, an automated rainfall object identification procedure, first introduced by Baldwin and Lakshmivarahan (2003), was developed in order to analyze rainfall events beyond the training dataset in this work. A simple threshold (0.05 mm) was used to convert each rainfall image into a binary image. A connected component labeling algorithm (Klette and Zamperoni 1996) was applied to this binary image to locate individual objects within the full image and identify them with a separate label. This algorithm labels pixels that are connected to other pixels with the same label. As in the training dataset, the areal extent of each object was increased by an integer number of pixels in each direction such that the object's area increased by as close to 15% as possible. In addition, it is not unusual to find small gaps between nearby regions of measurable rain. Therefore, the definition of *connected* pixels was expanded so that pixels that were within five pixels (~20 km) of one another were considered connected, and therefore given the same label value.

This algorithm will be illustrated via an example. In Fig. 7a, a subset of the U.S. rainfall analysis domain is shown for a case from July 2002. Here, a fairly large contiguous area of rainfall covers most of southern Minnesota. There are other smaller areas of rainfall over North and South Dakota, and a few pixels of scat-

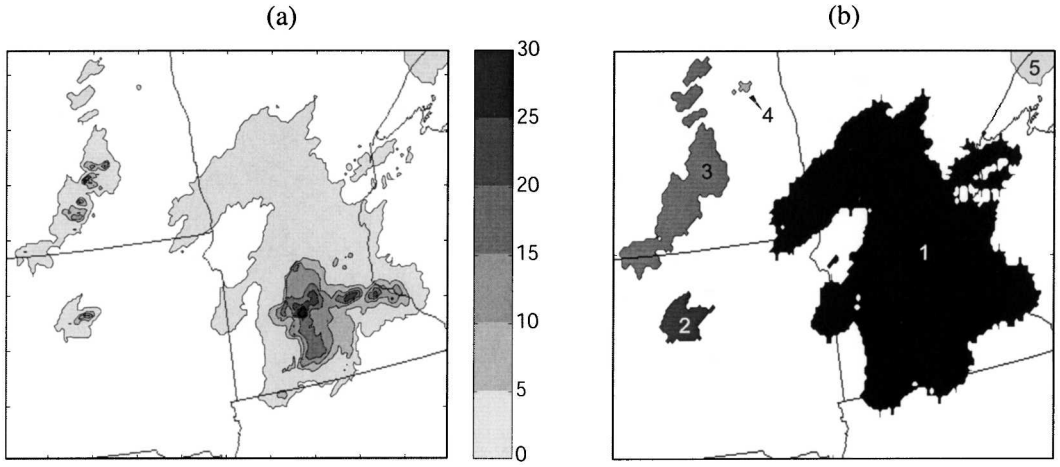

FIG. 7. Steps of the rainfall object identification process. (a) One-hour rainfall (mm) valid 2300 UTC 28 Jul 2002, with 5-mm contour interval. (b) Rainfall object labeling.

tered light rain appear in Wisconsin just east of the main heavy rainfall region. Figure 7b displays the final result of this process. There are five separate objects shown in this domain, since only a portion of object 5 in the figure is located within this domain; the four objects that are completely illustrated in this figure will be analyzed in more detail. The Minnesota rainfall has become a single object (object 1); the small region of intense rainfall in central South Dakota has also become a single object (object 2). The rainfall in North Dakota has been identified as two separated objects; object 3 contains the heavier rain plus the scattered light rain located adjacent and to the north, and object 4 is a small region of scattered light rain in eastern North Dakota.

b. Automated classification procedure

The automated classification procedure places previously unknown objects into one of five classes (two cellular classes, two linear classes, and one stratiform class), determined by the best cluster analysis results involving the training dataset discussed in section 2. Each of these classes was defined by the mean of the attribute vectors for the members of the five main clusters shown in Fig. 6. The mean attribute vectors for these clusters are provided in Table 2. Given an object for classification, the raw values of five attributes (gamma-scale parameter; eccentricity of 0.2, 0.4, 0.6, and 0.8 correlogram contours) were compared to these cluster means. The Euclidean distances between the object and each of the five cluster means were computed. The object was placed into the class represented by the nearest-neighbor cluster in terms of the smallest Euclidean distance.

For example, results of the automated classification procedure for the four rainfall objects represented in Fig. 7b are provided (attributes for these objects are listed in Table 3). In this example, the automated classification procedure classified the large contiguous region of rainfall over Minnesota (object 1) as cellular. The heavier rain within this object is not situated along a single axis, leading to low values of correlogram eccentricity. In terms of Euclidean distance, this object is closest to class 3, one of the cellular classes. Similarly, the scattered area of rainfall located in North Dakota (object 3) was also classified as cellular by the automated procedure. The heavier rain in object 3 was contained in two nearly circular blobs in the center of the object, resulting in relatively low eccentricity values and placement in class 3 (cellular). On the other hand, the smaller object in South Dakota (object 2) was classified as linear by the automated procedure. The heavier rain in object 2 was considerably elongated, leading to relatively high values of eccentricity (particularly for the 0.2 and 0.4 correlation contours). This object was

TABLE 2. Cluster mean attribute vectors, from five clusters denoted in Fig. 6.

	Identification	Gamma scale	Eccentricity 0.2 contour	Eccentricity 0.4 contour	Eccentricity 0.6 contour	Eccentricity 0.8 contour
Cluster 1	Stratiform	0.75	2.72	1.91	1.61	1.76
Cluster 2	Cellular	2.82	2.72	2.39	1.90	2.07
Cluster 3	Cellular	5.62	1.97	2.10	2.31	1.79
Cluster 4	Linear	3.56	6.36	8.08	3.66	3.09
Cluster 5	Linear	2.63	5.28	4.03	2.89	2.80

TABLE 3. A sample of attributes extracted from the four objects found in Fig. 7b.

