UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

AUTOMATED CLASSIFICATION OF RAINFALL SYSTEMS USING STATISTICAL CHARACTERIZATION

A Dissertation

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

By

Michael Eugene Baldwin

Norman, Oklahoma

2003

UMI Number: 3085711

UMI Microform 3085711

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

AUTOMATED CLASSIFICATION OF RAINFALL SYSTEMS USING STATISTICAL CHARACTERIZATION

A Dissertation APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY

Dr. Frederick Carr

Dr. Kelvin Droegemeier

Dr. John Kain

Dr. S. Lakshmivarahan

Dr. Mark Morrissey

Dr. Michael Richman

Acknowledgements

I could not have completed this work were it not for the help and inspiration of many people. Over the seven years that it has taken me to bring this work to a conclusion, there have been several professors, colleagues, friends, and family members who have played a important part in helping me to get to finish line. Unfortunately, the number of people who should be acknowledged is too large for me to be able to include them all. But, I will do my best.

Since I went through this process as a "part-time" student, I think it would be best to first thank my employers, whose support allowed me to act like a graduate student while still working full-time. For the first couple of years, I worked for General Sciences Corporation (SAIC) as a support scientist for NCEP/EMC located in Camp Springs, Maryland. I particulary appreciate the support of my supervisor, Mr. Mike Pecnick, I doubt anyone could have a better boss. Mike will always rather talk about the good things in life over work, such as golf, football, beer, movies, etc. Similarly, I'd like to thank my EMC supervisor, Dr. Geoff Dimego, for his friendship and for allowing me to pursue this goal. I'd also like to thank Dr. Tom Black of EMC for his friendship and strong support over the years. For the last half of this process, I have been employed at the University of Oklahoma/CIMMS, and affiliated with both the National Severe Storms Laboratory and the Storm Prediction Center. Thank you to the directors of these institutions, Dr. Peter Lamb of CIMMS, Dr. Jeff Kimpel of NSSL, and Dr. Joe Schaefer of SPC, for allowing me the freedom to work on this research as part of my work duties. For me to be able to obtain this position, I especially appreciate the support of Mr. Gary Grice and Dr. Dave Stensrud, who worked with their supervisors to create this unique co-affilliated position. I

also greatly appreciate the support of my supervisor at NSSL, Dr. Harold Brooks. Over the years, this work has been partially funded by a NOAA-University of Oklahoma Cooperative Agreement NA17RJ1227, and COMET Cooperative Project UCAR Awards 099-15805 and S01-32769.

Of course, I could not have graduated except for the guidance, patience, and inspriation provided by the members of my committee. To my committee chair, Dr. Fred Carr, who I've worked with for so many years, thank you for your patience, guidance, inspiration, wisdom, and for allowing me the freedom to modify the focus of my research (more than once). Not surprisingly, Dr. Carr and I share a common desire to apply our research towards improving operational forecasting. He has a keen understanding of the "real-world" issues involved with the operatational forecasting environment. He also has a great sense of humor and loves to tell a good story. He takes great pride and care of his students, which I have always appreciated.

I am also very thankful to Dr. Jack Kain of CIMMS, my unofficial co-advisor here at NSSL, for his patience, guidance, and understanding. Dr. Kain has several times worked tirelessly to obtain funding to support our research. Without his efforts, I doubt that I would have been able to continue my full-time employee/part-time student status for as long as I have. Jack also has an innate desire to apply his reserach towards improving operational forecasting, as evidenced by his numerous collaborative research studies with operational forecasters and numerical modelers. So it was more than a happy coincidence that we both ended up in Norman at about the same time. Most importantly, Jack has always been a good friend. He has always been very supportive and helpful during all of those times when I've run into problems or frustrations with work, research, or school. I

am looking forward to working with Jack for many years to come.

I have benefitted tremendously from the brilliance, energy, enthusiasm, time, knowledge, and the occasional kick in the pants provided to me by the outside member of my committee, Dr. S. Lakshmivarahan. I was impressed with Dr. Varahan early on in my studies while taking a course in data assimilation which he primarily taught. In addition, several years ago his lectures to the Storm Prediction Center on cluster analysis planted the initial seeds that eventually grew into the form of this work. Professor Varahan cheerfully spent hours and hours of his time teaching me several of the central subjects that make up the foundation of this work, data mining, statistics, parameter estimation, image processing, etc. He certainly deserves a lot of credit for the time he has spent working with me individually. He has an amazing thirst for knowledge on a wide variety of topics. This was evidenced by the several hours of discussion that we had on the subject of dynamic meteorology and numerical weather prediction, which he eagerly sought during the period when another of his students was researching cluster analysis techniques using meteorological data. I could not have completed, much less begun, this work without Dr. Varahan's help. More importantly, I value the friendship that we have built through this work, and I look forward to collaborating with him in the future on many other interesting projects.

My committee has always supplied a great deal of inspiration, encouragement, and guidance throughout this process. I appreciate the discussions on the topic of neural networks during a course taught by Dr. Mark Morrissey, several of the ideas raised there inspire this work as well as other research projects. Thanks to Dr. Mike Richman for his patient support and for sharing his knowledge of statistics with me. I also wish to thank

Dr. Kelvin Droegemeier for his enthusiastic support, his brilliance and hard work is an inspriation.

I am particularly grateful to several people who provided software and data for this project. Professor Daniel Wilks, Cornell University, kindly provided the software for maximum likelihood estimation of the parameters of the gamma distribution. Dr. Ying Lin of the Enviromental Modeling Center provided archived rainfall analysis data. Dr. Ahmed Alhamed provided cluster analysis software. Other software for solving non-linear least squares and matrix inversion was obtained from the Netlib.org repository. The GSLIB software library was used to compute geostatistical measures. The student version of MATLAB R11 was used for the bulk of the computation and visualization in this work.

Thanks to my friends for supporting me over these seven years. We've known for a long time that Norman is a wonderful place to live and work, and this is mainly due to the great friends that live and work here. Thanks to Dr. Kim Elmore and Dr. Lou Wicker for getting me to run (if you can call what we do *running*) and for lending an ear during many conversations we've had during those runs. Thanks to the Kolar family, Dr. Randy Kolar for encouraging me to continue in school and for inviting me to play basketball at the fieldhouse, and Mrs. Maria Kolar for being such a great friend and volleyball teammate. To the great teachers at St. Joesph's Early Childhood Development Center, thank you for taking such good care of our boys, Andrew and Maclain. In particular, thanks to Mrs. Aimee Bradley for your friendship and for making this wonderful center better and better each day. Thanks also to Dr. Jeff Trapp and Dr. Sonia Lasher-Trapp for your friendship and encouragement over all of these years.

Thank you to my parents, Bob and Bonnie, for their support and love over all of

these years. My parents always provided support for my explorations of a wide variety of interests while growing up. Dad, you always taught me to do things right the first time and to give my best effort. Mom, you taught me to care for others and that a smile was the best way to make friends. The appreciation that I have for all that you both gave me grows more and more each day.

Finally, and most importantly, I would like to thank my wife, Melissa, for her loving support, patience, thoughtfulness, understanding, and for deciding that it was time for us to move back to Norman to go to grad school. I couldn't have made it without you. To my boys, Drew and Maclain, you bring light to my life each and every day. Daddy has finally finished writing his book!

Table of Contents

List of Tables

List of Figures

Abstract

A general, completely automated procedure for classifying rainfall systems is developed. The technique is flexible and universally applicable, in that any rainfall system can be classified regardless of size, location, time of day or year, degree of organization, etc. The knowledge obtained from previous research was used to develop a relatively straightforward and unique classification system. To test the performance of the method, results were validated against a subjective classification based upon objective criteria. From an independent random sample, the automated classification system accurately placed events into stratiform, linear, and cellular classes 85% of the time.

Chapter 1

Introduction

1.1 Motivation

Classification is the process of systematically placing individual entities into categories or classes, based upon the similarity of an item to other members of a category. When considering rainfall (more generally, precipitation) systems, one is faced with a wide spectrum of *entities*, or phenomena. Several classes of rainfall systems have been previously defined, some based upon the underlying processes that produced the rainfall, such as the general parent classes of *convective* and *stratiform* (Houghton 1968). Other general classes are based upon the space and time scales associated with each system, such as *synoptic* and *mesoscale* (Austin and Houze 1972; Orlanski 1975). Sub-classes of these range from ordinary air-mass thunderstorms (Byers and Braham 1949) to mesoscale convective systems (MCSs, Zipser 1982) to mesoscale bands embedded within synoptic-scale circulations (Hobbs 1978). In many cases, the classes are delineated by various characteristics of the spatial patterns of rainfall. In particular, rainfall systems have been classified using characteristics related to the intensity, intermittancy, shape, structure, continuity, and organization of the rainfall amounts.

Rainfall systems have been classified for a variety of purposes. Numerous climatological studies (e.g., Austin and Houze 1972; Bluestein and Jain 1985; Johns and Hirt 1987; Houze et al. 1990; Blanchard 1990; Geerts 1996; Parker and Johnson 2000) based at least partially upon radar reflectivity data have examined the characteristics and behavior of various types of mesoscale precipitation systems. These studies were

motivated by the desire to increase understanding, and therefore improve forecasting of mesoscale convective precipitation systems. As a result of this research, conceptual models of various rainfall systems have been built. These models help in understanding the critical relationship between the mode of convection and the types of severe weather that may occur (e.g., Johns and Doswell 1990; Edwards et al. 2002). For example, Houze et al. (1990) found that MCSs classified as *moderately* or *weakly* similar to the archetypical "leading line-trailing stratiform" system with the most intense cells located along the southern portion of the line (*asymmetric*) were associated with the greatest number of severe weather reports, primarily tornadoes and damaging hail. Systems that were classified as *strongly* similar to the leading line-trailing stratiform category were mainly associated with flooding. Systems that were less organized (ironically classified by Houze et al. (1990) as "unclassifiable") were more often associated with severe hail.

Other more basic climatological studies of MCSs have focussed on the contribution that these systems make to the wet-season rainfall over the U.S. (Fritsch et al. 1986) or the tropics (Mohr et al. 1999). For example, Fritsch et al. (1986) found that 30-70% of the warm-season rainfall was produced by MCSs. Mohr et al. (1999) classified convective rainfall systems based upon satellite data in the microwave channel (85-GHz) and found that MCSs contributed 70-80% of the total wet-season rainfall in the tropics.

Beyond climatological studies, some other applications of rainfall classification involve forecasting precipitation systems directly. For example, one might want to track individual storms for short-term forecasting purposes or for use in a weather-related decision support system (e.g., Kessler 1966; Dixon and Wiener 1993; Peak and Tag 1994; Johnson et al. 1998; Wilson et al. 1998; Lakshmanan 2001). This might involve an expert

system that assimilates a large volume of data and automatically returns some form of interpretation of the data in real-time, speeding up the data analysis process so the human forecaster can concentrate on the decision-making task at hand. Most of these "nowcasting" tools focus on individual thunderstorm cells that may be located within a larger rainfall system.

Other researchers have been motivated by the desire to estimate vertical latent heating profiles or improve rainfall estimation from remote sensing (e.g., Steiner et al. 1995; Yuter and Houze 1997; Biggerstaff and Listemaa 2000; Rao et al. 2001). These classification schemes subdivide a rainfall system into convective and stratiform regions at the pixel-level. The motivation for this type of "micro-classification" is that convective and stratiform precipitation regimes are caused by vertical motions of greatly differing magnitudes. Houghton (1968) defines stratifrom precipitation as that related to vertical motions much less than the fall speed of snow, therefore the precipitation particles must grow as they fall through the cloud, primarily via aggregation. The stronger vertical motions associated with convective precipitation allow for different growth processes (accretion) to dominate. Since the precipitation growth mechanisms are so different, the vertical distribution of latent heating must also be different. In addition, the convective and stratifrom drop size distributions will likely be different, therefore different Z-R relations would be required for accurate rainfall estimation. In addition, by treating convective and stratiform regions differently, this type of classification of rainfall can improve data assimilation systems that utilize precipitation information (e.g., Zupanski and Mesinger 1995; Rogers et al. 2001).

There are other potential applications for a rainfall system classication procedure

that have not yet been developed. For instance, Elmore et al. (2002) showed that an ensemble of cloud-resolving numerical model forecasts produced skillful forecasts of storm lifetimes. An automated rainfall system classification system could be used to analyze a large set of high-resolution forecasts, providing meaningful infomation on the range of possible rainfall systems that were predicted by the ensemble members. Forecast verification and predictability studies may also benefit from such a classification system. For example, Anthes (1983) argued for expanding verification information to include the validation of the "realism" of a forecast. One specific method that Anthes (1983) suggested was to verify the characteristics of significant meteorological phenomena. Along these lines, several "object-oriented" or "feature-specific" approaches to verification have been attempted or proposed (Somerville 1977; Williamson 1981; Neilley 1993; Smith and Mullen 1993; Weygandt and Seaman 1994; Baldwin et al. 2001). In order to accomplish the task of verifying significant meteorological phenomena, an automated system for identifying, characterizing, and classifying such phenomena is required. Rainfall systems are certainly an excellent candidate for this type of verification technique. For example, information on errors of displacement, amplitude, orientation, convective mode, etc., related to specific classes of MCSs found in numerical guidance would be quite useful for operational forecasters, such as those at the Storm Prediction Center (Greg Dial 2003, personal communication). The development of a verification system of this kind is the primary motivation for this work.

The implementation of a national weather radar network along with real-time hourly raingage data has fostered the development of high-resolution hourly estimates of precipitation (Baldwin and Mitchell 1998). The existence of a national mosaic of high-

resolution rainfall data allows for the identification of rainfall systems across the lower 48 states over a wide ranges of scales, as well as the development of automated classification systems. The availability of several years of data also allows for comprehensive studies of the climatology of rainfall systems. The motivation for developing an automated procedure for rainfall system classification certainly exists. The next section will examine previous research related to rainfall system classification.

1.2 Previous work

In the previous section, various reasons for classifying rainfall systems were discussed. The manner in which rainfall systems have been classified in previous work will be detailed in this section. There are three basic methods that have been employed in the previous research to locate and identify specific classes of rainfall systems within meteorological data; subjective (using objective criteria), threshold-related, and agglomerative methods.

Several researchers have subjectively analyzed a relatively large number of precipitation systems for classification purposes. For example, Austin and Houze (1972) analyzed radar data from three radars in New England and classified rainfall systems based upon their size and intensity. They established four classes of systems: synoptic areas, large mesoscale areas, small mesoscale areas, and cells. They also described the characteristics of each class, including the relative contribution of the total rainfall from each class. In their conclusions, they comment that "clearly it would be desirable to find a more objective mode of defining and identifying mesoscale precipitation areas...if their characteristics could be analyzed by computer techniques, much more data could be handled, and more comprehensive statistics would emerge."

More comprehensive studies of MCSs were performed by Bluestein and Jain (1985), Bluestein et al. (1987), Blanchard (1990), Houze et al. (1990), Geerts (1998), and Parker and Johnson (2000). These studies examined MCSs subjectively via visual analysis of radar images. The systems were classified based upon how they developed over time and how closely they matched archetypical examples of MCS classes. For most of these, an objective criteria was used to define a *line* of convection, which was a length to width ratio of at least 5, at least 50km long and less than 50km wide (Bluestein and Jain 1985). In addition, most used an objective criteria for delineating the convective and stratiform regions. For example, Geerts (1998) used the 20dBZ threshold to delineate the convective region as long as there was a maximum reflectivity of at least 40dBZ embedded within it. Each study defined slightly different classes of MCSs, although there were a few classes in common worth noting. The *leading-line/trailing-stratiform* class (Houze et al. 1990) demonstrated the highest degree of linear organization and was the focus of several studies. On the opposite end of the alignment spectrum, Houze et al. (1990) established an *unclassifiable* class, similar to the *chaotic* class defined by Blanchard (1990). The common characteristic among these studies was the use of visual inspection of the radar images as the primary analysis tool. Since the goal of this work is developing an automated classification system, the subjective classification technique will obviously not be appropriate for this study.

The next group of classification tools uses some form of threshold in radar or satellite data to classify regions within rainfall systems. The specific class of thunderstorm cells have been located, classified, and tracked via reflectivity thresholds within weather-related decision support systems (e.g., Dixon and Weiner 1993; Johnson et

al. 1998). These routines are very class-specific, in fact, Johnson et al. (1998) does not recommend using their technique for larger rain systems. Peak and Tag (1994) use hierarchical threshold segmentation as a feature identification tool, which is a necessary step prior to classification. A hierarchy of "objects" within a satellite image is produced though the use of a set of thresholds. A neural network is used to train the system to decide when to subdivide a region and when not to. The resulting segmentation will depend on the expert used to train the network. The characteristics of the satellite image patterns that were used were size, boundary length, and "complexity" (related to fractal dimension) which is ratio of size to boundary length.

Steiner et al. (1995) apply a sort of adaptive thresholding technique to separate convective and stratiform regions within a rainfall system. The reflectivity value at a point is compared to the "background" value, which is an average of the reflectivities within a small radius of the point. If the reflecitivity is significantly higher than the background, or if it is > 40dBZ, the point is considered convective. This is referred to by Biggerstaff and Listemaa (2000) as a "peakedness method." The classification methods described by Yuter and Houze (1997) and Biggerstaff and Listemaa (2000) are based upon modified versions of the Steiner et al. (1995) routine. In addition, the analysis of the microwave channel of satellite data by Mohr and Zipser (1996) operates in a similar way, except with brightness temperatures. As mentioned in the previous section, the motivations for these types of classification routines are to estimate vertical latent heating profiles or modify Z-R relations to improve rainfall estimation. The physical basis for this type of "micro-classification" is that convective and stratiform precipitation are caused by vertical motions of greatly differing magnitudes (Houghton 1968), therefore the

associated vertical latent heating profiles should be different.

On the other hand, this work takes a "macro-classification" approach to classify entire rainfall systems, as opposed to the "micro-classification" of Steiner et al. (1995). The definition of a system is: "an organized integrated whole made up of diverse but interrelated and interdependent parts". In a typical MCS, the convective and stratiform regions are interrelated and interdependent parts of a system. The stratiform region would not exist if the convection was not there to transport ice crystals away from the strong updraft. In some cases, evaporation of rainfall within the stratiform region helps to enhance the mesoscale circulation that allows the convection to propagate, keeping the entire system alive (e.g., Zhang and Gao 1989). In this work, a mesoscale convective system of this type will be considered a convective system and a complete entity and not sub-divided into convective/stratiform regions. Therefore, the Steiner et al. (1995) and related algorithms will not be used in this work.

The final method of rainfall system classification that will be discussed is the *agglomerative* or cluster analysis technique. An image processing technique, an agglomerative region-growing algorithm operates by grouping together portions of an image with similar characteristics. A recent example of an agglomerative routine for processing weather-related images is provided by Lakshmanan (2001). Here, the *texture* of the image, represented by a vector of local statistical measures in the neighborhood of each pixel, is analyzed. Pixels that are similar in terms of their texture and spatial location are grouped together to form a set of clusters. This technique produces a hierarchy of objects over a range of spatial scales, where the number of clusters/objects is cut in half at each step in order to reach the next level of hierarchy. At some point, a subjective

decision as to the desired number of clusters must be made. This method does produce favorable results for weather radar and satellite images, and is currently being tested for its potential use in radar feature tracking algorithms at the National Severe Storms Laboratory (Lakshmanan 2003). However, it was not selected for this work since a subjective decision on the number of clusters or objects to keep for each image is required. While the selection of an acceptable threshold that would produce satisfactory results for any given rainfall image might be possible to obtain, perhaps via training of a neural network such as in Peak and Tag (1994), this would likely require a great deal of effort and tuning of the technique.