Label	Gamma scale	Eccentricity 0.2 contour	Eccentricity 0.4 contour	Eccentricity 0.6 contour	Eccentricity 0.8 contour
Object 1	6.95	1.43	1.42	1.50	1.58
Object 2	4.67	6.32	4.47	1.00	1.00
Object 3	5.05	2.85	3.16	2.00	1.00
Object 4	0.06	1.00	1.00	1.00	1.00

closest to class 5, one of the two linear classes. Finally, the small region of light rain in eastern North Dakota (object 4) was considered stratiform by the automated procedure. Object 4 contained only light rain, which resulted in a low value of gamma-scale parameter and placement in class 1, the stratiform class.

c. Analysis of 2002 data

Each hourly stage IV analysis from 2002 was processed using the automated rainfall system identification and classification procedures described in the previous sections. Out of a possible 8760 h over the course of the year, 8679 h, or 99.1% of the hours in the year, were included in the dataset that was obtained from NCEP. In total, 799 014 objects (rainfall systems), or an average of 92 objects per hour, were identified by the automated system. The distribution of objects as a function of their size (number of pixels) is shown in Fig. 8. The histogram (see Fig. 8) clearly shows that the majority of objects are relatively small in size. By approximating the length scales of mesoscale phenomena suggested by Orlanski (1975), the objects can be grouped into three size-related categories, small (meso-γ) objects of size 150 pixels (approximately 50×50 km^2) or less, medium-sized (meso-β) objects greater than 150 pixels and less than or equal to 2000 pixels ($\sim 200 \times 200$ km^2), and large (meso-α) objects of size greater than 2000 pixels. For instance, in the 2002 data there were 524 224 small objects (65.6% of the total, an average of 60 h^{-1}), 242 914 medium-sized (30.4%, average 28 h^{-1}) objects, and 31 876 large (4%, 3.7 h^{-1}) objects. In terms of areal coverage, the large objects were responsible for 73% of the precipitation area, medium objects 23%, and small objects 4% of the total areal coverage for 2002. In terms of total precipitation volume, the large objects represented nearly 87% of the total precipitation, medium objects 12%, and small objects produced only 1% of the total precipitation. Although the large objects only represented a small fraction of the total number of objects, they produced the majority of precipitation amount and areal coverage during 2002. For reference, Fig. 7b provides examples of typical objects in each size regime, a large object over Minnesota (object 1: 173 380 km^2), medium-sized objects in North Dakota (object 3: 26 084 km^2) and South Dakota (object 2: 6237 km^2), and a small object in eastern North Dakota (object 4: 635 km^2).

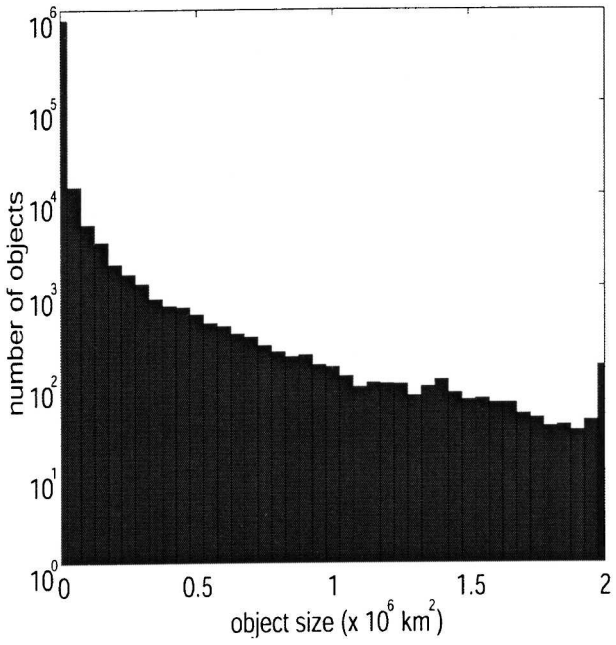

FIG. 8. Distribution of 2002 rainfall objects by size (x 10^6 km^2).

d. Testing dataset

To test the automated classification procedure, a random sample of objects from the 2002 data was selected. Since the 2002 data are dominated by objects of small size, one would expect a sample taken from the entire population to consist mostly of small-sized objects. Examination of the attributes associated with the small objects found these to be quite consistent, characterized by eccentricity values nearly equal to one and very small values of the gamma-scale parameter. Medium-sized objects also showed consistent characteristics, although not to the same extent as the small objects. These objects were dominated by small values of the scale parameter, and a large fraction had eccentricity equal to one as well. The vast majority of the small- and medium-sized objects were classified as stratiform because of the low values of gamma-scale and correlogram eccentricity. This finding seems counterintuitive, since one would expect a significant fraction of meso-β- and meso-γ-scale systems to be convective in nature. Cursory examination of a small number of small-sized events showed them to be likely associated with noise,

anomalous propagation, or ground clutter. Given the large number of events to consider, further analysis is necessary to determine the reasons for this counterintuitive result. However, it is likely that alternate methods of parameter estimation and characterization are required for objects that contain a relatively small number of pixels. Further refinement of the classification procedure will also be necessary in order to properly classify the smaller-scale rainfall systems.