As previously noted, there are a wide variety of applications for an automated rainfall system classification procedure. Many of the previous studies related to automated rainfall classification perform what has been defined as a "micro-classification" similar to categorizing parts of an entity, using a microscope. While there are certainly valid reasons for executing a classification of this kind, those are not the primary focus of this work. Instead, a broader "macro-classification" approach to the classification problem is followed in this work, considering classes of rainfall systems as separate types of entities or species of animals. For example, linear and chaotic MCSs are different species of the same family, they are associated with different types of mesoscale circulations, different environmental conditions, and tend to produce different types of severe weather (Houze et al. 1990). The differences between these two approaches to studying the morphology of rainfall systems is similar to the difference between *anatomy*, the study of parts of the body, and *taxonomy*, the classification of organisms in an ordered system. Of the past research described previously, only the subjective approaches to

classification followed this "macro-classification" philosophy. Therefore, in order to realize the goal of performing a more general automated classification, a unique classification procedure must be developed.

The specific objective of this work is to develop a general, completely automated procedure for classifying rainfall systems. A desirable property of such a technique is that it will be universally applicable, that is, any rainfall system can be classified regardless of size, location, time of day or year, degree of organization, etc. The knowledge obtained from previous research will be synthesized while a relatively simple, yet unique classification system is developed. To ensure that the method performs well, results of this technique will be validated against subjective classes based upon objective criteria, similar to those described in Bluestein and Jain (1985). The process of developing the automated classification system will be described in the next section.

1.3 Statement of work

Since there are many different applications for an automated rainfall system procedure, it is likely impossible to develop a universal method that will satisfy every user. The primary users that are the focus of this work are those interested in the classification of rainfall systems in their entirety. This section outlines the general framework that will be followed to perform an automated rainfall system classification. For this, we naturally turn to the discipline of data mining.

This work will take advantage of the well-established techniques found in the field of *knowledge discovery in databases* (KDD, Fayyad et al 1996) and *data mining* (Adriaans and Zantinge 1996). The concept of KDD is defined by Fayyad et al (1996) as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately

understandable patterns in data." Data mining is a specific part of the KDD process, referring to the application of algorithms for extracting patterns from data, and classification is but one of several specific data mining tasks. KDD is a multi-disciplinary field with roots in machine learning, expert systems, databases, statistics, and data visualization. The general steps of the KDD process are listed in table 1-1 (Fayyad et al 1996).

Table 1-1: General steps of the KDD process (from Fayyad et al. 1996).

1. Develop an understanding of the application and the goals of the end-user.
2. Create a target data set.
3. Preprocess the data set; remove noise and outliers and decide how to treat missing data.
4. Data reduction and projection; find useful features that represent the data with a smaller number of variables or dimensions.
5. Choose the data mining task; classification, regression, clustering, change detection, etc.
6. Choose the data mining algorithm.
7. Execute the data mining.
8. Interpret the mined patterns, possibly repeating previous steps as a result.
9. Consolidate discovered knowledge.

These general steps provide the broad outline that has been followed in developing an automated classification scheme for this research. The first step involves developing an understanding of the application and the goals of the end-user. This understanding has been established in this chapter. Here, the goal is to classify rainfall systems. For several reasons, rainfall has been selected as the variable for analysis. The spatial patterns found in precipitation fields often represent important and significantly different meteorological phenomena. There are several potential applications for an automated rainfall

classification system, including climatological studies, verification, data assimilation, feature tracking, and forecasting. The next steps in the KDD process involve the selection of a target data set and preprocessing the data. A relatively large dataset should be used, one that is richly populated with a variety of interesting and important phenomena that span a large portion of the entire range of possible events. Here, forty-eight cases from a high-resolution precipitation analysis produced operationally by the National Centers for Environmental Prediction (NCEP) are used to create the target data set. The use of operationally available data will make this work more relevant and allow for faster implementation into an operational forecasting environment.

Once the target data set has been selected and processed, the next steps in the KDD process involve data reduction and data mining. In this case, the data mining task is classification. Data reduction addresses the methods used to extract features of a relatively small dimension within the large-dimensional dataset that allow for proper classification of objects. Here, statistically-based attributes will be used exclusively. The determination of useful attributes that possess good discrimination and classification properties represents the most substantial portion of this work. The selection of attributes that characterize rainfall systems feeds off the lessons learned in previous research. For example, to attempt to separate convective and stratiform systems, histogram analysis was performed. Not only does this provide information on the overall intensity of the rain, but the "peakedness" (Biggerstaff and Listemaa 2000) as well. For example, convective systems will contain a relatively high number of heavy rainfall observations, which will be represented in the histogram by a distribution with a "thick tail." In order to separate convective events into linear and chaotic (cellular) classes, estimates of the degree of

linear organization of the rainfall system will be obtained via geostatistical measures. Once a useful set of attributes has been determined, algorithms will be developed to automate the identification of rainfall systems and extract the attributes associated with them, and automate the classification of each system. These procedures are related to the final steps of the KDD process, involved with selecting and executing the data mining algorithm, interpreting the results, perhaps repeating the previous steps, and finally consolidating the discovered knowledge. The automated classification procedure will be used to examine rainfall systems over the course of an entire year. The summary statistics of these data will be considered, and the classification method will be validated by an independent, representative sample.

In summary, to develop an automated rainfall system classification procedure, the KDD process has been followed in this work. A brief outline of the remainder of this dissertation folllows. A detailed description of the mathematical tools used throughout this work will be provided in chapter 2. Results from classification experiments involving histogram-related attributes are discussed in chapter 3. Experimental classification results using summary measures of geostatistics will be presented in chapter 4. A detailed description of procedures developed to identify, analyze, and classify rainfall systems in a completely automated fashion will be offered in chapter 5. In addition, summary statistics obtained from analysis of an entire year of rainfall data as well as an independent validation of the classification results will also be documented. Finally, concluding remarks and a discussion of future work related to this automated classification system will be provided in chapter 6.

Chapter 2

Mathematical tools

2.1 Introduction

There is a long history in the meteorological literature on the use of pattern recognition/classification techniques for a wide variety of applications. A few examples include: detecting patterns in atmospheric soundings in the near-tornado environment (Schaefer and Livingston, 1988), distinguishing polar ice cap cloud cover types in satellite data (Ebert 1987), locating frontal zones in numerical model output (Fine and Fraser 1990), classifying or clustering ensemble forecast data (Eckert et al 1996, Alahmed et al. 2002), locating significant circulation features in radar data (Weckwerth et al 1997, Stumpf et al 1998), and pattern analysis of climate data (Gong and Richman, 1995). There are common threads among these studies. Each begins with a complex data set of fairly large dimension. The goal of each is develop an objective method to extract useful information found within that large database. These goals are consistent with the general disciplines of *data mining* and *knowledge discovery in databases* (KDD) (Adriaans and Zantinge 1996, Fayyad et al 1996).

In the previous chapter, the general process for developing a rainfall classification system was outlined. A key aspect to this work involves comparing the results of an automated classification using trial attributes with a subjective classification. If the two classifications agree, the trial attributes will then be considered useful. Also, an automated rainfall system identification and analysis system is required. Several steps within this process require the use of specific mathematical tools and algorithms. The purpose of this

chapter is to introduce those concepts and tools. The KDD process (Table 1-1) defines the general framework that has been followed in developing a rainfall pattern classification technique. For example, step #2 (Table 1-1) involves creating a target data set. While the details of the type of rainfall observations that were used in the data set will be left to the next chapter, the definition of terms related to the target data set will follow.

2.2 Data Matrix

Classification involves placing subjects into groups based upon their similarity to other individuals found within a particular class. The similarity between individuals is determined by some function of the characteristics asssociated with each object. Analysis of a set of objects to be classified and the attributes associated with them is typically performed through the use of a *data matrix*. Define $X = [x_{ij}]$, $1 \le i \le m$, $1 \le j \le n$ as a data matrix. In such a matrix, each column represents an *object*, and each row represents an *attribute*. Therefore, x_{ij} represents the i^{th} attribute of the j^{th} object. In the context of meteorological data, objects might also be called events, phenomena, features, realizations, or cases. Attributes refer to the observations, parameters, characteristics, or measurements that describe various aspects of the objects of interest. For example, in ensemble forecasting (Alhamed et al 2002), objects would be forecasts from specific members of an ensemble, and attributes would be the values of the predicted variable from each member. Other examples of objects would include two-dimensional fields of heights on constant pressure surfaces, vertical soundings of temperature, time-series of wind speed, or snowdepth at a specified location. Each object can be described by a vector of m-dimension, where m is the number of attributes. Meteorological data consisting of multiple spatial dimensions (such as a 2-D field of temperature) is often visualized in a grid-

ded form, where, for example, the 1st dimension represents the east-west spatial direction and the 2nd dimension represents the north-south spatial direction. When placed in a data matrix, such data will be converted into a *m*-dimension vector by proper row-major or column-major storage methods. Additionally, each row of the data matrix can also be viewed as a vector of *n*-dimension, where *n* is the number of objects. A *target data set* is a data matrix containing some set of trial objects and attributes.

The choices of methods of comparing the similarity of objects will depend on the manner in which values are assigned to the attributes that describe each object. The rules for distinguishing among different attribute values are known as the *scales* of measurement. There are four scales for data that analysts typically use to establish the meaning of comparisons between attribute values. For example, when attributes can only be determined to be equal or unequal, this is called the *nominal scale*. Examples of nominal scale values include colors, binary (true or false) variables, and gender. When the scale allows for the ordering of attribute values, it is known as an *ordinal scale*. The ordinal scale further distinguishes the equality/inequality relation of the nominal scale, by describing values as greater or smaller than other variables. Examples of ordinal scale variables include letter grades (A through F) or opinion ratings on a scale of 1 to 10. The nominal and ordinal scales are also known as *qualitative* scales, since they only allow comparisons in a qualitative sense. For meaningful *quantitative* measures of the difference between two values, the *interval scale* is used. Finally, the *ratio scale* also allows for meaningful comparisons of the ratio of two values. For example, temperature values in degrees Celsius follow the interval scale, since the difference between two values is meaningful. Temperature in degrees Celsius does not follow the ratio scale since the ratio of two values is not

meaningful. An example of a ratio scale variable is temperature in degrees Kelvin. As mentioned earlier, the choices of methods of comparing objects containing nominal scale attributes will differ greatly from those containing ratio scale attributes. One must take care in comparing objects with attributes using a variety of different scales. Before measures of similarity can be computed, attributes should be converted into one type of scale. In most meteorological applications, ratio scale variables are available and widely used. In this work, all attributes reside in the ratio scale.

The units used in the assignment of various attributes can strongly impact measures of similarity among objects. For example, differences in rainfall amounts cast in terms of millimeters will appear to be 25.4 times as large as those given in units of inches. One can arbitrarily inflate the importance of a particular attribute by simply recasting it in terms of some other unit of measurement. To avoid this arbitrary affect, attributes can be transformed by converting them into dimensionless numbers. Transforming the data matrix allows each attribute to contribute more equally to the overall measure of similarity (Romesburg 1984). There are several choices of data matrix transformation methods, each transfers the information in the data matrix X into a new data matrix Z that is the same size as X. The transformed data matrix Z can then be use in the analysis of the data. For example, a simple type of transformation involves *centering* the data, that is, subtracting the mean value from each attribute.

$$z_{ij} = x_{ij} - \frac{1}{n}\sum_{j=1}^{n} x_{ij} \tag{2.1}$$

The centered data retain their units, however, and can be thought of as anomalies or perturbations. The data matrix can be *normalized* by dividing each centered attribute by its

standard deviation.

$$z_{ij} = \frac{x_{ij} - \frac{1}{n}\sum_{j=1}^{n} x_{ij}}{\sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(x_{ij} - \frac{1}{n}\sum_{j=1}^{n} x_{ij}\right)^2}} \tag{2.2}$$

Normalization produces attributes which have zero mean and unit variance. Other methods of transformation include scaling by the maximum or minimum value of each attribute, or by the maximum of the entire data matrix. In general, transformation of the data matrix is an optional step in data analysis. Analysis techniques such as cluster analysis or principal component analysis can be performed on either the original or transformed data matrices.

2.3 Cluster Analysis Overview

Once the data matrix has been populated with objects of interest and attributes that describe those objects, the next steps in the KDD process (Table 1-9) involve choosing the data mining task and algorithm. For this work, the data mining task is classification, and the tool that will be used to accomplish this task is cluster analysis. The following paragraphs will describe various cluster analysis methods, and discuss the specific algorithm that was chosen to perform the classification task. The bulk of description found in this section is adapted from the excellent summary of clustering methods found in the appendices of Alhamed et al (2002) as well as Alhamed (2000).

Cluster analysis is a descriptive statistical method of analyzing the similarity of objects in a data matrix. There are two classes of cluster analysis techniques, *hierarchical* and *partitional* methods. Hierarchical methods discover the relationships between groups

of objects by constructing a hierarchy of clusters. This can be visualized as a tree (Figure 2-1), where the ends of each individual branch represents each individual object (or a clus-

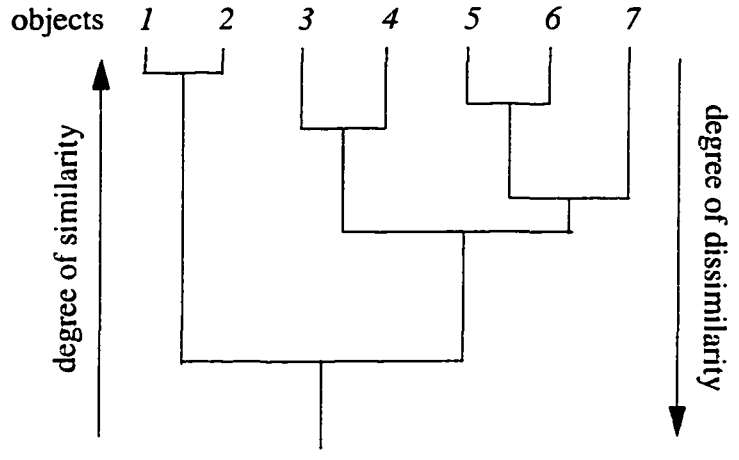


Figure 2-1: Hypothetical hierarchical clustering tree, also known as a *dendrogram.*

ter containing a single member), and as branches come together on the tree, objects are grouped together to form clusters. As you move further down the tree, the degree of similarity between clusters becomes less and less while clusters grow to contain more members. Eventually all objects are combined into a single cluster, which could be thought of as the trunk of the tree. Hierarchical methods are useful when the true number of clusters is not known. On the other hand, partitional (non-hierarchical) methods generate a single partition of objects into a pre-determined number of clusters. Ideally, objects grouped together to form a cluster will appear quite similar, and objects found in different clusters will appear quite different. The various clustering algorithms each follow this general idea in attempting to form ideal clusters. However, the manner in which this idea is implemented will differ for different algorithms.

2.4 Similarity measures

In order to discover the relationships between differing groups of objects, the similarity among objects must be computed. Measures of similarity between objects, also known as *resemblance coefficients* (Romesburg 1984), establish the degree of similarity or dissimilarity between two objects. These pair-wise similarity measures are computed for each pair of objects and are arranged in the form of a *similarity matrix* where the $(i,j)^{th}$ element indicates the resemblance coefficient between objects i and j. A similarity matrix is a square, symmetric, n x n matrix, where n is the number of objects. This matrix is symmetric since the similarity of objects i and j is identical to the similarity between objects j and i. A resemblance coefficient can be either a measure of similarity or dissimilarity. For a similarity measure, the larger the value, the more similar the objects will be. The opposite is true for a dissimilarity measure, larger values indicate less similar objects. Cluster analysis algorithms often operate directly with the similarity matrix, not the original data matrix, therefore the algorithms must properly account for the type of resemblance coefficients used.

A commonly used measure of dissimilarity is the *Euclidean distance*, denoted as d_{jk}, which measures the distance between two objects, j and k. Using the data matrix notation defined in section 2.2, the *Euclidean distance* is defined as:

$$d_{jk} = \sqrt{\sum_{i=1}^{m} (x_{ij} - x_{ik})^2} = \|x_{:k} - x_{:j}\|_2 \tag{2.3}$$

If we denote the j^{th} object as $x_{:j}$ this can also be rewritten as the norm of the difference vector between $x_{:j}$ and $x_{:k}$, where $\|x\|_2$ indicates the Euclidean norm or 2-norm of the vector x. Geometrically, the Euclidean distance is simply the length of the vector connecting

two points in space. Another distance measure is called the *Manhattan* or "taxi-cab" distance or 1-norm, h_{jk},and is defined as:

$$h_{jk} = \sum_{i=1}^{m} |x_{ij} - x_{ik}| = \|x_{:k} - x_{:j}\|_1 \tag{2.4}$$

It is known as the Manhattan distance since it is the distance that one would travel if the path between points was taken along the city blocks of a major city, while the Euclidean distance is more of a straight-line or "as-the-crow-flies" distance. The *Chebyshev* distance, p_{jk},also known as the ∞-norm, is the maximum absolute value of the difference of all attributes:

$$p_{jk} = \underset{1 \le i \le m}{MAX} |x_{ij} - x_{ik}| = \|x_{:j} - x_{:k}\|_\infty \tag{2.5}$$

The generalized version of the previous three distance measures is known as the *Minkowski* distance, m_{jk},(for $p>0$):

$$m_{jk} = \sqrt[p]{\sum_{i=1}^{m} |x_{ij} - x_{ik}|^p} = \|x_{:j} - x_{:k}\|_p \tag{2.6}$$

which equals the Euclidean distance for p=2, the Manhattan distance for p=1, and the Chebyshev distance for p=∞. Another distance measure is the *Energy norm* or *generalized Euclidean distance* d_{Ajk}

$$d_{Ajk} = [(x_{:j} - x_{:k})^T A(x_{:j} - x_k)]^{1/2} \tag{2.7}$$

which is a weighted inner product of the difference between two objects (x^T will be used in this work to denote the transpose of a vector or matrix). For this distance measure to

meet the qualifications of a distance metric, **A** must be a positive-definite symmetric matrix. The use of the weight matrix allows one to weigh certain attributes more heavily than others, or to account for differences in the units between attributes. When the identity matrix is used for the weight matrix **A**, the familiar Euclidean distance d_{jk} (eq. 2.3) is obtained. A more specific example of this type of measure is the *Mahalanobis distance*, where **A** is replaced by the inverse of the covariance matrix S^{-1} of the data matrix X. This distance measure has the advantage of taking into account the covariance between variables.

A commonly used measure of similarity is the *correlation coefficient*, r_{jk}, defined as:

$$r_{jk} = \frac{\sum_{i=1}^{m} x_{ij}x_{ik} - \frac{1}{m}\left(\sum_{i=1}^{m} x_{ij}\right)\left(\sum_{i=1}^{m} x_{ik}\right)}{\left[\sum_{i=1}^{m} x_{ij}^2 - \frac{1}{m}\left(\sum_{i=1}^{m} x_{ij}\right)^2\right]^{1/2}\left[\sum_{i=1}^{m} x_{ik}^2 - \frac{1}{m}\left(\sum_{i=1}^{m} x_{ik}\right)^2\right]^{1/2}} \tag{2.8}$$

This is the familiar Pearson product-moment correlation between objects j and k. Geometrically, the correlation coefficient is the cosine of the angle between centered vectors, where the mean of all attributes for each object is subtracted from each object (centered by column mean). This is unlike equation 2.1 above, where the centering was done by row mean (the mean of all objects for each attribute was subtracted). In addition, the *cosine coefficient*, c_{jk}:

$$c_{jk} = \frac{\sum_{i=1}^{m} x_{ij}x_{ik}}{\left(\sum_{i=1}^{m} x_{ij}^2\right)^{1/2}\left(\sum_{i=1}^{m} x_{ik}^2\right)^{1/2}} \tag{2.9}$$

is another measure of similarity. Geometrically, the cosine coefficient is simply the cosine of the angle between the unit vectors in the same direction as objects j and k.

2.5 Hierarchical cluster analysis

Conceptually, there are two approaches that can be used to accomplish hierarchical cluster analysis: *agglomerative* and *divisive*. Referring back to figure 2-1, this difference between these two approaches can be visualized as moving through the tree either in a top-down or a bottom-up direction. In the case of agglomerative clustering, each object is initially placed in its own cluster and the algorithm joins similar clusters together. The remaining clusters gradually contain more and more objects, until finally one cluster is formed that contains all of the objects in the data matrix. On the other hand, the divisive approach begins with a single cluster containing all objects and subdivides dissimilar groups until eventually each cluster contains a single object. In this work, an agglomerative method is used.

Basically, agglomerative clustering algorithms proceed as follows.

Step 1: Assemble a similarity matrix $S = [s_{ij}]_{n \times n}$ where resemblance coefficients are computed for all possible pairs of objects in the data matrix. Element s_{ij} represents the similarity/dissimilarity measure between objects i and j. Since this matrix is symmetric, only the lower triangular part of the matrix needs to be stored.

Step 2: Construct n clusters by placing each object in an individual cluster. At this point, cluster C_i contains only object i.

Step 3: Find the most similar pair of clusters in the similarity matrix. Let C_i and C_j be the most similar pair where $i > j$.

Step 4: Merge the two clusters and reduce the number of clusters by 1. Label the new cluster as C_i and update the similarity matrix appropriately to account for the modified similarities between this new cluster and all other existing clusters. Remove the row and column of S corresponding to the cluster C_j.

Step 5: Repeat steps 3 and 4 until only one cluster remains.

At each iteration, a record of the objects found within each cluster and the level of similarity found when clusters were merged is maintained. This allows for visualization of the tree structure (as in figure 2-1) where analysts can quickly examine the relationships between groups of objects. Within this basic framework, different methods could be used at each step, thereby developing different specific clustering algorithms. For instance, in step 3, the definition of the most similar pair will depend on whether similarity or dissimilarity measures were used in the similarity matrix. There are several possible methods for updating the similarity matrix in step 4. One of the more common methods is called the single linkage (SLINK) method, where the similarity between clusters is replaced by the most similar value of the resemblance measures between all elements of the two clusters. The complete linkage (CLINK) method takes a similar approach, except using the least similar value of the similarity measures between the elements of the two clusters. One can also use the average value of all of the similarity measures between the elements of the two clusters, or average linkage.

The agglomerative method used in this work is based on the variance conservation property for a group of objects, called *Ward's method* (Ward 1963) or the *minimum variance* method. The total variance among all of the objects found in the data matrix will be conserved regardless of how these objects are grouped into clusters. The total variance can also be divided into two components, the inter-cluster (between cluster) variance and the intra-cluster (within cluster) variance. The inter-cluster variance is defined as the scatter between the centroids of the clusters, while the intra-cluster variance is defined as the overall scatter between objects within each individual cluster. Since the total variance is constant, if the inter-cluster variance is relatively large, the intra-cluster variance must be relatively small, and vice versa. The criteria for merging clusters is minimizing the within-cluster variance, and therefore maximizing the between-cluster variance. This forces the objects found within a cluster to be similar while keeping the clusters as distinct as possible. Ward's clustering algorithm proceeds as follows:

Step 1: Assign each object to separate clusters, each of which contains only one object. The intra-cluster variance is zero at this point.

Step 2: Computed the intra-cluster variance for every possible merger of two clusters.

Step 3: Create a new cluster by merging the two clusters that produce the smallest increase in the intra-cluster variance.

Step 4: Repeat steps 2 and 3 until a single cluster containing all objects remains.

Ward's method has been found to produce good results for meteorological data in previous research (Alhamed et al. 2002). For this reason, Ward's method was chosen as the hierar-

chical cluster analysis algorithm for the entirety of this work.

2.6 Partitional cluster analysis

Although non-hierarchical cluster analysis methods were not used in this work, a brief description is included for completeness. Hierarchical cluster analysis results in a branching sequence of clusters, organized in order of the degree of similarity. This kind of analysis is useful when the overall number of clusters expected in the data is not known in advance; i.e., the clustering hierarchy allows the analyst to visualize how the objects are organized in terms of their similarity. On the other hand, partitional cluster analysis methods are designed to group objects together into a single set of k clusters, where k is specified ahead of time or is determined via the clustering algorithm. The problem of partitional clustering can be summarized as follows: given a set of n objects, determine a partition of the objects into k clusters such that the objects within each cluster are more similar to each other than to objects in different clusters (Jain and Dubes 1988). The practical issues related to partitional clustering involve the choice of the initial partition, and the criterion used to decide if the resulting partition is optimal. The initial partition can be formed by identifying an initial set of k seed points, from which the partitions will grow. The seed points could be the first k objects in the data matrix, or k randomly selected objects, or k subjectively chosen objects, if the analyst has some expertise with the data. From the seed points, the initial partition could be created by assigning each object to the partition belonging to the nearest seed point. Another possibility is to take the results of hierarchical clustering to generate the k initial partitions. For the choice of criterion, as in the hierarchical Ward's method, the partitions that minimize the within-cluster variance is one that is most commonly used. An example of such an algorithm is known as the k-

means algorithm (Anderberg 1973). It proceeds as follows:

Step 1: Begin with an initial partition of k clusters.

Step 2: For each object, compute the distance to the centroids of every cluster. If the object does not belong to the cluster representing the nearest centroid, reassign the object and update the centroid values for those clusters affected by the move.

Step 3: Repeat step 2 until no objects are moved from one cluster to another.

This section has discussed the concepts and algorithms associated with cluster analysis. The basic structure of the data used in cluster analysis is the set of objects/attributes associated with the data of interest. The next section will discuss the mathematical concepts and tools that were used in the computation of attributes in this work.

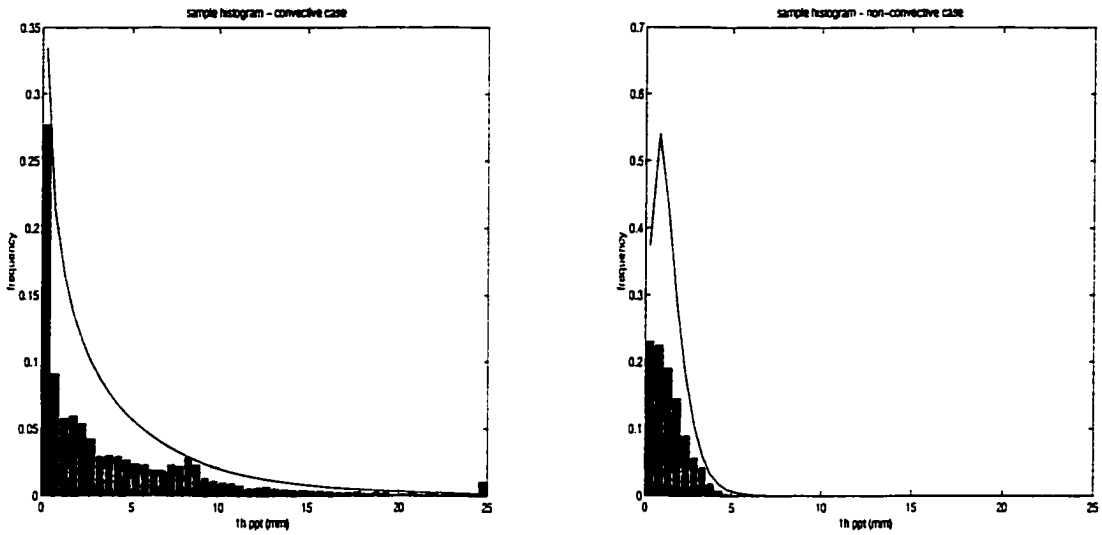
2.7 Computation of Attributes

As previously mentioned, the KDD process will be followed in order to determine useful attributes for an automated rainfall classification system. Once the goals of the end-user have been established and the target data set has been created, the next step involves data reduction and projection. In Table 1-1, data reduction is defined as finding useful features that represent the data with a smaller number of variables or dimensions. Ideally, the set of attributes should be relatively small to allow for faster and simpler analysis. In addition, attributes should measure significant and interesting aspects of the objects of interest, and also allow discrimination between significantly different phenomena. Attributes should be as simple to compute as possible, to reduce the amount of computation time required. To assist in the interpretation of the results, attributes should be conceptually easy to understand and explain to the users of the results (meteorologists). The following sections describe the mathematical concepts and tools used to derive the

attributes that were used in this work.

2.8 Gamma distribution

Data reduction involves objectively extracting useful features of a relatively small dimension within the large-dimensional dataset that allow for classification of objects. Many statistical analysis methods can also be considered to be data reduction techniques. For example, if one fits a Gaussian distribution to a sample data set, the mean and variance are all that is necessary to describe the distribution. Hence, the large-dimensional data set has been reduced to two dimensions. Following this simple idea, the parameters of a theoretical statistical distribution that fit the histogram representing the observed distribution of rainfall amounts were used as trial attributes. For example, Wilks (1989) mentions that for the Weibull distribution; $(f(x;\alpha,\beta) = (\beta/\alpha)(x/\alpha)^{\beta-1}[\exp-(x/\alpha)^{\beta}],\ x, \alpha, \beta > 0)$ "Smaller values of β [the so-called *shape* parameter] will reflect a tendency toward briefer and more predominantly convective precipitation, and larger values will indicate a greater tendency toward steadier precipitation derived from larger-scale processes." The distribution of rainfall tends to be highly positively skewed. For example, heavy rainfall is a rare event, and when large amounts of rain do occur, such as typically found intermittently in some convective systems, the resulting distribution possesses a long "tail" (Fig. 2-2a). It is also common to see widespread light rain, such as typically found in non-convective systems, resulting in a distribution that is "humped" near a low amount of rainfall with little or no "tail" (Fig. 2-2b). These characteristics limit the choices of theoretical distributions as potential models for the observed distribution. For this work, we selected the gamma distribution since it is positively skewed and non-negative, provides a reasonable representation with only two parameters (Eq. 1), and has been widely used in the meteorologi-



(a) Left panel (b) Right panel

Figure 2-2:Sample histograms of non-zero rainfall from a convective case (#1, left panel) and a non-convective case (#45, right panel) from the target data set. Curves indicate gamma distribution fit using method of moments parameter estimation described in section 2.9.1.

cal literature for the analysis of precipitation data (e.g., Wilks 1990). The probability density function (f(x)), moment generating function ($M_x(t)$), and first four moments (μ_i) of the gamma distribution are (Freund and Walpole 1987):

$$f(x;\alpha,\beta) = (x/\beta)^{\alpha-1}[\exp(-x/\beta)][\beta\Gamma(\alpha)]^{-1}, x \geq 0, \alpha, \beta > 0 \quad (2.10)$$

$$M_x(t) = (1-\beta t)^{-\alpha} \quad (2.11)$$

$$\mu_1 = \alpha\beta \quad (2.12)$$

$$\mu_2 = \alpha\beta^2(\alpha+1) \quad (2.13)$$

$$\mu_3 = \alpha\beta^3(\alpha^2+3\alpha+2) \quad (2.14)$$

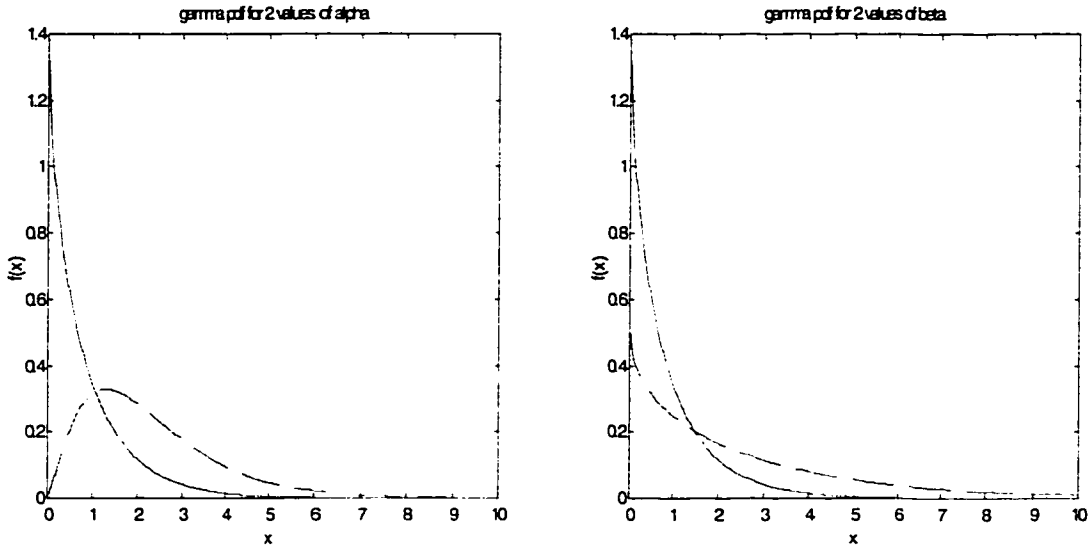
$$\mu_4 = \alpha\beta^4(\alpha^3+6\alpha^2+11\alpha+6) \quad (2.15)$$

where $\Gamma(\alpha)$ is the standard gamma function, which equates to the factorial for integer val-

ues of α.

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt \tag{2.16}$$

The parameter α is commonly referred to as the "shape" parameter and β is referred to as the "scale" parameter. Figure 2-3a shows two example probability density function curves for the gamma distribution for varying values of α. For values of the parameter α



(a) Left panel (b) Right panel

Figure 2-3:Plots of the gamma probability density function for (a, left panel) $\alpha = 0.9$, $\beta = 1.0$ (solid) and $\alpha = 2.3$, $\beta = 1.0$ (dashed), (b, right panel) $\alpha = 0.9$, $\beta = 1.0$ (solid) and $\alpha = 0.9$, $\beta = 3.0$ (dashed).

< 1.0, the distribution is skewed strongly to the right with f(x) approaching infinity 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 α the gamma distribution is similar to the Gaussian distribution. The role of the parameter β (Fig. 2b) is to "pull" the distribution to the right for larger β, increasing the frequency of larger values of x and creating a thicker tail. For smaller β, the frequency of smaller values of x is increased, creating a thinner tail and "pushing" the distribution towards the left. It seems reasonable

to expect that the shape and scale parameters might be useful attributes for describing the overall distribution of rainfall intensity. For example, since higher values of β produces a thicker tail in the distribution, one would expect to see such high β values associated with intense rainfall events. The usefulness of these attributes in discriminating between convective and non-convective rainfall events will be evaluated in Chapter 3.

2.9 Parameter estimation

In order to fit this theoretical distribution to the observed rainfall values, it is necessary to estimate the parameters of the distribution. The following sections outline various parameter estimation techniques.

2.9.1 Method of moments

In the method of moments, a set of equations are developed to estimate the number of unknown parameters found in the model. In the case of the gamma distribution, there are two unknown parameters, α and β, therefore two equations relating these to known quantities are needed. Here, these two equations are found by equating the first two computed sample moments to the population moments. For example, the population mean of the gamma distribution is $\alpha\beta$ and the sample mean is $\bar{x}$ (which is known, computed from the observed data). The population variance (related to the second moment) is $\alpha\beta^2$ and the sample variance is σ^2. Equating these sample and population values provides a set of two equations and two unknowns. This system can easily be solved to find that $\alpha=\bar{x}^2/\sigma^2$ and $\beta=\sigma^2/\bar{x}$. These parameters fit the observed mean and variance exactly, but higher-order moments are not taken into account. Wilks (1990) also points out that for smaller values of α ($\alpha < 10$), the parameter estimates resulting from the method of moments are subject to a relatively high degree of variability from one data sample to another. As dem-

onstrated in figure 2-3, small values of α correspond to strongly skewed distributions. Since it is common for distributions representing short-term rainfall amounts to be highly positively skewed, a more precise method of parameter estimation is desired.

2.9.2Maximum likelihood estimation

As the name suggests, the method of maximum likelihood estimation (MLE) seeks to maximize the *likelihood function*, which is the joint distribution of values of the unknown parameters given the observations of the random variable (denoted as $f(\theta|W)$, where θ is the vector of unknown parameters and W is the vector of size T containing the observations, the i^{th} observation is denoted by w_i). The multiplicative law of probability (Wilks 1995) states that for independent events, the joint probability is equal to multiplying all of the probabilities of the individual events.

$$f(\theta|W) = f(\theta|(w_1, w_2, w_3, \ldots, w_T)) = \prod_{i=1}^{T} f(\theta|w_i) \tag{2.17}$$

Note that independent events are uncorrelated; the occurrence of one event does not depend on the occurrence of another. Spatial rainfall data will undoubtedly violate this assumption. The likelihood function for a single observation appears identical to the probability density function. The distinction between the likelihood and probability density functions is a technical one; the probability density is a function of the observations given the fixed values of the parameters. The likelihood function is a function of the parameters given the fixed values of the observations. In the case of the independent events following the gamma distribution, the likelihood function can be written as:

$$L(\alpha, \beta; W) = \prod_{i=1}^{T} (w_i/\beta)^{\alpha-1} \frac{\exp(-w_i/\beta)}{\beta\Gamma(\alpha)} \tag{2.18}$$

Since the logarithm is a strictly increasing function, maximizing a function f is the same as maximizing $\log(f)$. Therefore it is convenient to take the logarithm of the likelihood function prior to maximization, the *log-likelihood* function becomes a sum rather than a product.

$$\Lambda(\alpha, \beta; W) = (\alpha - 1) \sum_{i=1}^{T} \ln(w_i) - \frac{1}{\beta} \sum_{i=1}^{T} w_i - T[\ln\Gamma(\alpha) + \alpha \ln(\beta)] \tag{2.19}$$

For some theoretical distributions, such as Gaussian, maximization of the log-likelihood function can be solved analytically by taking partial derivatives of that function with respect to each unknown parameter and setting those equal to zero. This is not possible in the case of the gamma distribution since the derivative of the standard gamma function must be evaluated numerically. Therefore, maximizing the log-likelihood function is performed iteratively using standard multivariate optimization techniques. These also involve computation of partial derivatives of the log-likelihood function. Again, in the specific case of the gamma distribution, there will be a term involving the sum of the logarithm of the observed values. For rainfall, this is problematic since zero observations are quite common. Therefore an modified method is required.

Wilks (1990) outlines the method of maximum likelihood estimation of the gamma distribution parameters for data containing values of zero[1]. This uses the statistical concept of *censored* data. Type I censored data (Kendall and Stuart 1977) contain a known

1. Professor Daniel S. Wilks of Cornell University kindly supplied a Fortran subroutine that implements this algorithm. This code was used in this work.

number of observations above or below some detection limit, with unknown numerical values. Typical applications of censored data in statistics involve censoring "on the right" for "survival" data, such as the number of people who did not die as part of a medical trial. In the case of rainfall, "left censored" data is involved, where, due to characteristics of the observing system, a given number of observations fall below the detection limit of the sensor. For example, in a tipping bucket raingage, zero rainfall will be reported until the amount of rain reaching the bucket is greater than the sensitivity of the instrument (e.g., 0.01"). There is no way of knowing exactly how much rainfall was observed in this instance, it could be any value between zero and 0.01 inches. Operational gridded radar rainfall estimates are also censored below a small amount of rain (0.25mm). For MLE of the gamma distribution using data that include N_c values of zero, and N_w non-zero values ($T=N_c+N_w$), Wilks (1990) shows how the likelihood function is modified to allow for censored data:

$$L(\alpha, \beta; W) = [F(C;\alpha, \beta)]^{N_c} \prod_{i=1}^{N_w} (w_i/\beta)^{\alpha-1} \frac{\exp(-w_i/\beta)}{\beta\Gamma(\alpha)} \tag{2.20}$$

$F(C;\alpha,\beta)$ is the cumulative distribution function:

$$F(C;\alpha, \beta) = \int_0^C (x/\beta)^{\alpha-1} \frac{\exp(-x/\beta)}{\beta\Gamma(\alpha)} dx = \Pr\{x \le C\} \tag{2.21}$$

which is the probability that an observation is less than or equal to the censoring threshold. The logarithm of the likelihood function is taken as before, except now all terms involving the sum of the logarithm of the observed values use the N_w non-zero values only. The usual optimization procedures for finding the maximum are followed from this point on.