The training dataset that the automated classification system was built upon consisted entirely of large objects (smallest object had just under 3250 pixels). Large-sized objects from the 2002 data had attributes whose values varied across a wide range, including time of day, time of year, and location. As mentioned previously, the large objects represented the majority of precipitation amount and areal coverage during 2002. Therefore, the testing data sample was taken entirely from the large-sized object regime.

Because it was not feasible to perform expert classification on the entire dataset, which consisted of 31 876 large-sized objects, a sample of 100 objects was randomly selected to create an evaluation, or testing dataset. To confirm that this sample was representative of the entire population, the distributions of various attributes associated with these objects were compared to the summary statistics associated with the large objects (see Figs. 9 and 10). As shown in Figs. 9a and 9b, but for an anomalous peak in the early morning, the diurnal cycle of the validation sample was quite similar to the overall large object distribution, generally decreasing during the evening and overnight hours, then increasing to a peak in the late afternoon. The distribution of objects from the random sample during the course of the year (see Figs. 9c and 9d) was also representative of the entire population, with relatively high frequency in the warm season and low frequency in the cool season. The test sample was also well distributed across the United States (see Figs. 9e and 9f) with somewhat dense clusters of sample objects in South Florida and the Pacific Northwest in the same vicinity of maximum density in the overall distribution. The distribution of the validation sample objects in attribute space (see Fig. 10) also appeared to be representative of the entire 2002 population of large objects, with scale values ranging from 0.1 to 10, shape values ranging from 0.1 to 1, and eccentricity values ranging from 1 to 10. However, one object in the testing sample did appear to be an outlier, with a very small gamma-scale value, large shape parameter, and eccentricity slightly greater than one. These results showed that this sample was representative of the population and exhibited an interesting range of attribute values. The automated classification procedure was tested by comparing the results of an expert classification of this 100-object sample to those from the automated classification system.

e. Testing results

Each object from the evaluation sample was classified into five classes by the automated classification procedure. Class 1 was a stratiform class, classes 2 and 3 were both cellular, and classes 4 and 5 were linear. Figure 11a shows the results of this classification. The most popular individual class was the stratiform class, where 39% of the objects were classified. However, combining classes 2 and 3 (cellular) showed that 46% of the objects were considered cellular by the automated system. Linear events were the most rare: combining classes 4 and 5 resulted in 15% of the objects in the validation sample. Comparing this with the automated classification results for all of the large objects in the 2002 data (see Fig. 11b) further confirmed that the testing dataset was representative of the population (43% stratiform, 39% cellular, and 18% linear).

The expert classification of the evaluation sample separated the rainfall systems into three classes: linear, cellular, and stratiform. These results were compared with the automated classification results, where classes 2 and 3 were combined into a cellular class and classes 4 and 5 were combined into a linear class. Overall, 89% of the objects were correctly classified into the parent convective/nonconvective classes (two-class case), and 85% of the objects were correctly classified in the three-class case (stratiform, linear, cellular).

Further examination of these results showed that the majority of incorrectly classified cases were considered nonconvective by the expert classification and classified as convective by the automated procedure. In fact, seven of the cases were classified as linear by the automated classification and stratiform in the expert classification. Closer examination of these cases showed that high values of correlogram eccentricity tended to place them into a linear class even though the value of the gamma-scale parameter was small, indicating a lack of heavy rainfall. An example of an error of this type is provided by case 3 of the sample, an object located over northern Michigan at 0500 UTC 16 May 2002 (see Fig. 12). The rainfall associated with this object is generally light and widespread, which lead to the expert stratiform classification and a value of the gamma-scale parameter of 0.7. At the same time, the rainfall is somewhat organized along a line, represented by the relatively high eccentricity values, in particular, $a/b = 4.6$ for the 0.4 correlogram contour. In terms of Euclidean distance to the five cluster means, this object was closest to the linear cluster 5 due mainly to the high values of eccentricity. Through detailed analysis of other errors of this type, further improvements to the classification procedure will likely be realized in future work.

4. Conclusions

The overall goal of this work was to develop a completely automated rainfall system classification proce-

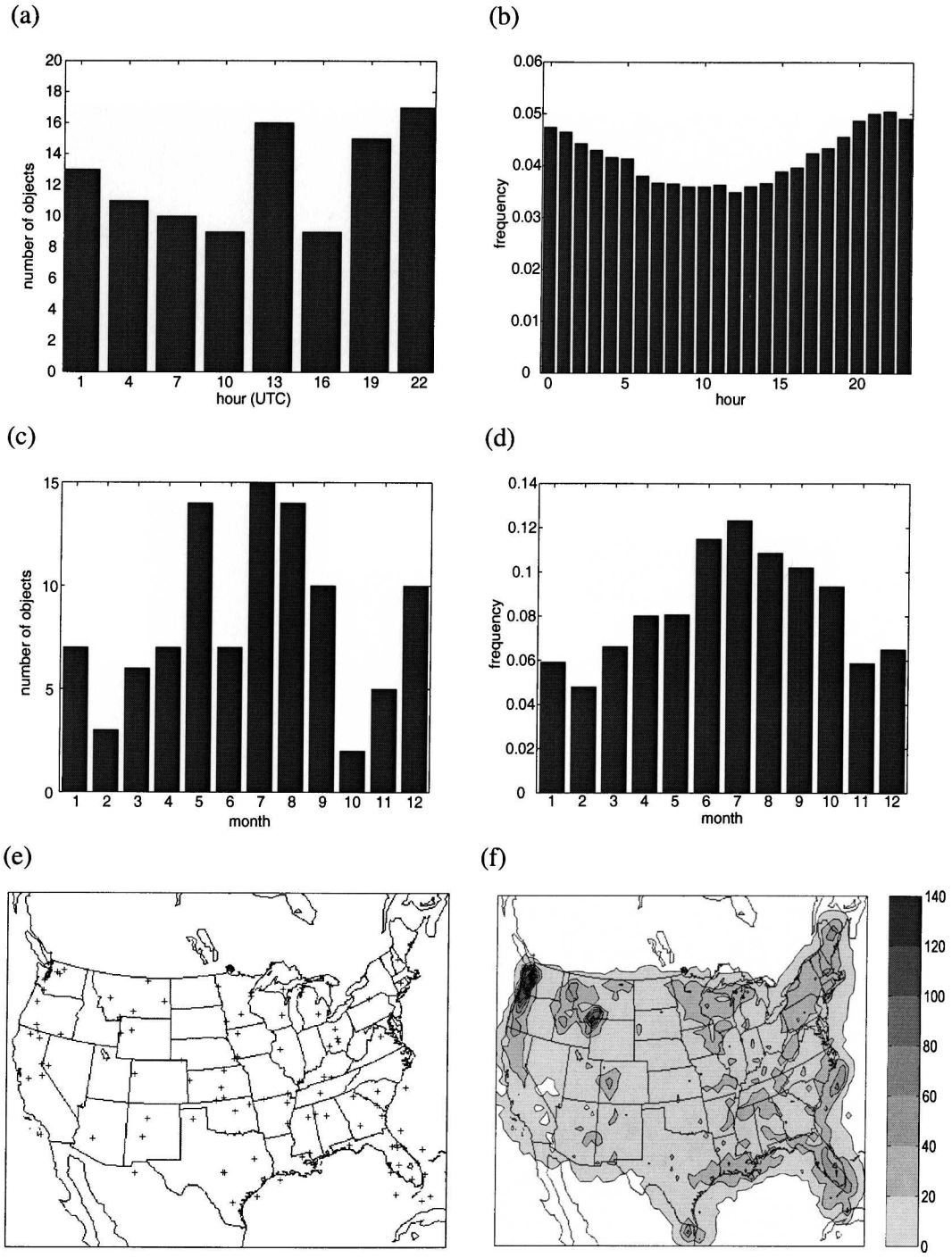

FIG. 9. Comparison of characteristics of (left column) evaluation sample and (right column) large objects in 2002 dataset. (a), (b) Distribution of objects as a function of time of day. (c), (d) Distribution of objects by month. (e), (f) Distribution of object center of mass locations.

dure. To accomplish this task, the discovery of a set of attributes that allow for characterization of morphological aspects of rainfall patterns was required. This task was accomplished via a relatively small training dataset, comparing the results of various cluster analysis experiments with an expert classification. These experiments involved reducing the dimension of the data by analyzing the "bulk" global distribution of rainfall values across each object, using the histogram of rainfall values representing each object. The gamma distri-