2.9.3Generalized method of moments

Rainfall data, like many meteorological variables, are spatially correlated. As previously mentioned, a key assumption in the method of maximum likelihood estimation is that the data are independent and identically distributed. Independence implies that the occurrence of one event does not depend on the occurence of another, or that the data are serially uncorrelated. Spatial rainfall data will undoubtedly violate this assumption. For this reason, a robust method of parameter estimation is desired that does not rely upon an assumption of independence, for example, the *generalized method of moments* (GMM, Hansen 1982; Hamilton 1994). GMM can allow correlation in the data to affect the parameter estimation. The generalized method of moments 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 N unknown parameters in the theoretical distribution, N such equations are necessary. The resulting parameters will produce a theoretical distribution that fits those N moments exactly. In some cases, however, it may be desirable for the parameters to provide a better fit to the observed skewness (related to the 3rd moment) or kurtosis (related to the 4th moment). For example, if two parameters are unknown, one might desire to produce parameter estimates that fit the first, second, third, and fourth moments of the sample as closely as possible. A non-linear vector function $g(\theta,w)$ could be produced representing the difference between the sample moments and the population moments, using the gamma distribution for example:

$$g(\theta, w) \equiv \begin{bmatrix} \{\hat{\mu}_1 - \alpha\beta\} \\ \{\hat{\mu}_2 - \alpha\beta^2(\alpha+1)\} \\ \{\hat{\mu}_3 - \alpha\beta^3(\alpha^2+3\alpha+2)\} \\ \{\hat{\mu}_4 - \alpha\beta^4(\alpha^3+6\alpha^2+11\alpha+6)\} \end{bmatrix} \tag{2.22}$$

Here, $\theta = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$ is the vector containing the parameters of the gamma distribution, w_t represents the values of the sample, and $\hat{\mu}_n = \langle 1/T \rangle \sum_{t=1}^{T} (w_t)^n$ is the n^{th} sample moment.

One can create an objective scalar function $\Phi(\theta) = g(\theta, w)^T A g(\theta, w)$ which represents the weighted sum of squared errors of the estimates of the parameters, where **A** is a symmetric positive-definite weighting matrix that represents the relative importance of fitting each of the moments. In this work, the θ that minimizes this function was found iteratively using the bounded truncated-Newton method (Nash 1984).

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

$$\hat{S}_T = (1/T) \sum_{t=1}^{T} g(\hat{\theta}, w) g(\hat{\theta}, w)^T \tag{2.23}$$

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):

$$\hat{S}_T = \hat{\Gamma}_{0,T} + \sum_{v=1}^{q} \{1 - [v/(q+1)]\}(\hat{\Gamma}_{v,T} + \hat{\Gamma}^T_{v,T}) \tag{2.24}$$

where:

$$\hat{\Gamma}_{v,T} = (1/T) \sum_{t=v+1}^{T} [g(\hat{\theta}, w_t)][g(\hat{\theta}, w_{t-v})]^T \tag{2.25}$$

and q is the lag-correlation length. Newey and West (1987) show that eq. 2.24 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 (θ) is needed. An iterative procedure is followed where an initial estimate of the parameters (θ_0) are obtained using an arbitrary weighting matrix such as the identity matrix A_0=I. This estimate of θ is used in eq. 2.24 to produce an initial estimate of S_T, which is inverted to produce A_1. The objective function Φ is minimized using A_1 to produce a new estimate θ_1, which is then used to estimate A_2. These iterations continue until convergence is reached. For all cases in this work convergence was reached in five iterations or less. To my knowledge, this work is the first example of the use of GMM with rainfall data in the meteorological community.

2.10 Geostatistics

In the previous section, attributes related to the overall distribution of rainfall across an object were presented. These attributes are expected to provide useful information on the intensity of rainfall within each object. However, these attributes will not provide information on the spatial continuity and variability of the rainfall within an object. The

same observed histogram could be realized from different looking events that could be either randomly unorganized or spatially continuous, since the distribution ignores information on the location of rainfall amounts. In order 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.

To find such attributes, the place to turn to is the field of *geostatistics*. Geostatistics is concerned with the study of phenomena that fluctuate in space, of which rainfall is certainly an example. The primary interest for this work is explaining and characterizing the spatial variance and continuity of a rainfall object; however much of the field of geostatistics is also concerned with estimating unknown values of a spatial field at any random location given a set of observations in space. There are several measures of spatial variability and continuity to choose from (Isaaks and Srivastava 1989; Deutsch and Journel 1988). For this work three were examined: two-dimensional plots of the *semivariogram, correlogram, and covariance*. All three measure some aspect of the spatial field as a function of a two-dimensional separation vector h (Fig 2-4). All possible pairs of values that are separated by h on the original field will be used to compute the various statistics. The semivariogram $\gamma(h)$ is defined as half of the average squared difference between the pairs of all values separated by h (Eq. 2.26). The covariance $C(h)$ is the traditional covariance (Eq. 2.27) between all possible pairs of "tail" and "head" values separated by h. The correlogram $\rho(h)$ is also known as the auto-correlation, which is the covariance normalized by the respective tail and head standard deviations (Eq. 2.28).

$$\gamma(h) = \left(\frac{1}{2N(h)}\right) \sum_{N(h)} (w_t - w_h)^2 \qquad (2.26)$$

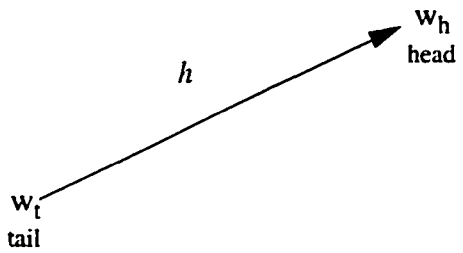


Figure 2-4:A conceptual example of the separation vector h

$$C(h) = \left(\frac{1}{N(h)}\right) \sum_{N(h)} (w_t w_h - m_t m_h) \tag{2.27}$$

$$\rho(h) = \frac{C(h)}{\sigma_t \sigma_h} \tag{2.28}$$

Here m_t and m_h are the means of the tail and head values, respectively, and σ_t and σ_h are the standard deviations of the tail and head values, respectively. N(h) is the total num-

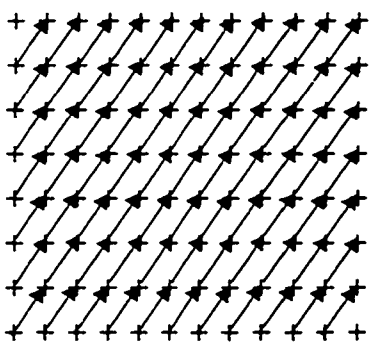


Figure 2-5: Example of data pairs for separation vector h = (1,1). Adapted from Isaaks and Srivastava (1989).

ber of possible pairs of tail and head values for a given separation vector. Figure 2-5 provides an example of how data can be paired to compute a statistic for a specific value of h. Some of the analysis results using these statistics were computed using GSLIB, a freely

available library of software packages for geostatistics developed at Stanford University (Deutsch and Journel 1988).

Examples of 2-D semivariogram, covariance, and correlogram plots are found, along with the corresponding rainfall field, for an example shown by Figure 2-6. The rainfall field (Fig 2-6a) shows fairly continuous heavier precipitation organized along a line ori-

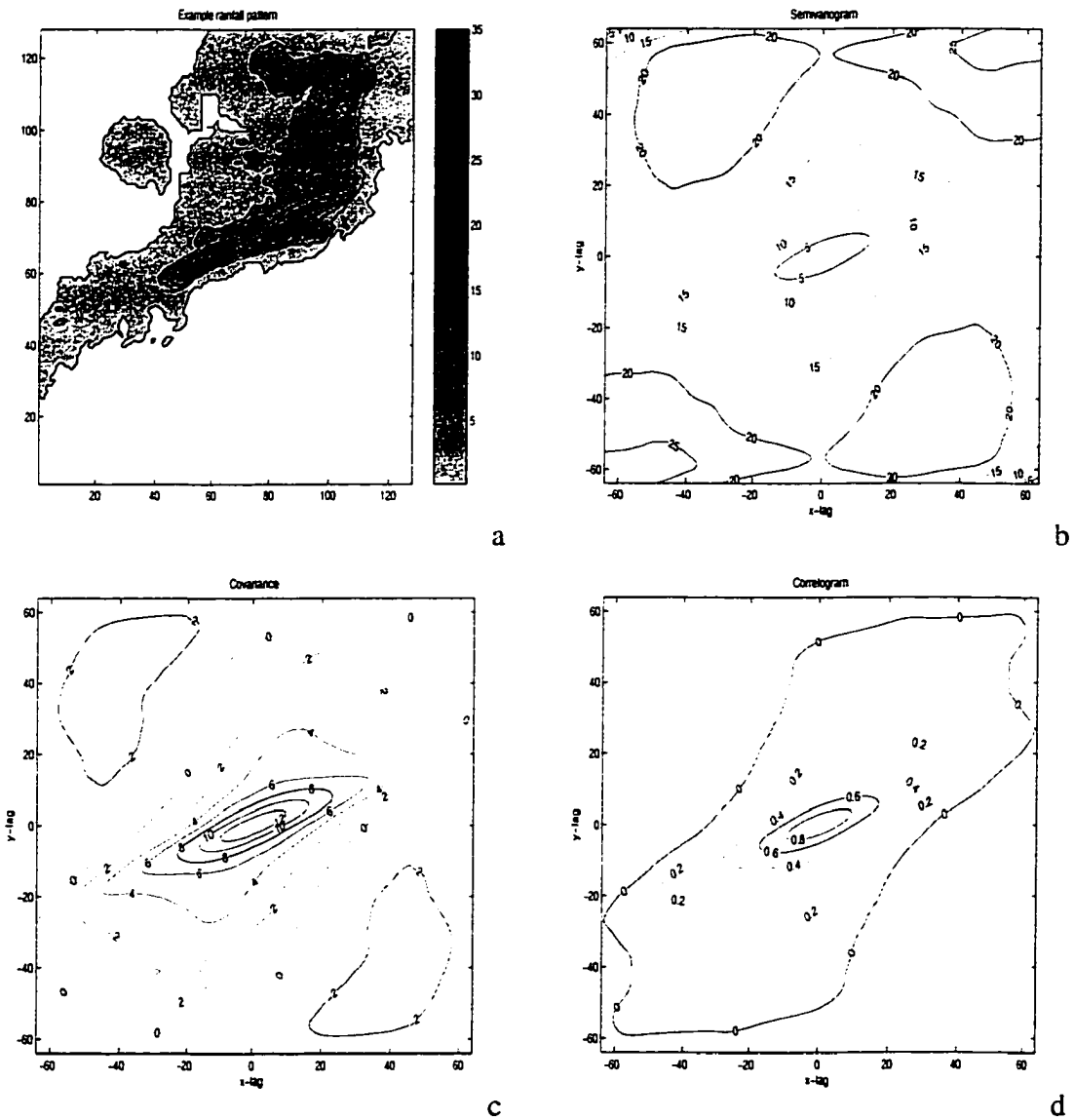


Figure 2-6: Example 1h accumulated rainfall field (a) with corresponding semivariogram (b), covariance (c), and correlogram (d) plots.

ented approximately west-southwest to east-northeast, with strong variations in amounts

normal to this line. The two-dimensional plots of these statistics can be interpreted by using the separation vector concept. For example, the correlation between all pairs of points separated by h=(0,20), that is 20 grid points to the north, is approximately 0.2. The plots are symmetric about any line passing through the origin since these statistics are even functions (for example, $\gamma(h)=\gamma(-h)$). This symmetry is due to the fact that the same pairs of points will be compared when the head and tail of the separation vector are switched. Plots of the semivariogram, covariance, and correlogram (Figs 2-6b-d) provide fairly consistent information, that rainfall values are similar over a large distance in the direction approximately parallel to the x-axis, and similar to other values only over a short distance in other directions. The semivariogram (Fig 2-6b) provides information on the average squared difference, therefore the value at the origin (h=(0,0)) is zero and values increase as h moves further from the origin. The covariance (Fig 2-6c) plot works in the opposite sense, indicating how pairs of values simultaneously vary from their means, the value at the origin is the variance of the overall field. The correlogram (Fig 2-6d) operates in a similar fashion to the covariance plot, except the value at the origin is normalized to 1.0.

There is a long history in the literature of research using geostatistical tools to examine the characteristics of spatial radar/rainfall data. Kessler and Russo (1963) and Kessler (1966) computed the two-dimensional auto-correlation of radar reflectivity. From this, ellipses were fit to the contours of the correlogram, and statistics such as the lengths of the major and minor axes of the average autocorrelation coefficient and the orientation of the major axis were computed. Kessler and Russo (1963) noted how the ellipticity 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. They also found that statistics of this type did not vary greatly during the lifetime of a particular storm. The usefulness of these sorts of attributes in discriminating among different modes of organization in rainfall patterns will be evaluated in Chapter 4. Zawadzki (1973) proposed using a slightly different form of the two-dimensional auto-correlation, by not subtracting the head and tail mean values in the computation of the covariance (Eq. 2.27) and variance. The cross-correlation of reflectivity images in time was used to determine the velocity of a given storm. This idea was the foundation of radar "nowcasting" approaches that are still being researched to this date (i.e., Germann and Zawadzki 2002; Wilson et al 1998). Germann and Joss (2001) examined one-dimensional variograms of radar reflectivity. In this work, all separation vectors of the same length were combined and the resulting variogram provides information on the spatial continuity of the field as a function of distance alone, independent of direction. No information on the anisotropic nature of the field can be obtained. Through the use of variograms, Germann and Joss (2001) show how different precipitation phenomena produce different variograms, as well as how variograms can help to determine the representativeness of a point observations of rainfall, estimate observation error variance, and find preferred regions for convective rainfall. Harris et al (2001) examined several multiscale statistical measures of observed and predicted rainfall fields, including Fourier spectra, generalized structure function, and moment-scale analyses. The generalized structure function using the second moment is equivalent to the semivariogram. These statistical measures were used to determine whether or not a forecast model produced similar spatial variability in the rainfall field as what was observed. Here, a power-law scaling regime is found within various multiscale statistical measures, simi-

larly to previous work examining the fractal properties of rain and cloud fields (e.g., Lovejoy 1982). Zepeda-Arce et al (2000) also performed a moment-scale analysis of observed and predicted rainfall fields, examining the variance of normalized rainfall fluctuations as a function of spatial scale. It seems reasonable to expect that geostatistical parameters might be useful attributes for characterizing the spatial structure and variability of rainfall events.

2.11 Principal Component Analysis

Another method of multivariate statistical analysis that will be used in this work is *principal component analysis* (PCA). PCA is a multi-purpose tool, it allows one to represent a data set in terms of a basis that is uncorrelated (orthogonal). In addition, PCA also helps to explain the variance contained within the data set. PCA is based upon the eigen-analysis of the variance-covariance matrix; eigenvectors of this matrix are orthogonal, and the eigenvector associated with the largest eigenvalue points in the direction that the data set exhibits the most variability (called the first principal component). PCA can be used as a data reduction tool, a data set can be reduced in dimension by retaining only a subset of the eigenvectors. Since the eigenvalues reveal the relative amount of variance explained by their respective eigenvectors, a specified amount of the total variance contained within the data can be preserved. If a large fraction of the total variance is explained by the first few principal components, a great deal of data reduction can be accomplished. Richman (1986) showed how rotation can be applied to PCA to create a basis for the data where the basis vectors point towards clusters of highly related variables. The rotated principal components do not necessarily have to retain the orthogonality of the original eigen-vectors. Gong and Richman (1995) argue that cluster analysis and rotated PCA have the same

goals.

The operation of PCA proceeds as follows. Define $X = [x_{ij}]$, $1 \le i \le m$, $1 \le j \le n$ as a $m \times n$ data matrix with n objects and m attributes. The Grammian Σ of X is a real, symmetric matrix defined as:

$$\Sigma = X^T X \tag{2.29}$$

The Grammian could use the raw values of X (Σ=second moment matrix), centered values (Σ=covariance matrix), or normalized unit-variance values (Σ=correlation matrix). Let (λ_i, p_i) represent an eigenvalue/eigenvector pair for Σ:

$$\Sigma p_i = \lambda_i p_i \tag{2.30}$$

The full eigen-system represented by i=1,2,...n of the above equation can be denoted as:

$$\Sigma P = P\Lambda \tag{2.31}$$

where $P = [p_1, p_2, \ldots, p_n]$ is the matrix of eigenvectors and Λ is the diagonal matrix of eigenvalues:

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix} \tag{2.32}$$

Since Σ is a real and symmetric matrix, P is orthogonal, that is:

$$P^T P = I = P P^T \tag{2.33}$$

Therefore, equation 2.33 can be rewritten as:

$$\left.\begin{aligned} P^T\Sigma P &= \Lambda \\ P^T X^T X P &= \Lambda \\ (XP)^T XP &= \Lambda \end{aligned}\right\} \tag{2.34}$$

Define $\bar{Z} = XP$, then this equation becomes:

$$\bar{Z}^T\bar{Z} = \Lambda \tag{2.35}$$

Assuming Σ is positive-definite, the eigenvalues will be positive, therefore the "square-roots" of Λ can be taken:

$$\left.\begin{aligned} \bar{Z}^T\bar{Z} &= \Lambda^{1/2}\Lambda^{1/2} \\ \Lambda^{-1/2}\bar{Z}^T\bar{Z}\Lambda^{-1/2} &= I \\ (\bar{Z}\Lambda^{-1/2})^T\bar{Z}\Lambda^{-1/2} &= I \\ F^T F &= I \end{aligned}\right\} \tag{2.36}$$

where $F = \bar{Z}\Lambda^{-1/2}$. F is an uncorrelated transformation of the data matrix X called the principal component *scores*. The objects found in X are projected in the directions represented by the eigenvectors found in P, normalized by dividing by the square root of the eigenvalues, λ_i. Further:

$$\left.\begin{aligned} F &= \bar{Z}\Lambda^{-1/2} \\ F &= XP\Lambda^{-1/2} \\ F &= X(P\Lambda^{-1/2}) \end{aligned}\right\} \tag{2.37}$$

Solving for X, we find:

$$\left.\begin{aligned} X &= F(P\Lambda^{-1/2})^{-1} \\ X &= F(\Lambda^{1/2}P^{-1}) \\ X &= F\Lambda^{1/2}P^T \end{aligned}\right\} \tag{2.38}$$

or $X = FA^T$ where $A^T = \Lambda^{1/2}P^T$ or $A = P\Lambda^{1/2}$. A is called the principal component *loading*, which represents the covariances (assuming Σ represents the covariance matrix, correlations if Σ is a correlation matrix) between F and X.

As shown above, PCA can be used to transform the data matrix, projecting X onto a basis or coordinate system that is uncorrelated. In this work, principal component analysis will be used to visualize a multi-dimensional data set. Conceptually, visualizing a high-dimensional data set using only the leading two components of the PCA scores is equivalent to slicing a 2-D plane through the data in such a way that the maximum possible amount of variance contained in the original data set is displayed on that plane.

2.12 Image Processing

To this point, the mathematical tools and concepts that have been outlined have primarily focused on providing attributes that characterize aspects of objects and analyzing the similarity of objects. The method of locating and identifying individual objects within a full realization of rainfall has not yet been discussed. For this task, the discipline of image processing provides a wide range of tools. Image processing tools are also used in this work to assist in further data reduction in the automated analysis of correlogram information. Although the concepts of an image are not foreign to meteorologists, some definitions of common terms will be provided at the onset.

An *image* is a representation of values onto a set of spatial coordinates (x,y). Typically, image values are often associated with grey scales or some other color map to allow for visualization. Formally (Klette and Zamperoni 1996), an image is a function f defined on a set of *image points* $p=(x,y)$. The *image value* $f(p)=f(x,y)$ is the numerical value of the function at point p. A *pixel* is an element of the image, represented by its location and

value *(x,y,f(x,y))*. Optical systems generate *analog images*, where the spatial coordinates *(x,y)* and image values can be considered continuous variables. Typical computer-based images consist of data stored as grids with a finite number of spatial coordinates and image values, known as *digital* or *discrete images*. A Cartesian coordinate system is often used, where x and y take on integer values with intervals $1 \le x \le M, 1 \le y \le N$ for an image sized $M \times N$. The origin is typically taken to be the lower left corner of the grid. Although most meteorological analyses consist of continuous variables represented on discrete grids, in this work meteorological fields (i.e., rainfall) will be treated as discrete images. In fact, due to bit packing used in the standard formats for transferring gridded weather data (Stackpole 1994) continuous variables are truncated to discrete values, whose number is determined by the number of bits used in the packing process.

Many image processing tools involve computations of image values in a neighborhood surrounding a particular point. To assist in the description of these algorithms, a commonly used (Davies 1997) template or moving window will be defined as follows. Assume original image values are stored in image space P, where P(*i,j*) is the local pixel value. In the neighborhood around P(*i,j*), define a template where P0=P(*i,j*), P1=P(*i*+*1*,*j*),

P4	P3	P2	$j+1$
P5	P0	P1	j
P6	P7	P8	$j-1$
$i-1$	i	$i+1$	

Figure 2-7: Neighborhood numbering scheme.

and so on (figure 2-7). Assume that the results of processing the original image P will be stored in space Q. Various basic image processing algorithms can proceed by moving the

template across all pixels found in the original image and operating on the image values found within the window defined in Figure 2-7. For example, a simple process involves copying an image from one space into another. In FORTRAN, this might look like:

```
do j=1,N
do i=1,M
Q(i,j)=P(i,j)
end do
end do
```

For sake of brevity, loops across the entire image will be represented by double brackets [[]]. In addition, the numbering scheme used in Figure 2-7 will also be used. Therefore, the example given above would be replaced with the following *pseudo-code*:

$$\text{COPY:} \qquad [[Q0 = P0]] \tag{2.39}$$

Another simple example would be to shift the image to the left by one pixel:

$$\text{LEFT:} \qquad [[Q0 = P1]] \tag{2.40}$$

Other basic image processing algorithms can be found in Davies (1997) and Klette and Zamperoni (1996). For this work, a more useful example is the creation of a binary (1=dark, 0=light) image through the use of a simple threshold. This algorithm is commonly used for object detection.

$$\text{THRESH:} \qquad [[\text{IF } (P0 > \text{thresh}) \; Q0=1 \text{ ELSE } Q0=0 \;]] \tag{2.41}$$

Once a binary image is obtained, one may wish to shrink or expand the extent of the dark regions of the image. For these types of processes, the sum of the values surrounding the center point (sigma) is introduced.

$$\text{SHRINK:} \qquad \begin{array}{c} [[\text{sigma} = P1 + P2 + P3 + P4 + P5 + P6 + P7 + P8 \\ \text{IF } (\text{sigma} < 8) \; Q0=0 \text{ ELSE } Q0=P0]] \end{array} \tag{2.42}$$

One can easily see how the SHRINK process would shrink the darker portion of the image. The only way the new pixel would be dark (=1) would be if all 9 points found in the moving window on the original image were also dark. The converse would be true for the EXPAND process:

$$\text{EXPAND:} \quad \begin{array}{c}[[\text{sigma} = P1 + P2 + P3 + P4 + P5 + P6 + P7 + P8 \\ \text{IF (sigma} > 0) \text{ Q0=1 ELSE Q0=P0]]}\end{array} \tag{2.43}$$

Edge detection for a binary image also makes use of the sigma sum. The purpose of the edge detection algorithm is to remove all pixels that are not on the edge of the dark region (assumed to be an object). If a pixel is in the middle of a dark region, one can easily see that the sigma sum will be equal to 8. If a pixel is in the midst of a bright region, the sigma sum will be equal to 0. Therefore, the edge points will be in the range of 1 to 7. Edge points are defined to be part of the original dark object.

$$\text{EDGE:} \quad \begin{array}{c}[[\text{sigma} = P1 + P2 + P3 + P4 + P5 + P6 + P7 + P8 \\ \text{IF (sigma} = 8) \text{ Q0=0 ELSE Q0=P0]]}\end{array} \tag{2.44}$$

This edge detection algorithm will be used extensively in this work.

Given an image that contains multiple objects, one might wish to locate individual objects and label each one uniquely. Here, an object in a binary image is defined as a contiguous region of dark pixels. Each contiguous region will be labeled with a unique integer number. For this task, a connected component labeling algorithm is required. The algorithm used here is adapted from Klette and Zamperoni (1996). Since it is considerably more complicated than the simple algorithms introduced in the previous paragraph, the method will be described without providing detailed pseudo-code. The algorithm proceeds in several steps. First, a binary image A is obtained, perhaps by using the THRESH operator (2.41) on the original digital image P. Next, a image B containing values of the

object labels is created with default values set to zero. Scanning through binary image A, once an unlabeled dark pixel of A has been found, a region growing algorithm is used to apply the same label to other neighboring dark pixels. The region growing algorithm is similar to the EXPAND (2.43) algorithm above, except the routine searches in an outward reaching spiral from the "seed" point, and unlabeled dark pixels (where $A(i,j)=1$ and $B(i,j)=0$) will obtain the same label as the seed point if any of its 8 neighbors also have the same label as the seed point. As depicted in figure 2-8, one pass of the region growing

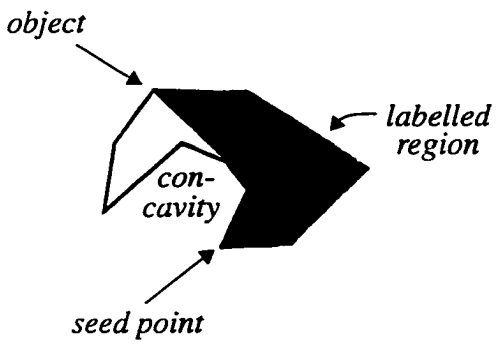


Figure 2-8: Impact of concavity on first pass of label algorithm. The entire object is the union of the dark and light shaded regions. The portion of the object that is given the same label as the seed point is solid black.

algorithm will not completely "grow" the entire contiguous region for objects containing concavities. A set of points is *concave* (or non-convex) if there exists at least two points in the set where a straight line connecting the two would not fall entirely within the set. For a given label value, one could repeat the region growing algorithm several times for each object until no new pixels are assigned that label; however this would be computationally expensive. A faster alternative is to allow different parts of objects to be labelled with different values while compiling a table of neighboring points possessing different labels. Each time a pixel in B is assigned a value, the surrounding 8 neighbor points are checked

for different label values, and if any are found, those values are stored in a table. The final step of the connected component algorithm is to reconcile the multiple labels so that each pixel within a contiguous object has the same label value. This step consists of assigning the lowest label value of all connected labels to each pixel determined to be connected in the "multiple label" table.

Many other image processing operations, including image enhancement and noise suppression, are *window functions* or *local operators*, where the processed image pixel is some function of the values of the original image within a window or neighborhood surrounding the pixel location. For example, to smooth an image, various filters can be used that compute a weighted sum of the pixel values in the vicinity of each location. In order to examine the impact of these types of functions on the image, it is convenient to use the mathematical concept of *convolution*. The continuous form of the convolution of (one-dimensional) function f with function h is:

$$g(t) = \int_{-\infty}^{\infty} f(\tau)h(t-\tau)d\tau \tag{2.45}$$

which is typically denoted by $g = f*h$. In terms of using convolution for smoothing, f is the input signal or image, h is the filtering function (impulse response), and g is the resulting smoothed signal. In discrete form, the integral is replaced by a sum:

$$g(i) = \sum_{k=-\infty}^{\infty} f(k)h(i-k) \tag{2.46}$$

which, again in terms of a filter, the smoothed signal g is obtained by a weighted sum of the input signal f, where the weights are determined by h. Typically, h has what is known as *compact support*, where the values of h are zero for all absolute values of k greater than

some small value *w*. Therefore, the infinite sum can be replaced by:

$$g(i) = \sum_{k=-w}^{w} f(k)h(i-k) \tag{2.47}$$

where *w* is the *window size*. For a two-dimensional function, the convolution can be performed along one dimension at a time, or the convolving functions can be combined to make a two-dimensional function:

$$g(i,j) = \sum_{l=-w_y}^{w_y} \sum_{k=-w_x}^{w_x} f(k,l)h(i-k,j-l) \tag{2.48}$$

Given a 3x3 convolution function h and the numbering notation found in figure 2-7, this can be replaced by the following pseudo-code, where P is the input image and Q is the resulting smoothed image:

$$\text{CONVOLVE:} \quad \left[\left[Q0 = \sum_{i=0}^{8} P_i h_i \right]\right] \tag{2.49}$$

In order to design a filter function, the effect of the filter in the frequency (for time domain variables) or wavenumber (space domain) space must be established. For example, a *low-pass* filter allows the low frequency waves through, but smooths out high-frequency waves. In general, the response for waves that should be kept is relatively high and the response for waves that one wants to remove is near zero. In frequency space, the filter can be considered to be such a function multiplying the signal (figure 2-9). In order to apply this approach, a Fourier transform of the input would be required, then multiplying this result by the filter response function, and finally an inverse Fourier transform of the result would be needed to get back into the time domain. Note that this involves two

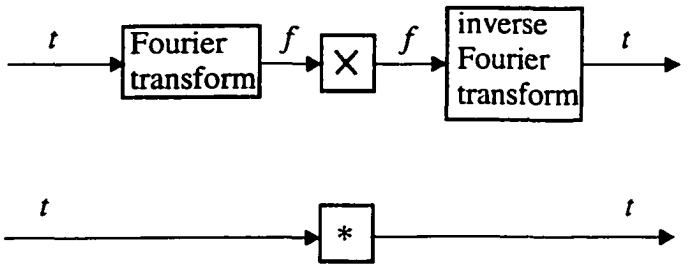


Figure 2-9: Two equivalent methods of filter application. Here, *t* indicates a signal in the time domain, *f* indicates frequency domain, x indicates multiplication by a filter response, * indicates convolution with the Fourier transform of the same filter response. Adapted from Davies (1997).

Fourier transforms, which can be computationally expensive. By taking advantage of the convolution theorem, which says that the Fourier transform of a convolution is the same as the product of the Fourier transforms, one could obtain the same result by convolving the input with the Fourier transform of the response function (figure 2-9). If we define a Fourier transform (Boas 1983) by:

$$g(\alpha) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(x)e^{-i\alpha x}dx \tag{2.50}$$

the product of two Fourier transforms becomes:

$$\begin{aligned} g_1(\alpha)g_2(\alpha) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} f_1(u)e^{-i\alpha u}du \cdot \frac{1}{2\pi}\int_{-\infty}^{\infty} f_2(v)e^{-i\alpha v}dv \\ &= \left(\frac{1}{2\pi}\right)^2 \iint_{-\infty}^{\infty} f_1(u)f_2(v)e^{-i\alpha(v+u)}dudv \end{aligned} \tag{2.51}$$

By applying a change of variables $x = v + u$, $dx = dv$ in the v integral, (2.51) becomes:

$$g_1(\alpha)g_2(\alpha) = \left(\frac{1}{2\pi}\right)^2 \iint_{-\infty}^{\infty} f_1(u)f_2(x-u)e^{-i\alpha x}dudx \tag{2.52}$$

Plugging in (2.45), (2.52) becomes:

$$g_1(\alpha)g_2(\alpha) = \frac{1}{2\pi}\left[\frac{1}{2\pi}\int_{-\infty}^{\infty} f_1 * f_2 e^{-i\alpha x} dx\right] \tag{2.53}$$

which is $\frac{1}{2\pi}$ times the Fourier transform of $f_1 * f_2$. If the convolving function has compact support, this will likely not require as much computation time as the preceding "Fourier transform-multiply-inverse Fourier transform" method.

As shown in equation (2.49), the convolution operator simplifies to an application of a template across the image. Some examples of commonly used templates will now be presented. Marr and Hildreth (1980) show that for smoothing purposes, an ideal filter should be smooth and compact in both the spatial and wavenumber domains. For example, filters that are not smooth in the spatial domain, such as a square-wave (equivalent to a box-average in 2-D), will have a Fourier transform that is not compact, containing sidelobes in the wavenumber domain. These two requirements are conflicting and related by the Heisenberg uncertainty principle, which states that both the spatial and wavenumber domains cannot be measured with arbitrarily high precision. Marr and Hildreth (1980) find the filtering function that optimizes the uncertainty principle is Gaussian, which can be written in two-dimensions:

$$G(x, y) = \frac{1}{2\pi\sigma^2}\exp\left[\frac{-(x^2 + y^2)}{2\sigma^2}\right] \tag{2.54}$$

Here, σ^2 is the variance and determines the "width" of the window when turned into a discrete operator. A commonly used template which closely approximates a Gaussian function is:

$$\frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} \tag{2.55}$$

which is the 2-D version of the well-known "1-2-1" filter.

Besides smoothing, convolution templates are used for a variety of purposes in image processing. Image sharpening and enhancement along with the detection of lines, edges, and other specific shapes are processes often performed using local window templates. In this work, edge detection operators were tested for their effectiveness at locating regions of convective rainfall embedded within larger areas of stratiform precipitation. The *contra-harmonic filter* (Klette and Zamperoni 1996) is one example that highlights edge pixels based upon the difference between estimates of the local maximum and minimum values within a window. These estimates are non-linear calculations of averages (contra-harmonic average) of the local pixel values. The basic idea behind this filter is that, in the vicinity of edges, large differences between local maximum and minimum values should exist. The filter is simply:

$$h(i,j) = C_M - C_m \tag{2.56}$$

where, assuming a 3x3 window and numbering scheme from figure 2-7:

$$C_M = \frac{\sum_{n=0}^{8} P_n^{r+1}}{\sum_{n=0}^{8} P_n^{r}} \text{ and } C_m = \frac{\sum_{n=0}^{8} P_n^{-r+1}}{\sum_{n=0}^{8} P_n^{-r}}, r>0 \tag{2.57}$$

The C_M and C_m values can be thought of as weighted averages of the local pixel values, where the weights are P_n^r for C_M and P_n^{-r} for C_m. From this perspective, for positive real

values of r, one can see that in the C_M case higher pixel values are enhanced, while lower values are enhanced in the C_m case. The impact of this filter on a hypothetical one-dimensional edge function can be found in figure 2-10. Assuming an object is defined as the set of pixels with relatively high values, and an edge is defined to be part of an object, in this case (figure 2-10a), the edge is located near i=12. The result of (2.56) and (2.57) using r=1.2 and a window consisting of 3 points is shown in figure 2-10b. The maximum of this filter correctly indicates the location of the edge, near i=12.

Another example of a commonly used edge detection operator (Marr and Hildreth 1980) is based upon well-known property of the second-derivative of a function in the vicinity of a edge, that is, edges are co-located with the zero-crossings of the second-derivative of a function. While based upon calculus, the development of the operator was motivated by research into mammalian vision systems (Marr and Hildreth 1980), as the authors desired to make the algorithm as consistent with human vision processes as possible. The calculus-related properties that the operator is based upon can be illustrated using the preceding example. Again, a hypothetical one-dimensional edge function can be found in figure 2-10a, where the edge is located near i=12. Here, the function increases in a step-wise fashion. The first-derivative of this function (figure 2-10c) begins increasing on the "low" side of the step, reaches a maxima near "mid-step" where the slope of the step function is highest, then decreases back to zero as the top of the step is approached. The second-derivative of the function (figure 2-10d) has large positive values on the "low" side of the step, where the first-derivative is increasing, a zero-crossing near i=12, and large negative values on the "high" side of the step. Again, this example illustrates where one intuitively expects to find an edge, somewhere between where the function increases

and decreases the fastest, which corresponds to the location of the zero-crossing of the second-derivative. The Marr-Hildreth operator combines the smoothing properties of the Gaussian with the edge-finding properties of the Laplacian (second-derivative), in fact, it is also known as the Laplacian of Gaussian (LOG) filter. The LOG filter function can be obtained by taking the Laplacian of (2.54), which results in:

$$G''(x,y) = \frac{-1}{2\pi\sigma^4}\left(2 - \frac{(x^2+y^2)}{\sigma^2}\right)\exp\left[\frac{-(x^2+y^2)}{2\sigma^2}\right] \tag{2.58}$$

Again, σ^2 is the variance and determines the "width" of the window ($= 2\sigma$)when turned into a discrete operator, as well as the spatial scales that are included/discarded by the smoothing properties of the Gaussian part of the function. Klette and Zamperoni (1996) provide a detailed implementation of a discrete template version of the LOG function. One of the advantages of the LOG operator is the scale-selectivity of the function, allowing the procedure to be applied at a variety of scales. Large-scale edges will be consistently found via applications of this operator for a wide range of values of σ^2, while smaller-scale edges will disappear for larger values of σ^2. Multiple edge maps can be created at different scales, which allows for the creation of a hierarchical representation of the edge features found within an image (defined as *blobs*, *bars*, and *edges* by Marr and Hildreth 1980).

2.13 Summary

The KDD process (Table 1-9) defines the general framework that has been followed in this work in order to develop a rainfall pattern classification technique. In this chapter, several mathematical tools and concepts needed to perform this task were introduced and

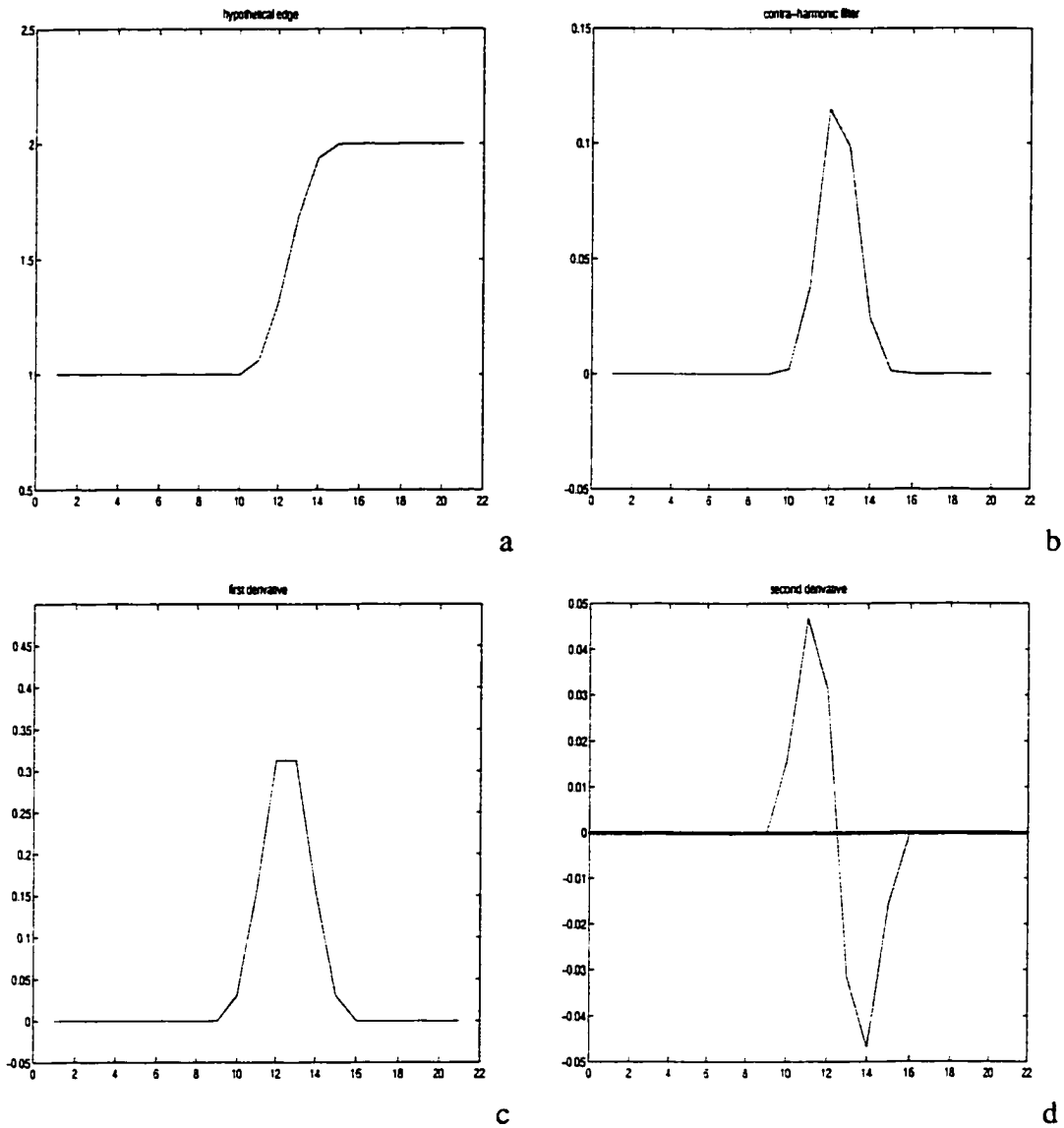


Figure 2-10: Hypothetical example of an edge. Smoothed step function (a) with corresponding contra-harmonic filter of that function (3 point window, r=1.2) (b), finite-difference approximation of first-derivative (c), and finite-difference approximation of second-derivative (d).

discussed in some detail. These involve the creation of a target data set, choice of data mining task and algorithms, and data reduction. In addition, several image processing tools required for object location and identification were introduced. The final steps in the KDD process involve interpreting the results of patterns discovered by the data mining algorithms, possibly modifying the system, iterating previous steps, and consolidating the

knowledge obtained through the process. These steps will be taken in the following chapters where analysis of a target data set will be performed.