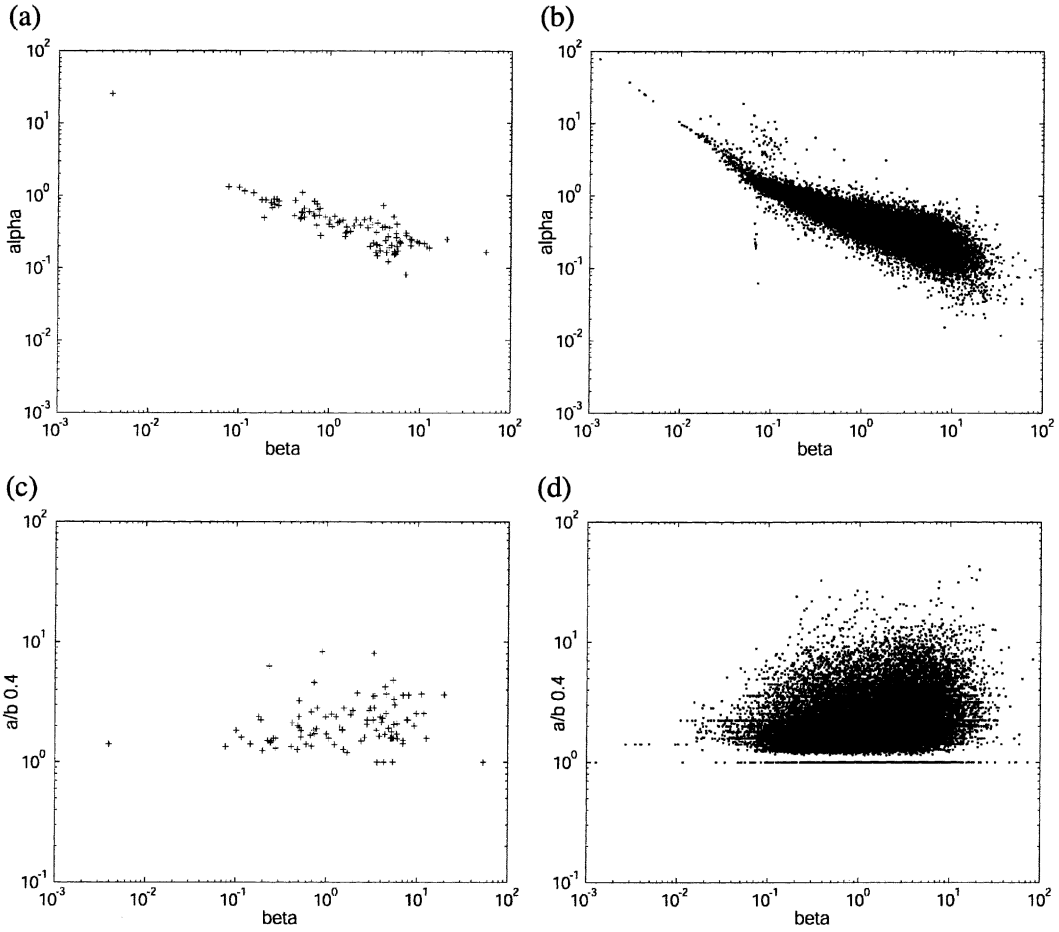

FIG. 10. Comparison of characteristics of (left column) evaluation sample and (right column) large objects in 2002 dataset. (a), (b) Object distribution density in a gamma-shape, gamma-scale (alpha, beta) plane. (c), (d) Object distribution density in gamma-scale, 0.4 correlogram contour eccentricity plane.

bution was selected as a compact model of the observed histograms. The shape and scale parameters of the gamma distribution were fit to each histogram using the generalized method of moments technique. Information regarding the degree of elongation of the rainfall systems was obtained via geostatistical measures. The correlogram was selected for analysis because it is independent of the magnitude of the rainfall values. By estimating the area and eccentricity of various correlation contours in the correlogram, useful information on the degree of linear organization within each rainfall system was obtained. The cluster analysis using the gamma-scale parameter along with the eccentricity of four correlogram contours as attributes successfully separated the training dataset into linear, cellular, and stratiform classes. These results were used as the basis for an automated classification procedure.

The classification system was based upon a nearest-neighbor approach, using the best results from the previous cluster analysis experiments using the training dataset. An independent evaluation of the procedure was required. To obtain a representative, random sample of the rainfall object population, analysis of the characteristics of rainfall systems across an entire year was performed. Rainfall objects (or systems) were defined as contiguous regions of precipitation. Summary statistics of attributes from this year were examined, and a random sample of interesting objects was pulled from the 2002 data. The distribution of the random sample was compared with the summary statistics in order to confirm that this validation dataset was representative of the population. The results of an expert classification were compared to those from the automated classification procedure and showed that the automated classification accurately placed 85% of the objects into correct linear, cellular, and stratiform classes, and 89% of objects into their correct parent convective/nonconvective class. Therefore, an independent test of the performance of this classification system was obtained.

While the automated rainfall system classification procedure developed in this work produced satisfactory

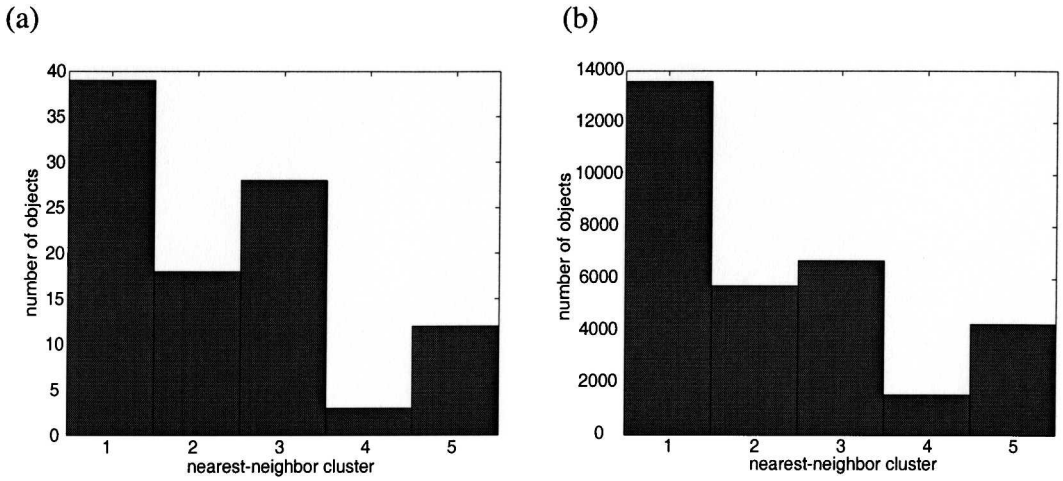

FIG. 11. Distribution of objects by the automated classification system. (a) Results from the evaluation sample; (b) results for all large objects from 2002. Class 1 is the stratiform class, classes 2 and 3 are cellular, and classes 4 and 5 are linear.

results, further refinement of the methods may result in a variety of improvements. Image segmentation routines, such as those proposed by Peak and Tag (1994) and Lakshmanan (2001) may prove to be beneficial in locating rainfall systems within the full analysis domain. These may be especially useful in subdividing synoptic-scale, contiguous areas of rainfall, which are currently defined to be a single rainfall system. One might wish to separate a convective line associated with a strong surface cold front from one that is connected to warm frontal bands within the stratiform region of a cyclone. The inclusion of other sources of rainfall-related data, such as lightning, radar reflectivity, satellite radiances, etc., may also help to improve the classification.