Chapter 3

Histogram analysis

3.1 Introduction

The outline of the KDD process (Table 1-1) defines the general framework that has been followed in developing an automated rainfall pattern classification technique. The first step involves developing an understanding of the application and the goals of the end-user. In this work, the goal is to classify rainfall systems. For several reasons, rainfall has been selected as the variable for analysis. The spatial patterns found in precipitation fields often represent important and significantly different meteorological phenomena. There are several potential applications for an automated rainfall classification system, including climatological studies, verification, data assimilation, feature tracking, hydrology, etc. The next steps in the KDD process involve the selection of a target data set and preprocessing the data. A relatively large dataset should be used, one that is richly populated with a variety of interesting and important phenomena that span a large portion of the entire range of possible events. Here, several cases from a high-resolution precipitation analysis produced operationally by the National Centers for Environmental Prediction (NCEP) are used to create the target data set. The use of operationally available data will make this work more relevant and allow for faster implementation into an operational forecasting environment.

Once the target data set has been selected and processed, the next steps in the KDD process involve data reduction and data mining. Data reduction addresses the methods used to extract features of a relatively small dimension within the large-dimensional dataset that allow for proper classification of objects. The determination of useful

attributes that possess good discrimination and classification properties represents the most substantial portion of this work. It is not clear from the outset which attributes will be suitable for use in a classification system. Therefore, appropriate attributes will be discovered by trial and error. The results of an automated classification using trial attributes will be compared to a subjective classification. If these results are in agreement, the trial attributes will be considered useful. To provide a baseline for comparison, no data reduction will be performed and the raw values of analyzed rainfall at every point in space will be tested for their classification ability. As the next step in this multi-faceted analysis process, bulk statistical measures representing the distribution of rainfall values will be tested as trial attributes. As the choice for a theoretical distribution, the gamma distribution is selected since it is well suited for rainfall data and has been widely used for rainfall histogram analysis in the meteorological literature. Due to the spatially correlated nature of rainfall, a robust method of parameter estimation of the gamma distribution is required, therefore the generalized method of moments (GMM) estimation technique was used. Hierarchical cluster analysis is then performed using the parameters of the gamma distribution as attributes to classify the objects in the target data set, and those results are compared to a subjective classification of the rainfall patterns. The results show that this system successfully classifies the cases in the target data set into convective and non-convective events with over 95% accuracy. Much of the work described in this chapter was also included in Baldwin and Lakshmivarahan (2002). Further refinement of the classification will be discussed in chapter 4.

3.2 Target data set

To begin this work, a small target data set 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 for the period covering late summer/early fall of 2000. The Stage IV analysis is a national mosaic of optimal estimates of hourly accumulated rainfall using radar and raingage data, which is available on a 4km x 4km mesh covering the contiguous 48 states. The data analysis routines include a mean radar bias correction, separate radar-only and gage-only analysis mosaics, and a "multi-sensor" analysis combining the radar and gage estimates using an optimal estimation technique (Seo, 1998).

Every observation platform and analysis system contains imperfections and Stage IV analyses are certainly no exception. The primary sources of information are the radar estimates of precipitation, which are potentially fraught with errors (Wilson and Brandes, 1979; Austin, 1987). Typical sources of error include beam blockage by terrain, beam overshoot at long range or where the radar is sited at a high elevation, anomalous propagation, overestimation due to melting snow or hail, radar calibration problems, underestimation of snow, and sensitivity to the Z-R relationship. The Stage IV analysis performs very little quality control (Baldwin and Mitchell, 1998) on the raingage data and no quality control on the radar estimates. The raingage density is fairly sparse (~2500 gages total), particularly in the mountainous Western U.S. where the radar also has the most problems with beam blockage and overshoot. A radar bias adjustment is applied across the entire radar umbrella, no attempt is made to account for bias as a function of range (Smith and Krajewski, 1991). The encoding of the radar estimates causes a truncation of light precipitation amounts, the smallest reported precipitation rate is 0.25 mm/hr, which converts to approximately 0.25 inches/day (the analysis will be rounded to the nearest 0.1mm when

packed into the standard GRIB format for transmission). However, since many of the errors affect the precipitation estimates over a large area in a similar fashion, the spatial structure of the field is assumed to be well-observed and that using the analysis to determine a spatial pattern will be generally valid.

In this case, the method of selecting the rainfall objects (the terms *objects* and *systems* will be used interchangeably when referring to rainfall entities) for target data set was not automated and clearly cannot be used to create an automated system, which is the ultimate goal of this research. However, the purpose of the initial target data set is to test the usefulness of various trial attributes in an automated classification system. This will be accomplished through the use of a data set that is relatively small and manageable but still well-populated with interesting features that are desirable for classification. Forty-eight separate precipitation events occurring at different times and locations across the United States were selected for inclusion in the target data set. The selection criteria was based upon the occurrence of typical rainfall patterns that are often found across the U.S. during the year. The late summer-fall time period was selected due to data availability and the fact that this represents a transition period from warm-season convection to cool-season stratiform precipitation regimes. The size of the domain was chosen to be fixed at 128 x 128 4km grid boxes, which is approximately 500km by 500km. For each case, the domain was centered visually near the event of interest. For each of these 48 events, the rainfall pattern from the entire 500km by 500km domain will be considered the object for classification.

The domain size was chosen for a variety of reasons. Subjectively, 500km was deemed large enough to capture a wide variety of events across a range of spatial scales,

from a significant portion of a synoptic-scale stratiform event to a collection of small-scale cells. Making the domain too large will often mean that many different types of events will be found together in the same domain, making classification more difficult. Making the domain too small will also create problems, since only a fraction of an event will be observed and many different types of events will appear similar when examined with a "zoom lens."

Figures 3-1 through 3-4 display the rainfall patterns for all of the 48 cases found in the target data set. These plots demonstrate that a wide range of times, geographic locations, and rainfall phenomena were included in the target data set. For some cases, (e.g. figure 3-1 case #03) the range of the radar/raingage analysis is indicated by a solid red line. Rainfall outside of this range was assumed to be zero for purposes of this study.

3.3 Subjective classification

Each case was subjectively classified (by the author) into three main event classes; *linear, cellular,* and *stratiform.* In the *linear* class, there is larger-scale organization of smaller-scale elements of heavy precipitation. The smaller-scale elements (usually called "cells") appear to be arranged approximately along a line. For the *cellular* class, there is very little large-scale organization of smaller-scale heavy precipitation elements (similar to "unclassifiable" class of Houze et al. (1990) and "chaotic" class of Blanchard (1990)). The rainfall field typically consists of "cell" features somewhat randomly positioned in a disorderly fashion. In the *stratiform* class, there is large-scale organization of light rainfall, in which the precipitation field shows little variation in any direction over a large area. Linear and cellular events are considered to be related, due to the existence of smaller-scale heavy rainfall elements. These events fall under the *convective* precipitation

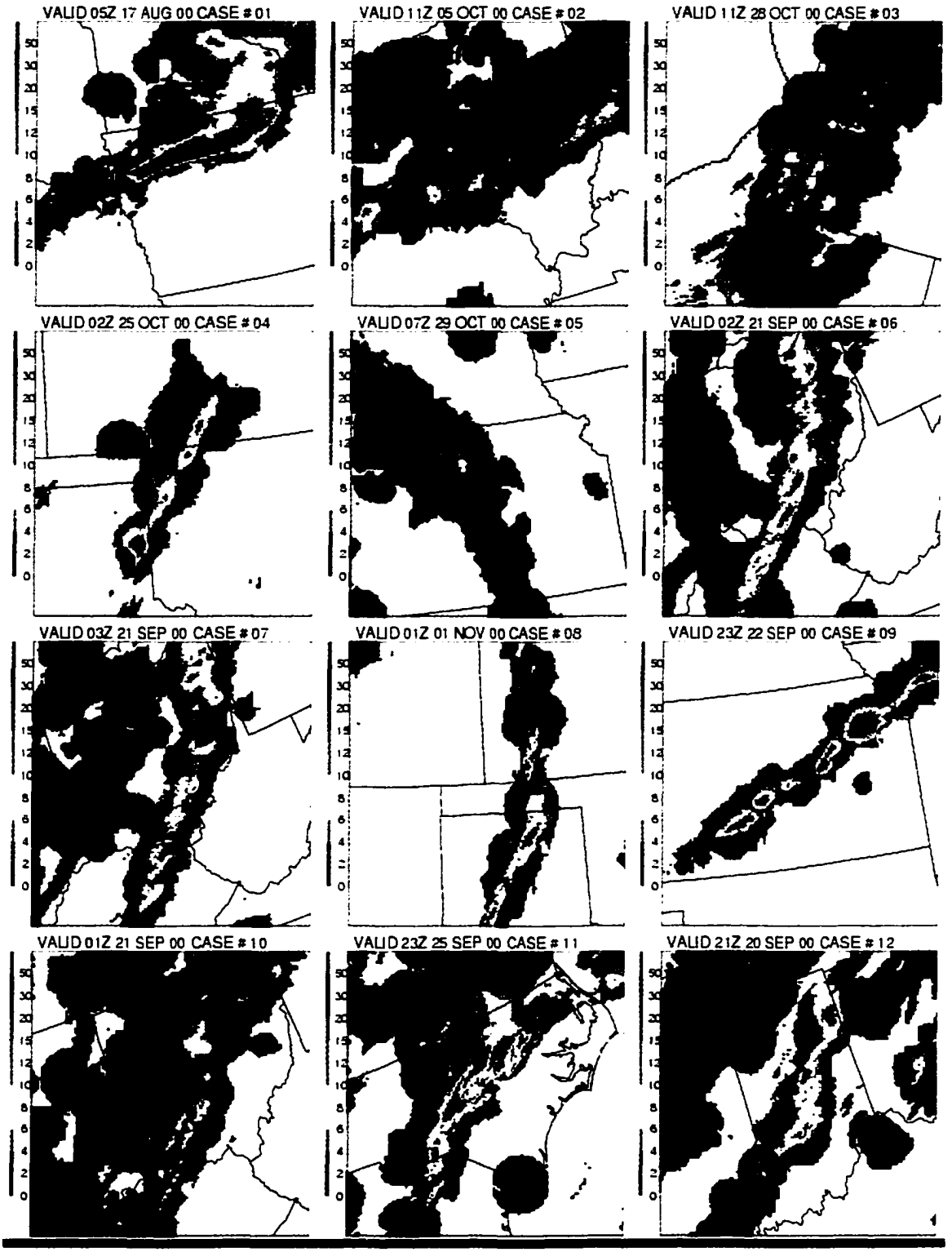


Figure 3-1: Cases (objects) 1 through 12 of the target data set, from NCEP Stage IV 1h accumulated rainfall analyses. Domain consists of 128 x 128 4km grid boxes. Colorbar on the side of each image indicates rainfall amounts in mm. Valid times and case numbers are indicated at the top of each image.

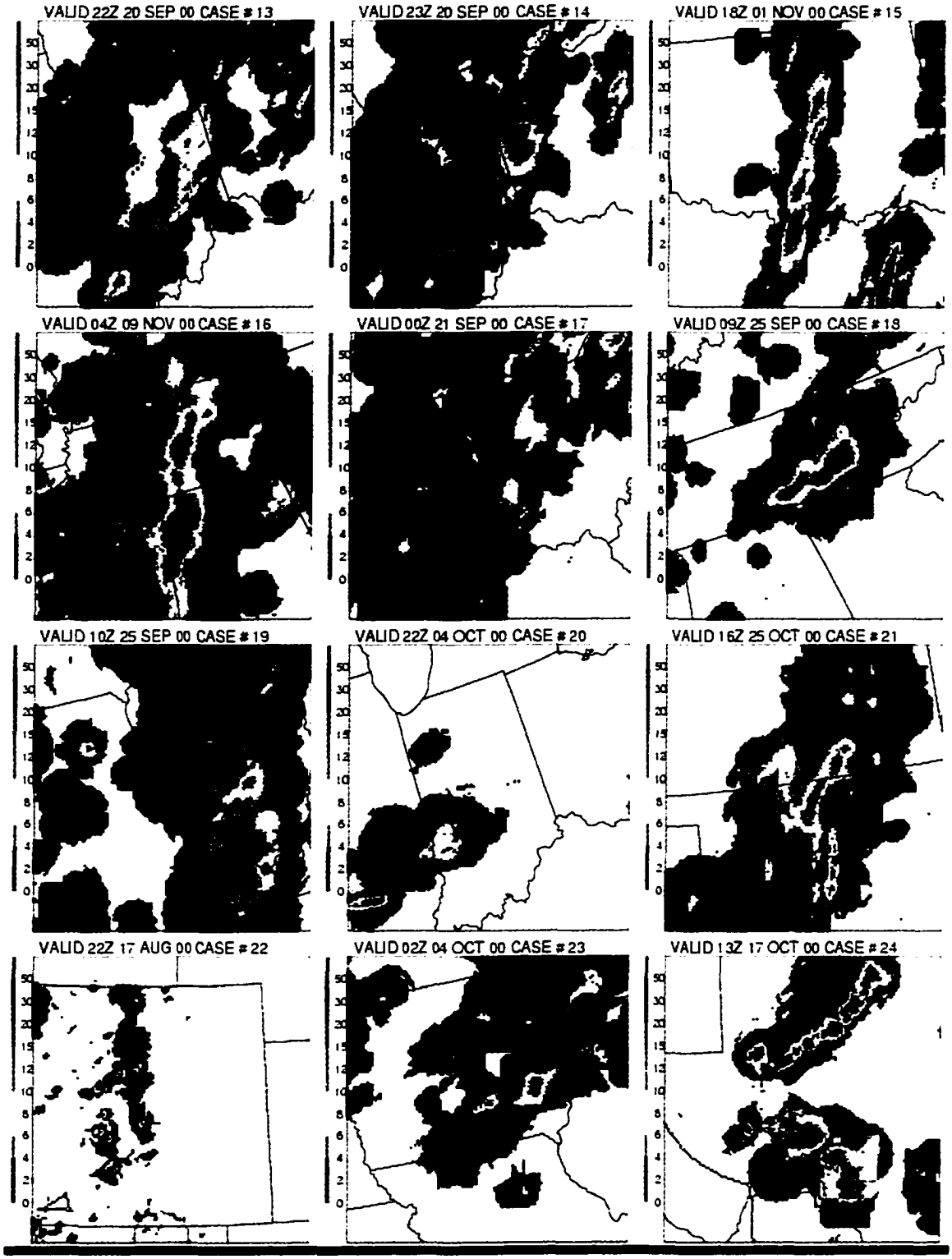


Figure 3-2: Cases (objects) 13 through 24 of the target data set, from NCEP Stage IV 1h accumulated rainfall analyses. Domain consists of 128 x 128 4km grid boxes. Colorbar on the side of each image indicates rainfall amounts in mm. Valid times and case numbers are indicated at the top of each image.

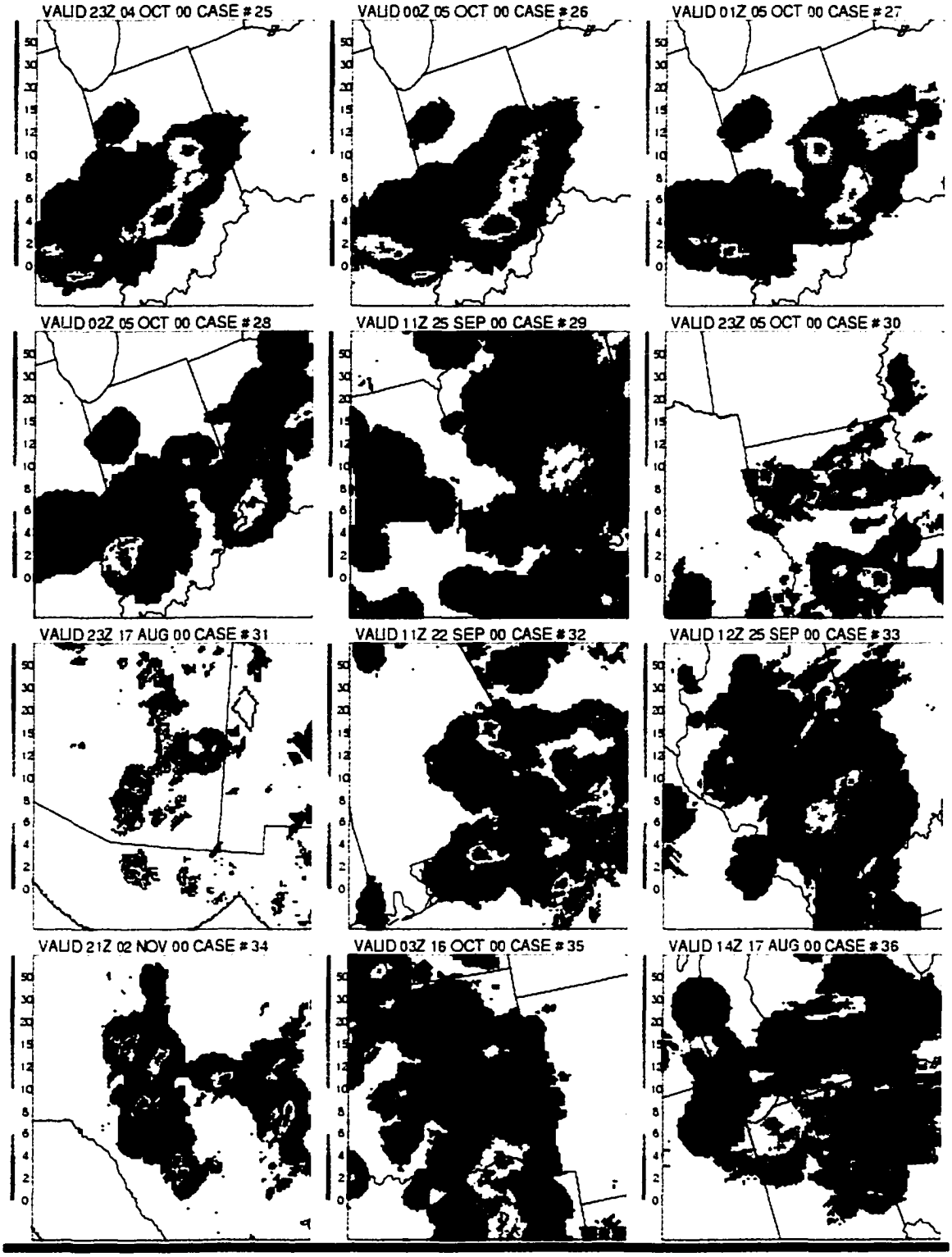


Figure 3-3: Cases (objects) 25 through 36 of the target data set, from NCEP Stage IV 1h accumulated rainfall analyses. Domain consists of 128 x 128 4km grid boxes. Colorbar on the side of each image indicates rainfall amounts in mm. Valid times and case numbers are indicated at the top of each image.

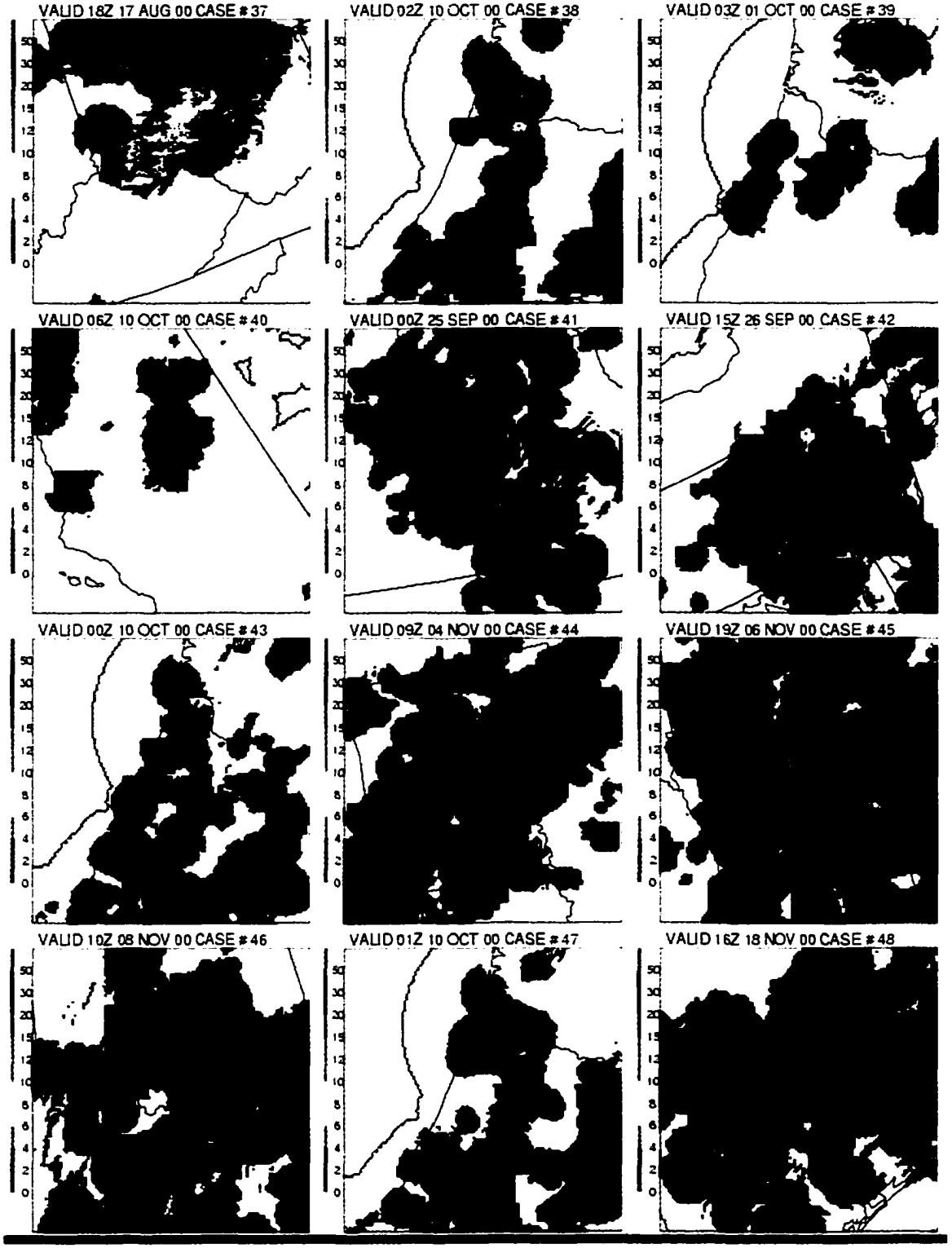


Figure 3-4: Cases (objects) 37 through 48 of the target data set, from NCEP Stage IV 1h accumulated rainfall analyses. Domain consists of 128 x 128 4km grid boxes. Colorbar on the side of each image indicates rainfall amounts in mm. Valid times and case numbers are indicated at the top of each image.

class, which consists of rainfall produced by small-scale (wavelengths on the order 100km and smaller), convectively-driven atmospheric circulations. Stratiform events fall under the general *non-convective* precipitation heading, where rainfall is produced by upward vertical motion resulting from large-scale (wavelengths on the order 1000km and larger) forcing mechanisms. The subjective classification of the 48 cases was based entirely upon the rainfall pattern; no other information (such as meteorological conditions, location, time of year, time of day) associated with the events was provided to the human analyst. This forces the subjective classification to sample from the same "attribute space" as the automated classification systems.

Objective criteria are used to subjectively classify the cases in this work, similar to Bluestein and Jain (1985). For convective events, a significant fraction of the rainfall system must observe 5mm/hr or higher rain rates, otherwise the system will be classified as stratiform. For convective events, the region of heavier rainfall will be surrounded by a rectangular bounding box, and if the ratio of length to width of such a box is 3 or greater (relaxed from the 5 to 1 ratio used by Bluestein and Jain (1985)) the system will be classified as linear, otherwise it will be considered cellular. The determination of what fraction of heavy rain would be considered "significant" when determining whether a system was convective was left to the discretion of the analyst. In addition, the determination of the region of heavier rainfall to be outlined by the bounding box was also left to the discretion of the analyst. Different analysts will have different criteria for determining these aspects of the subjective classification. Table 3-1 shows the distribution of the subjectively classified events across the three main classes. Note that the events are not uniformly distributed, the majority of the events were convective in nature. As described previously, the

linear and cellular classes are considered sub-classes of the convective precipitation class.

3.4 Classification on the raw values

To begin this work, a "baseline" automated classification will be performed. An objective classification of the target data set is performed without any data reduction in order to provide a baseline experiment to compare other classification experiments against. The degree of similarity between the raw values of rainfall at each point in space for the 48 individual events will be analyzed, using cluster analysis. While cluster analysis methods were described in detail in chapter 2, a brief explanation will also be provided here. Hierarchical cluster analysis has been selected as the primary classification tool. Here, similar objects will be grouped together into clusters where objects are defined as rainfall events over regions of fixed size. In general, an object consists of a vector of length m consisting of attributes that describe the object. In this baseline experiment, attributes are rainfall values at each point in space within the object domain, therefore m=16384 (=128×128). The rainfall objects are simply the 2-D rainfall analyses stored as a vector in row-major order. The data matrix for this experiment is therefore of dimension $m \times n = 16384 \times 48$. In this particular case, the similarity of objects will be measured by a single number representing the difference of rainfall values at each point in space, in a point-by-point sense. For example, the *Euclidean distance* between objects j and k would be:

$$d_{jk} = \sqrt{\sum_{i=1}^{m} (x_{ij} - x_{ik})^2} \tag{3.1}$$

(x_{ij} being the rainfall amount at location i for object j) which is almost identical to the root

Table 3-1: Central location, time, date, and subjective event classification for the 48 cases of the target data set.

case number	central location (latitude degrees north, longitude degrees west)	date, time (UTC)	subjective classification
1	42.8, 94.9	17 Aug 2000, 0500	linear
2	39.0, 90.0	05 Oct 2000, 1100	linear
3	42.9, 123.7	28 Oct 2000, 1100	linear
4	36.6, 99.4	25 Oct 2000, 0200	linear
5	39.2, 97.3	29 Oct 2000, 0700	linear
6	39.5, 82.5	21 Sep 2000, 0200	linear
7	39.5, 82.5	21 Sep 2000, 0300	linear
8	37.0, 102.0	01 Nov 2000, 0100	linear
9	38.4, 97.2	22 Sep 2000, 2300	linear
10	40.0, 83.7	21 Sep 2000, 0100	linear
11	35.5, 78.5	25 Sep 2000, 2300	linear
12	39.9, 86.3	20 Sep 2000, 2100	linear
13	40.0, 85.7	20 Sep 2000, 2200	linear
14	40.1, 85.1	20 Sep 2000, 2300	linear
15	34.8, 97.2	01 Nov 2000, 1800	linear
16	35.3, 87.6	09 Nov 2000, 0400	linear
17	40.2, 84.5	21 Sep 2000, 0000	linear
18	35.5, 85.2	25 Sep 2000, 0900	linear
19	38.8, 90.8	25 Sep 2000, 1000	cellular
20	40.0, 86.0	04 Oct 2000, 2200	cellular
21	36.9, 97.7	25 Oct 2000, 1600	cellular
22	39.1, 104.2	17 Aug 2000, 2200	cellular
23	41.4, 92.8	04 Oct 2000, 0200	cellular
24	31.2, 101.6	17 Oct 2000, 1300	cellular
25	40.0, 86.0	04 Oct 2000, 2300	cellular
26	40.0, 86.0	05 Oct 2000, 0000	cellular
27	40.0, 86.0	05 Oct 2000, 0100	cellular
28	40.0, 86.0	05 Oct 2000, 0200	cellular
29	38.8, 90.8	25 Sep 2000, 1100	cellular
30	32.4, 93.0	05 Oct 2000, 2300	cellular
31	32.3, 110.0	17 Aug 2000, 2300	cellular
32	31.8, 85.8	22 Sep 2000, 1100	cellular
33	39.3, 88.9	25 Sep 2000, 12000	cellular
34	30.0, 99.6	02 Nov 2000, 2100	cellular
35	35.0, 95.5	16 Oct 2000, 0300	cellular
36	41.89, 86.1	17 Aug 2000, 1400	cellular
37	38.5, 83.2	17 Aug 2000, 1800	cellular
38	44.6, 123.3	10 Oct 2000, 0200	stratiform
39	45.1, 123.1	01 Oct 2000, 0300	stratiform
40	36.1, 118.7	10 Oct 2000, 0600	stratiform
41	38.7, 94.5	25 Sep 2000, 0000	stratiform
42	41.2, 76.3	26 Sep 2000, 1500	stratiform
43	44.6, 123.3	10 Oct 2000, 0000	stratiform
44	34.6, 92.1	04 Nov 2000, 0900	stratiform
45	42.3, 93.6	06 Nov 2000, 1900	stratiform
46	34.2, 97.4	08 Nov 2000, 1000	stratiform
47	44.6, 123.3	10 Oct 2000, 0100	stratiform
48	30.3, 97.7	18 Nov 2000, 1600	stratiform

mean squared error (RMSE) between two analyses. One might expect that this measure would be sensitive to large differences at a small number of points. This kind of point-by-point comparison might also be sensitive to small errors in phase lag or displacement. For example, two objects containing the same intense rainfall cell, with one object displaced slightly when compared to the other, will produce a large value of d_{jk}, even though the same type of rainfall event is occurring in both cases.

The hierarchical cluster analysis method that will be used is Ward's method (Ward 1963), which is built upon the fact that the total sum of squares (or variance) of all objects in the data matrix is constant and can be partitioned into the between-cluster and within-cluster scatter. The criteria for combining objects into a cluster is minimizing the total sum of squared error, which is the same as minimizing the within-cluster variance, and therefore maximizing the between-cluster variance. This forces objects found within a cluster to be similar while keeping the clusters as different as possible. Ward's clustering algorithm proceeds as follows:

Step 1: Assign each object to separate clusters, each of which contains only one object. The total within-cluster sum of squared error is zero at this point.

Step 2: Compute the increase in the within-cluster sum of squares for every possible merger of two clusters.

Step 3: Create a new cluster by combining the two clusters that produce the smallest increase in the within-cluster sum of squares.

Step 4: Repeat steps 2 and 3 until a single cluster containing all objects is created.

Ward's method has been found to produce satisfying results for meteorological data in pre-

vious research (Alhamed et al. 2002). The statistical toolbox component of the student version of MATLAB (R11) is used to execute Ward's method throughout this work.

The results of a hierarchical clustering algorithm can be displayed as a tree or dendrogram. Figure 3-5 shows a hypothetical example of such results. In this case (unlike

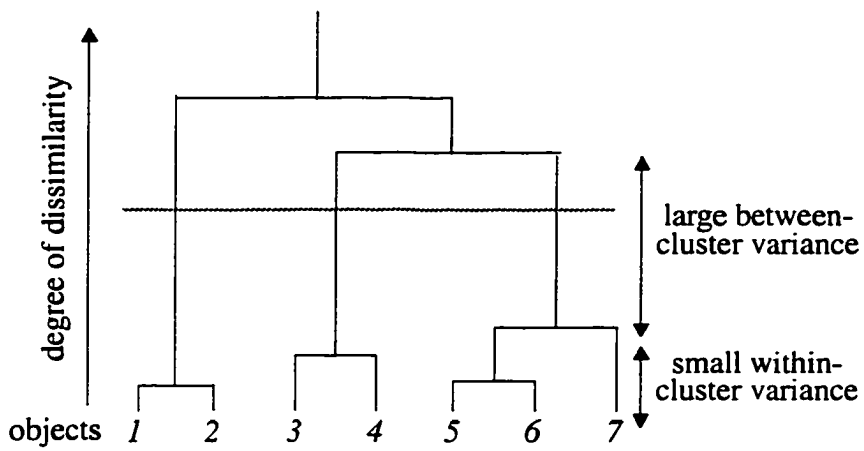


Figure 3-5: Hypothetical hierarchical clustering dendrogram, indicating ideal clustering. Ideally, the within-cluster variance will be relatively small, while the between-cluster variance will be relatively large. An ideal *cut-level*, indicated by the dashed line, can be made in the gap separating the major within-cluster and between-cluster variances.

figure 2-1), the y-axis indicates increasing degrees of dissimilarity, where more similar objects are grouped on the tree at lower parts of the graph. The first question that arises is, how does one interpret the dendrogram? Hierarchical cluster analysis provides information on the relationship of the similarities between objects/clusters, but it does not automatically provide information on the number of clusters found in the data. The number of clusters must be determined *subjectively.* In fact, Kalkstein et al (1987) mention that the term objective maybe should be replaced with automated when referring to this type of analysis technique. The automated classification relies upon several subjective decisions; choice of cluster analysis algorithm, number of clusters, cut-level, etc. Ideally, objects

will be grouped into clusters at a high level of similarity, and a relative small number of major clusters will ultimately be grouped at a high level of dissimilarity. On a dendrogram, this ideal clustering tree might look something like the hypothetical tree found in figure 3-5. An experienced analyst can examine the dendrogram and determine a "cut-level", or a degree of similarity/dissimilarity where the tree can be cut, forming a discrete number of major clusters each containing a number of objects. Returning to the hypothetical example (figure 3-5), the ideal cut-level is somewhere along the "long branches" of the tree, where there is a gap between the level where the major clusters are found to be relatively dissimilar and the level where the objects within each cluster are found to be relatively similar. This would produce clusters that contain objects which appear similar to other objects within a cluster, but different from objects found in other clusters. Of course, this is an idealized example and, as we will soon see, results from real data do not always follow idealized examples.

Using hierarchical cluster analysis as a classification tool requires a subjective decision in order to determine the number of clusters. In this work, this subjective decision will be made as objectively as possible, with the goal of producing groups of objects that are as close to ideal as possible. The cut-level will be made such that a small number of clusters (3 or 4) contain as many objects in the target data set as possible. Every attempt will be made to cut the dendrogram at a point where there is substantial separation between the intra-cluster and inter-cluster variance. This cut-level will likely result in a number of outlier objects that do not belong to any of the major clusters. The outliers will not be classified. Each major cluster will be considered a different *class* of objects. For each of these main clusters, the class definition will be determined by the highest

percentage of subjectively classified cases detected for that particular cluster. For example, a cluster that contains a large fraction of the events that were subjectively classified as stratiform will be considered a stratiform class for purposes of the automated classification. The percent of objects that are "correctly" classified by their membership in the dominant subjective class for each cluster will be used as the metric for determining the skill of the automated classification.

Figure 3-6 shows the results of Ward's cluster analysis on the baseline (no data reduction) target data set. At first glace, there does not appear to be an ideal cut-level that results in three to four main clusters with small within-cluster variance and large between-cluster variance. For many of the objects, the increase in the variance for combining two objects is nearly as great as the increase in the variance found for merging a large number of objects. In addition, there is no cut level that results in a small number of outliers and three or four main clusters. It is assumed that at least three main clusters should be used since three subjective classes are provided. Given these difficulties, a cut level was determinined on the dendrogram tree at approximately the 300 level (indicated that the square root of the increase in the total sum of squared errors is ~300) that results in separating the 48 cases into three main clusters and eleven outliers. This number of outliers represents a significant fraction (22.9%) of the total number of objects in the data set, which is undesirable. For each of the main clusters, the dominant class was defined as the class (linear, cellular, or stratiform) that contained the highest percent of cases detected for that particular cluster. For instance, the dominant class for cluster #1 was stratiform, since the cluster detected 9 of the 11 (82%) cases belonging to that class. Only 1 of the 18 linear cases (5.6%) and 3 of 19 (15.8%) of the cellular cases were part of this first cluster.

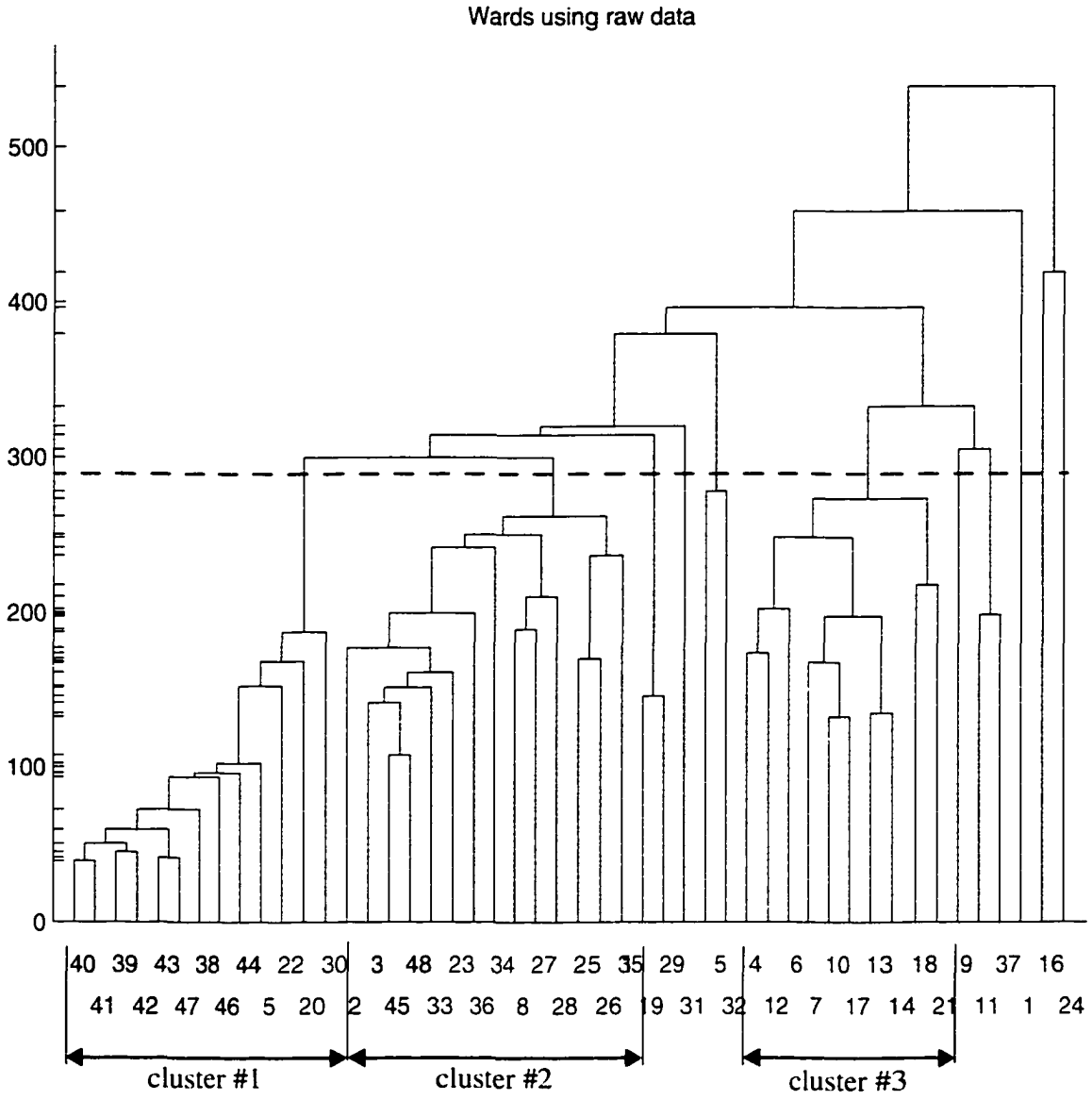


Figure 3-6:Dendrogram produced by Ward's method with target data set using raw rainfall data as attributes. Each object is color-coded by its subjective classification, linear events are green, cellular events are red, stratiform events are blue. Tick marks on the y-axis indicate the square root of the increase in the total sum of squared errors at which two clusters are combined to form a new cluster.