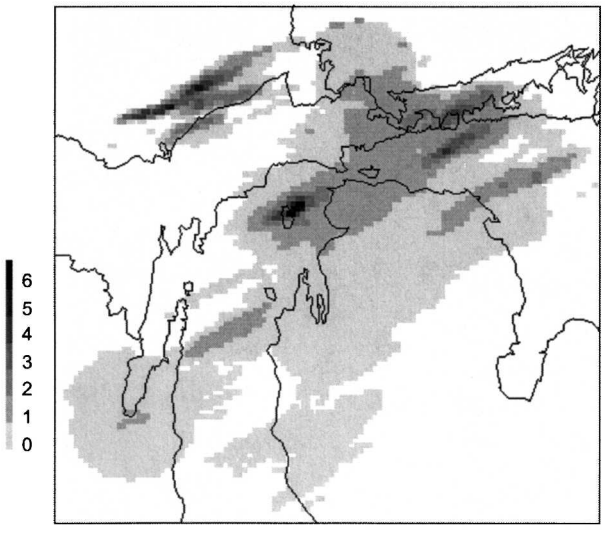

FIG. 12. Object 3 from the testing dataset, 1-h accumulated rainfall (mm) valid 0500 UTC 16 May 2002. Only rainfall amounts contained within the object are plotted.

Further refinements in the classification hierarchy are also desirable. For example, the degree to which the attributes used in this work will divide the linear class into more refined classes [such as symmetric/asymmetric as in Houze et al. (1990) or leading stratiform, parallel stratiform, and trailing stratiform as in Parker and Johnson (2000)] should be determined. If the attributes currently in use do not have the power to further discriminate among desired subclasses, additional attributes that have this ability must be discovered.

There are many potential applications for the automated rainfall system classification procedure developed in this work. For example, verification, predictability, and climatological studies may benefit from such a classification system. Since a multiyear archive of stage IV analyses is available, interannual variability of rainfall events could be studied. Through the use of operational gridded analyses of environmental conditions [such as those produced by the Rapid Update Cycle at NCEP (Benjamin et al. 1994)], the relationship between system types and the thermodynamic and environment flow conditions associated with them could be studied further (e.g., Perica and Foufoula-Georgiou 1996). Severe weather reports could also be associated with the various classes of collocated rainfall systems, possibly leading to improved forecasts of hazardous weather.

Acknowledgments. This research was performed as part of the lead author's doctoral dissertation. Accordingly, the lead author wishes to acknowledge his coauthors along with the other members of his doctoral committee (Dr. Frederick Carr, Dr. Kelvin Droege-

meier, Dr. Mark Morrissey, and Dr. Michael Richman) for their inspiration and patient guidance.

This manuscript benefited greatly from the constructive comments and suggestions provided in a preliminary review by Dr. V. Lakshmanan of the National Severe Storms Laboratory and by four anonymous reviewers. The authors are particularly grateful to several people who provided software and data for this project. Dr. Ying Lin of the Environmental Modeling Center provided archived rainfall analysis data. Dr. Ahmed Alhamed provided cluster analysis software. Other software for solving nonlinear least squares and matrix inversion was obtained from the Netlib.org repository. The GSLIB software library was used to compute geostatistical measures. This work was partially funded by a NOAA–University of Oklahoma Cooperative Agreement NA17RJ1227, and COMET Cooperative Project UCAR Award S01-32769.

APPENDIX

Generalized Method of Moments

The GMM (Hansen 1982; Hamilton 1994) can be considered an extension to the classical method of moments. In the method of moments, the parameters of the theoretical distribution are found by developing a system of equations that equate the population moments with their sample counterparts. If there are K unknown parameters in the theoretical distribution, K such equations are necessary. The resulting parameters will produce a theoretical distribution that fits those K moments exactly. However, rather than fitting K moments exactly, it may be desirable for the parameters to fit $K + L$ moments as closely as possible. For example, if two parameters are unknown, one might desire to produce parameter estimates that provide the best fit to the first, second, third, and fourth moments of the sample. A nonlinear vector function $g(\boldsymbol{\theta}, r)$ can be produced representing the differences between the sample moments ($\hat{\mu}_n$) and the population moments, using the gamma distribution, for example,

$$g(\boldsymbol{\theta}, \mathbf{r}) \equiv \begin{bmatrix} \{\hat{\mu}_1 - \alpha\beta\} \\ \{\hat{\mu}_2 - \alpha\beta^2(\alpha + 1)\} \\ \{\hat{\mu}_3 - \alpha\beta^3(\alpha^2 + 3\alpha + 2)\} \\ \{\hat{\mu}_4 - \alpha\beta^4(\alpha^3 + 6\alpha^2 + 11\alpha + 6)\} \end{bmatrix}. \tag{A1}$$

Here, $\theta = [\alpha\ \beta]$ is the vector containing the parameters of the gamma distribution, $\mathbf{r}$ represents the vector of length N containing the sample rainfall, and $\hat{\mu}_n = \langle 1/N \rangle \Sigma_{t=1}^{N} (r_t)^n$ is the nth sample moment. One can create an objective scalar function $\Phi(\theta) = g(\boldsymbol{\theta}, \mathbf{r})^T\ \mathbf{A} g(\boldsymbol{\theta}, \mathbf{r})$, which is the weighted sum of squared errors of the estimates of the parameters, where $\mathbf{A}$ is a symmetric positive-definite weighting matrix that represents the relative importance of fitting each of the moments. In this work, the parameter vector θ that minimizes this function was found iteratively using the bounded truncated-Newton method (Nash 1984).

GMM can allow correlation in the data to affect the parameter estimation. The optimal weighting matrix $\mathbf{A}^*$ is the inverse of the parameter error covariance matrix $\mathbf{S}$. If the data are serially uncorrelated, an estimate of the error covariance matrix is the second moment matrix:

$$\hat{\mathbf{S}} = (1/N) \sum_{t=1}^{N} g(\hat{\boldsymbol{\theta}}, \mathbf{r}_t) g(\hat{\boldsymbol{\theta}}, \mathbf{r}_t)^T, \tag{A2}$$

which is the mean outer product matrix of the errors of the estimated parameters. Serial correlation in the data can be taken into account by modifying the estimate of the second-moment matrix (Newey and West 1987):

$$\hat{\mathbf{S}} = \hat{\mathbf{G}}_0 + \sum_{\nu=1}^{q} \{1 - [\nu/(q+1)]\} (\hat{\mathbf{G}}_\nu + \hat{\mathbf{G}}_\nu^T), \tag{A3}$$

where

$$\hat{\mathbf{G}}_\nu = (1/N) \sum_{t=\nu+1}^{N} [g(\hat{\boldsymbol{\theta}}, \mathbf{r}_t)][g(\hat{\boldsymbol{\theta}}, \mathbf{r}_{t-\nu})]^T. \tag{A4}$$

Newey and West (1987) show that Eq. (A3) provides a consistent estimate of the covariance matrix if q grows as a fractional power of sample size ($q < T^{1/4}$).

Note that in order to compute the second-moment matrix, an estimate of the unknown parameters ($\boldsymbol{\theta}$) is needed. An iterative procedure is followed in which an initial estimate of the parameters ($\boldsymbol{\theta}_0$) is obtained using an arbitrary weighting matrix such as the identity matrix $\mathbf{A}_0 = \mathbf{I}$. This estimate of $\boldsymbol{\theta}$ is used in Eq. (A3) to produce an initial estimate of $\mathbf{S}$, which is inverted to produce $\mathbf{A}_1$. The objective function Φ is minimized using $\mathbf{A}_1$ to produce a new estimate $\boldsymbol{\theta}_1$, which is then used to estimate $\mathbf{A}_2$. These iterations continue until convergence is reached. For all cases in this work convergence was reached in five iterations or less.

REFERENCES

Alhamed, A., S. Lakshmivarahan, and D. J. Stensrud, 2002: Cluster analysis of multimodel ensemble data from SAMEX. *Mon. Wea. Rev.,* **130,** 226–256.

Anderson, C. J., and R. W. Arritt, 1998: Mesoscale convective complexes and persistent elongated convective systems over the United States during 1992 and 1993. *Mon. Wea. Rev.,* **126,** 578–599.

Anthes, R. A., 1983: Regional models of the atmosphere in middle latitudes. *Mon. Wea. Rev.,* **111,** 1306–1335.

Austin, P. M., and R. A. Houze Jr., 1972: Analysis of the structure of precipitation patterns in New England. *J. Appl. Meteor.,* **11,** 926–935.

Baldwin, M. E., and K. E. Mitchell, 1998: Progress on the NCEP hourly multi-sensor U.S. precipitation analysis for operations and GCIP research. Preprints, *Second Symp. on Integrated Observing Systems*, Phoenix, AZ, Amer. Meteor. Soc., 10–11.

——, and S. Lakshmivarahan, 2002: Rainfall classification using

histogram analysis: An example of data mining in meteorology. *Intelligent Engineering Systems through Artificial Neural Networks,* Vol. 12, C. H. Dagli et al., Eds., ASME Press, 429–434.

——, and ——, 2003: Spatial characterization of rainfall patterns for use in a classification system. *Intelligent Engineering Systems through Artificial Neural Networks,* Vol. 13, C. H. Dagli et al., Eds., ASME Press, 683–688.

Benjamin, S. G., K. J. Brundage, P. A. Miller, T. L. Smith, G. A. Grell, and D. Kim, J. M. Brown, and T. W. Schlatter, 1994: The Rapid Update Cycle at NMC. Preprints, *10th Conf. on Numerical Weather Prediction,* Portland, OR, Amer. Meteor. Soc., 566–568.

Biggerstaff, M. I., and S. A. Listemaa, 2000: An improved scheme for convective/stratiform echo classification using radar reflectivity. *J. Appl. Meteor.,* **39,** 2129–2150.

Blanchard, D. O., 1990: Mesoscale convective patterns of the southern high plains. *Bull. Amer. Meteor. Soc.,* **71,** 994–1005.

Bluestein, H. B., and M. H. Jain, 1985: Formation of mesoscale lines of precipitation: Severe squall lines in Oklahoma during the spring. *J. Atmos. Sci.,* **42,** 1711–1732.

——, G. T. Marx, and M. H. Jain, 1987: Formation of mesoscale lines of precipitation: Nonsevere squall lines in Oklahoma during the spring. *Mon. Wea. Rev.,* **115,** 2719–2727.

Davies, E. R., 1997: *Machine Vision: Theory, Algorithms, Practicalities.* Academic Press, 750 pp.

Deutsch, C. V., and A. G. Journel, 1998: *GSLIB: Geostatistical Software Library and User's Guide.* 2d ed. Oxford University Press, 369 pp.

Doswell, C. A., 1991: Comments on "Mesoscale convective patterns of the southern high plains." *Bull. Amer. Meteor. Soc.,* **72,** 389–390.

Duda, R. O., P. E. Hart, and D. G. Stork, 2000: *Pattern Classification.* 2d ed. Wiley Interscience, 654 pp.

Ebert, E. E., and J. L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrol.,* **239,** 179–202.

Evans, J. L., and R. E. Shemo, 1996: A procedure for automated satellite-based identification and climatology development of various classes of organized convection. *J. Appl. Meteor.,* **35,** 638–652.

Fulton, R. A., J. P. Breidenbach, D. J. Seo, D. A. Miller, and T. O'Bannon, 1998: The WSR-88D rainfall algorithm. *Wea. Forecasting,* **13,** 377–395.

Gamache, J. F., and R. A. Houze, 1983: Water budget of a mesoscale convective system in the Tropics. *J. Atmos. Sci.,* **40,** 1835–1850.

Geerts, B, 1998: Mesoscale convective systems in the southeast United States during 1994–95: A survey. *Wea. Forecasting,* **13,** 860–869.

Hamilton, J. D., 1994: *Time Series Analysis.* Princeton University Press, 799 pp.

Hansen, L. P., 1982: Large sample properties of generalized method of moments estimators. *Econometrica,* **50,** 1029–1054.

Hosking, J. G., and C. D. Stow, 1987: Ground-based, high resolution measurements of the spatial and temporal distribution of rainfall. *J. Climate Appl. Meteor.,* **26,** 1530–1539.

Houghton, H. G., 1968: On precipitation mechanisms and their artificial modification. *J. Appl. Meteor.,* **7,** 851–859.

Houze, R. A., 1997: Stratiform precipitation in regions of convection: A meteorological paradox? *Bull. Amer. Meteor. Soc.,* **78,** 2179–2196.

——, B. F. Smull, and P. Dodge, 1990: Mesoscale organization of springtime rainstorms in Oklahoma. *Mon. Wea. Rev.,* **118,** 613–654.

Isaaks, E. H., and R. M. Srivastava, 1989: *An Introduction to Applied Geostatistics.* Oxford University Press, 561 pp.

Johnson, R. H., and P. J. Hamilton, 1988: The relationship of surface pressure features to the precipitation and airflow structure of an intense midlatitude squall line. *Mon. Wea. Rev.,* **116,** 1444–1473.

Jirak, I. L., W. R. Cotton, and R. L. McAnelly, 2003: Satellite and radar survey of mesoscale convective system development. *Mon. Wea. Rev.,* **131,** 2428–2449.

Kessler, E., 1966: Computer program for calculating average lengths of weather radar echoes and pattern bandedness. *J. Atmos. Sci.,* **23,** 569–574.

——, and J. A. Russo Jr., 1963: Statistical properties of weather radar echoes. *Proc. 10th Weather Radar Conf.,* Washington, DC, Amer. Meteor. Soc., 25–33.

Klette, R., and P. Zamperoni, 1996: *Handbook of Image Processing Operators.* John Wiley and Sons, 397 pp.

Lakshmanan, V., 2001: *A Hierarchical, Multiscale Texture Segmentation Algorithm for Real-World Scenes.* Ph.D. dissertation, University of Oklahoma, 115 pp.

Locatelli, J. D., and P. V. Hobbs, 1974: Fall speeds and masses of solid precipitation particles. *J. Geophys. Res.,* **79,** 2185–2197.

Nash, S. G., 1984: Newton-type minimization via the Lanczos method. *SIAM J. Numer. Anal.,* **21,** 770–778.

Newey, W. K., and K. D. West, 1987: A simple, positive semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica,* **55,** 703–708.

Orlanski, I., 1975: A rational subdivision of scales for atmospheric processes. *Bull. Amer. Meteor. Soc.,* **56,** 527–530.

Parker, M. D., and R. H. Johnson, 2000: Organizational modes of midlatitude mesoscale convective systems. *Mon. Wea. Rev.,* **128,** 3413–3436.

Peak, J. E., and P. M. Tag, 1994: Segmentation of satellite imagery using hierarchical thresholding and neural networks. *J. Appl. Meteor.,* **33,** 605–616.

Perica, S., and E. Foufoula-Georgiou, 1996: Linkage of scaling and thermodynamic parameters of rainfall: Results from midlatitude mesoscale convective systems. *J. Geophys. Res.,* **101,** 7431–7448.

Romesburg, C. H., 1984: *Cluster Analysis for Researchers.* Life Time Learning, 334 pp.

Seo, D. J., 1998: Real-time estimation of rainfall fields using radar rainfall and rain gauge data. *J. Hydrol.,* **208,** 37–52.

Steiner, M., R. A. Houze Jr., and S. E. Yuter, 1995: Climatological characteristics of three-dimensional storm structure from operational radar and rain gauge data. *J. Appl. Meteor.,* **34,** 1978–2007.

Tsakraklides, G., and J. L. Evans, 2003: Global and regional diurnal variations in organized convection. *J. Climate,* **16,** 1562–1572.

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

Wilks, D. S., 1990: Maximum likelihood estimation for the gamma distribution using data containing zeros. *J. Climate,* **3,** 1495–1501.

——, 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.

Wilson, J. W., N. A. Crook, C. K. Mueller, J. Sun, and M. Dixon, 1998: Nowcasting thunderstorms: A status report. *Bull. Amer. Meteor. Soc.,* **79,** 2079–2099.

Yuter, S. E., and R. A. Houze Jr., 1997: Measurements of raindrop size distributions over the Pacific warm pool and implications for Z–R relations. *J. Appl. Meteor.,* **36,** 847–867.

Zawadzki, I. I., 1973: Statistical properties of precipitation patterns. *J. Appl. Meteor.,* **12,** 459–472.

Zhang, D.-L., and K. Gao, 1989: Numerical simulation of an intense squall line during 10–11 June 1985 PRE-STORM. Part II: Rear inflow, surface pressure perturbations, and stratiform precipitation. *Mon. Wea. Rev.,* **117,** 2067–2094.