For cluster #2, 47.4% of the cellular cases (9 of 19) were detected, while 16.7% of the linear cases (3 of 18) and 18.2% of the stratiform cases (2 of 11) were also detected, therefore this cluster was considered to be cellular. Cluster #3 was considered linear

dominant, with 50% (9 of 18) of all linear cases detected, 5.3% (1 of 19) of all cellular cases, and 0% (0 of 11) of the stratiform cases detected. The number of cases in each dominant class was summed to produce a total number of "correct" cases, and all cases not in the dominant class were considered "incorrect" cases. The overall percent correct was computed, this being equal to the total number of "correct" cases divided by the total number of cases minus the number of outlier cases (= 48 – 11 = 37 in this case). For the baseline classification experiment, 27 of 37 cases were correctly classified, or 72.3%.

In addition, one can determine the skill of the automated classification for separating the convective events from the non-convective events by combining the linear and cellular classes into a parent "convective" class. Again, for each of the main clusters, the dominant class was defined as either convective or non-convective based upon the highest percent of subjectively classified cases detected for that particular cluster. For instance, the dominant class for cluster #1 was non-convective, since the cluster detected 9 of the 11 (82%) cases belonging to that class. Only 4 of the 37 convective cases (10.8%) were also part of this first cluster. For cluster #2, 32.4% of the convective cases (12 of 37) were detected, while 18.2% of the stratiform cases (2 of 11) were also detected, therefore this cluster was considered to be convective dominant. Cluster #3 was also considered convective, since all 10 cases found within that cluster were subjectively classified as convective. For the two-class case, the baseline classification produced 31 of 37 correctly classified cases, or 83.7% percent correct.

These percent correct values (72.3% for the three-class and 83.7% for the two-class) will be used to compare with other classification experiments using trial attributes determined by various data reduction methods.

3.5 Data reduction experiments

3.5.1 Theoretical statistical distribution

There are a large number of potential choices of attributes that could describe the rainfall pattern over a region. An obvious choice is the amount of rainfall at every point in space obtained from a gridded analysis over the region of interest. This type of "no data reduction" classification was performed in section 3.4. Given this choice, one expects the clustering algorithm to produce groups of objects that are similar in a "point-to-point" sense. Since the goal of this work is to classify rainfall patterns, such as a heavy precipitation band oriented along a line or a disorganized collection of cellular convection, the precise locations of the maxima/minima are not of great importance. A more general characterization of the patterns may be more appropriate. Therefore, a logical choice for useful attributes might be some sort of bulk statistical measure of the overall distribution of rainfall across the region. To begin this work, the simplest choice of bulk statistical measures was selected, that is the parameters of a theoretical statistical distribution fit to the histogram representing the observed distribution of rainfall amounts across the region of each object.

As discussed in chapter 2, the distribution of rainfall tends to be highly positively skewed. Heavy rainfall is a rare event, and when large amounts of rainfall are observed the resulting distribution possesses a long "tail" (see for example figure 2-2a). On the other hand, widespread light rainfall would produce a distribution that is "humped" near a light amount of rainfall with little or no "tail" (for example, figure 2-2b). These characteristics limit the choice of theoretical distributions as potential models for the observed distribution. For this work, the gamma distribution was selected since it is positively skewed

and non-negative, provides a reasonable representation with only two parameters (α,β), and has been widely used in the meteorological literature for the analysis of precipitation data (e.g., Wilks 1990). The gamma probability density function is;

$$f(x;\alpha,\beta) = (x/\beta)^{\alpha-1}[\exp(-x/\beta)][\beta\Gamma(\alpha)]^{-1}, x \geq 0, \alpha, \beta > 0 \qquad (3.2)$$

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. For small values of $\alpha(\alpha < 1)$, 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 α, the gamma distribution is similar to the Gaussian distribution. The role of the parameter β, known as the *scale* parameter, is mainly to affect the tail of the distribution. For larger β, the distribution is "pulled" to the right, increasing the frequency of larger values of x and creating a thicker tail. For smaller β, the frequency of smaller values of x is increased, creating a thinner tail and "pushing" the distribution towards the left (see figure 2-3 for an illustration of the impact of varying α,β on the probability distribution). Refer to section 2.8 for more detailed information on the gamma distribution.

3.5.2 Determination of trace region

Precipitation observing systems, such as raingages and radar estimates, will not detect a non-zero value until the precipitation accumulation reaches some small amount. In the case of the Stage IV analysis used in this work, this detection limit is assumed to be 0.05mm, since the minimum reported value of the analysis is 0.1mm and values between 0.05 and 0.1mm are rounded up to 0.1mm by the gridded data packing routine. However, since the atmosphere is diffusive, it is physically sensible to expect a significant fraction

of locations receiving non-zero precipitation over a region will receive less than the detectable limit (also known as a "trace" amount). Studies involving high-resolution measurements of rainfall have confirmed this, for example, Hosking and Stow (1987) found that 30-40% of non-zero rain periods produced total accumulations less than the resolvable limit by conventional recording raingages. Spatially, it seems reasonable to expect that the size of the area receiving "trace" amounts of precipitation will be some fraction of the total area receiving detectable precipitation. It also seems reasonable to assume that the location of "trace" amounts of precipitation will be in the vicinity of the region receiving measurable precipitation. The determination of the number of data values below the detection limit that are associated with each rainfall "object" is critical in the estimation of the parameters of a statistical distribution. Since there are no widely used methods for estimating the size of the trace area, this area was determined by experiment.

For each of the 48 cases in the target data set, the Wilks (1990) method for maximum likelihood estimation (MLE) of the gamma distribution parameters using left-censored data was performed. Wilks (1990) showed examples of this technique on a time series of precipitation data at fixed locations, here we use the technique for the distribution of precipitation at fixed time across several locations. For more details on MLE of the gamma distribution, refer to section 2.9.2. For each case, the number of "trace" observation points (N_c) was assumed to be equal to a fraction of the total number of points greater than the detection limit across the region (N_w). This fraction (k) varied from 0 to 1, limited by the fact that N_c must be less than or equal to the total number of "zero" points available in the 128×128 domain ($= 16384 - N_w$). From the α,β parameters estimated by the MLE method, the mean absolute error (MAE) of the distribution fit to the observed histogram

was computed. The fraction k that produced the minimum MAE (k_{min}) was then determined for each case. For example, figure 3-7 shows an example of how MAE varied with

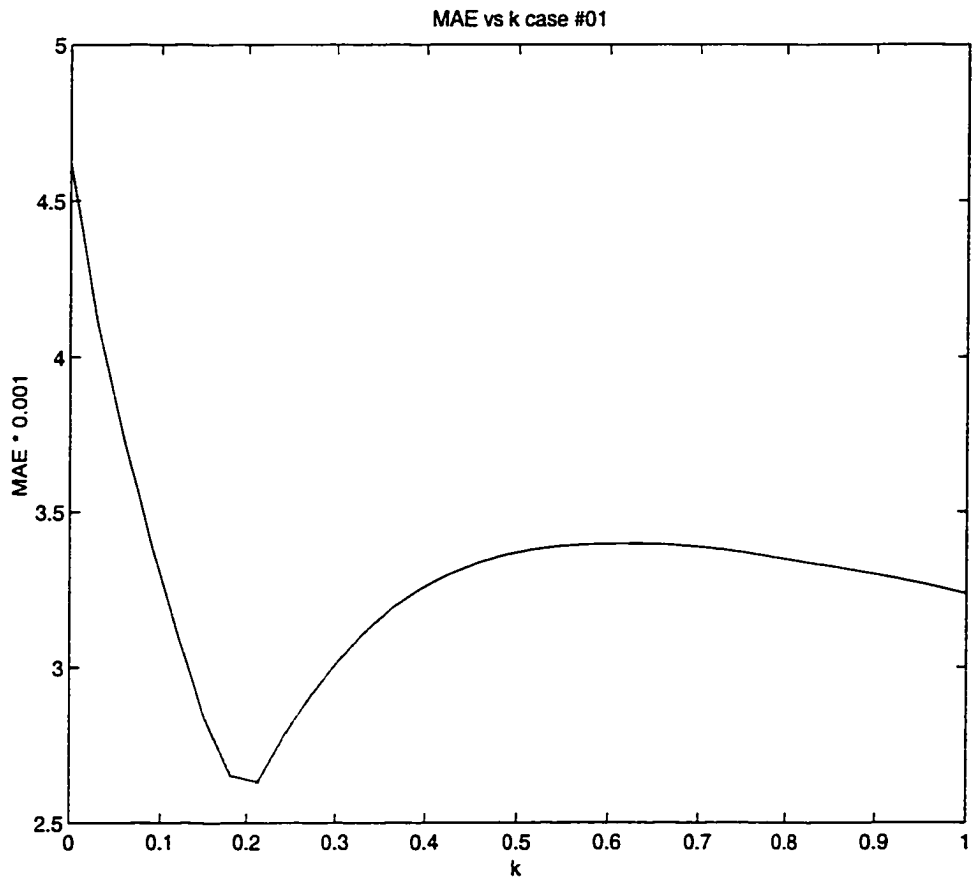


Figure 3-7: Mean absolute error of the gamma distribution fit using MLE (Wilks 1990) to the observed histogram for case #1 for 34 values of k. The k fraction is used in MLE to determine the number of censored (rainfall below the detection limit) data points ($N_c = k \times N_w$)

k for case #1. For this case, the minimum MAE was found near k=0.2. Figure 3-8 shows the distribution of k_{min} across the 48 cases in the target data set, plotted as a function of N_w. Most of the k_{min} values are clustered in the 0.1 to 0.2 range. However, note that for a minority of cases, k_{min} was close to 1.0. For many of these cases, MAE continuously

decreases as k increases. It is possible that the actual k_{min} for these cases is some value greater than 1.0, which would mean that the trace area was larger than the area that received detectable precipitation. In addition, note that many of these high k_{min} cases are

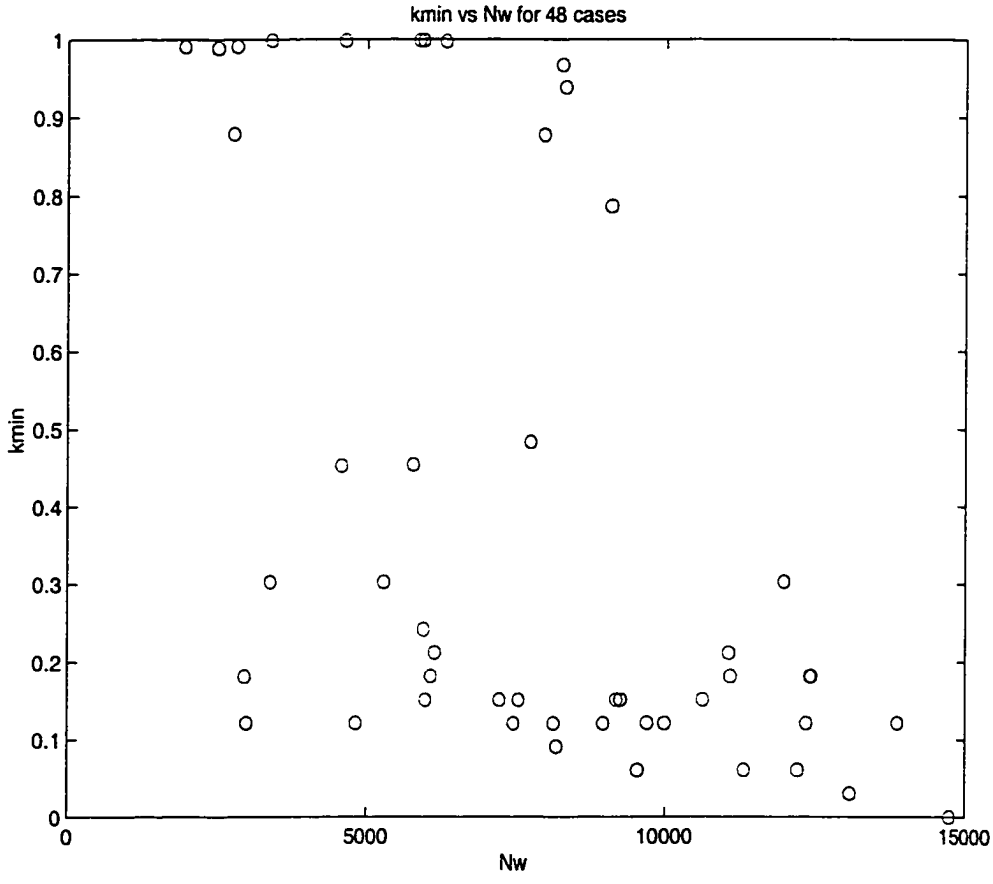


Figure 3-8: Distribution of *kmin* for the 48 cases in the target data set. Values are plotted as a function of N_w on the x-axis.

also associated with lower N_w, in other words, small areas of measurable precipitation. It seems unreasonable to expect the area observing non-zero rainfall below the detection limit should be as large or larger than the area receiving rainfall above the detection limit. Over the 48 cases, the median value for k_{min} was found to be 0.18. As an approximation to this, a value of k=0.15 was used to estimate the "trace" precipitation area for the exper-

imental results found in the remainder of this work.

In practice, k=0.15 area of trace precipitation was established in the gridded data by extending the detected rainfall region an integer number of grid points (denoted by *iext*) in each direction. This number of grid points is computed by assuming a circular area of pre-

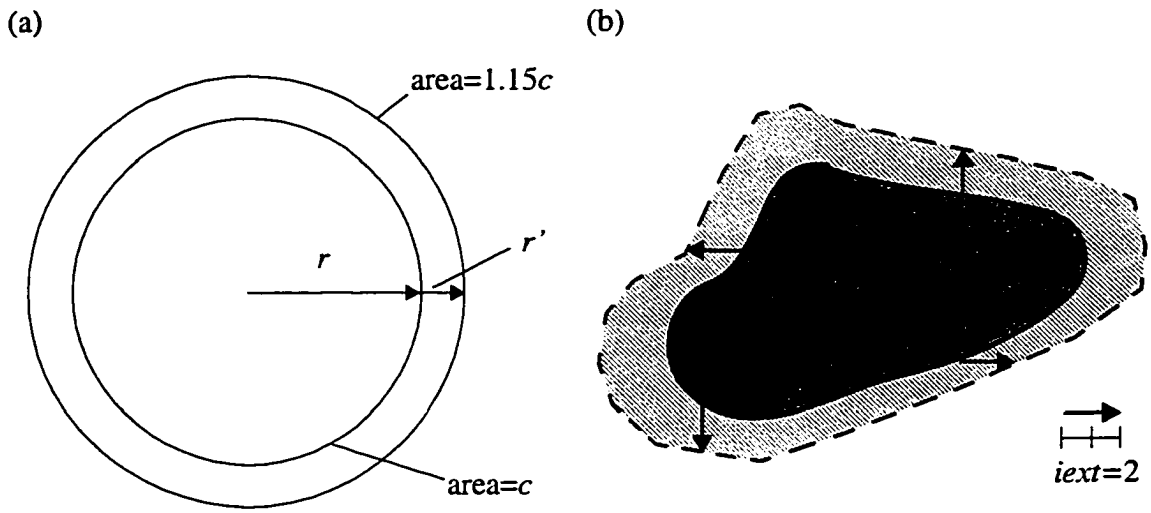


Figure 3-9: Explanation of how the trace precipitation region was applied to the region of detectable rainfall. The left panel (a) illustrates how the formula for the extension radius r' was calculated by assuming a circular region of rainfall. The right panel (b) shows how the edge of the detectable rainfall region (dark shaded) is extended by *iext* (= nearest integer value to r') grid points in each direction, represented by the arrows. In this example, *iext*=2.

cipitation (figure 3-9a). Assume the area of this circular region is = c, this could be considered to be equal to the number of grid points containing measurable precipitation if one assumes unit area for each grid point, therefore $c = N_w$. The radius of this assumed circular area is therefore $r = \sqrt{\frac{N_w}{\pi}}$. To produce a circle of radius = $r + r'$ that has an area that is 15% larger than c, one must use an extension radius

$$r' = (\sqrt{1.15} - 1)r = (\sqrt{1.15} - 1)\sqrt{\frac{N_w}{\pi}} \tag{3.3}$$

The nearest integer value to this extension radius becomes the actual number of grid points that are extended in each direction from the edge of the region of measurable precipitation (figure 3-9b).

3.5.3 Parameter estimation

Rainfall data, like many other meteorological variables, are spatially correlated. For this reason, a robust method of parameter estimation that does not rely upon an assumption of independence is desired. One such parameter estimation technique is known as the generalized method of moments, or GMM (Hansen 1982; Hamilton 1994). GMM can be formulated to allow correlation in the data to affect the parameter estimation. GMM can be considered an extension to the more familiar method of moments technique for parameter estimation. In the method of moments technique, a set of equations are developed to cover the number of unknown parameters found in the model. In the case of the gamma distribution, there are two unknown parameters, α and β, 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. In this case, parameters obtained via the method of moments technique will fit the observed mean and variance exactly, but higher-order moments will not be taken into account. In some cases, it may be desirable for the parameters to provide a better fit to the observed skewness (related to the 3rd moment) or kurtosis (related to the 4th moment). The GMM technique allows for this by adding higher-order moments to the equation set, resulting in an non-linear system of equations which can then be solved by least-squares methods. A detailed

description of GMM is provided in section 2.9.3. To my knowledge, this work is the first example of the use of GMM with rainfall data in the meteorological community.

The GMM estimates of the unknown parameters are those that minimize the *weighted* sum of squared errors of the parameter estimates. The optimal weighting matrix is the inverse of the parameter error covariance matrix. If the data are serially uncorrelated, a consistent (in the statistical sense of *consistency*, where the sample estimate approaches the true population value in the limit where the sample size approaches infinity) estimate of the error covariance matrix is the second moment matrix, $\hat{S}_T$, where:

$$\hat{S}_T = (1/T) \sum_{t=1}^{T} g([\hat{\alpha}, \hat{\beta}], w) g([\hat{\alpha}, \hat{\beta}], w)^T \tag{3.4}$$

which is the mean outer product matrix of the errors of the estimated parameters ($g([\hat{\alpha}, \hat{\beta}], w)$ as defined by eq. 2.22). Newey and West (1987) show how to modify the estimate of the second moment matrix to account for serial correlation in the data:

$$\hat{S}_T = \hat{\Gamma}_{0,T} + \sum_{v=1}^{q} \{1 - [v/(q+1)]\} (\hat{\Gamma}_{v,T} + \hat{\Gamma}^T_{v,T}) \tag{3.5}$$

where:

$$\hat{\Gamma}_{v,T} = (1/T) \sum_{t=v+1}^{T} [g([\hat{\alpha}, \hat{\beta}], w_t)][g([\hat{\alpha}, \hat{\beta}], w_{t-v})]^T \tag{3.6}$$

and q is the lag-correlation length. The question now becomes, what is the proper choice for q? Newey and West (1987) discuss how, in many studies, q is set equal to the number of non-zero autocorrelations in the data, which are known ahead of time. However, in many cases the number of non-zero autocorrelations is not known ahead of time or may in

fact be quite large. For example, figure 3-10a shows the autocorrelation function (ACF) for case #1 from the rainfall target data set when stored in vector form in row-major order. Here, the ACF drops off to zero at a lag of approximately 35, then returns to a value near

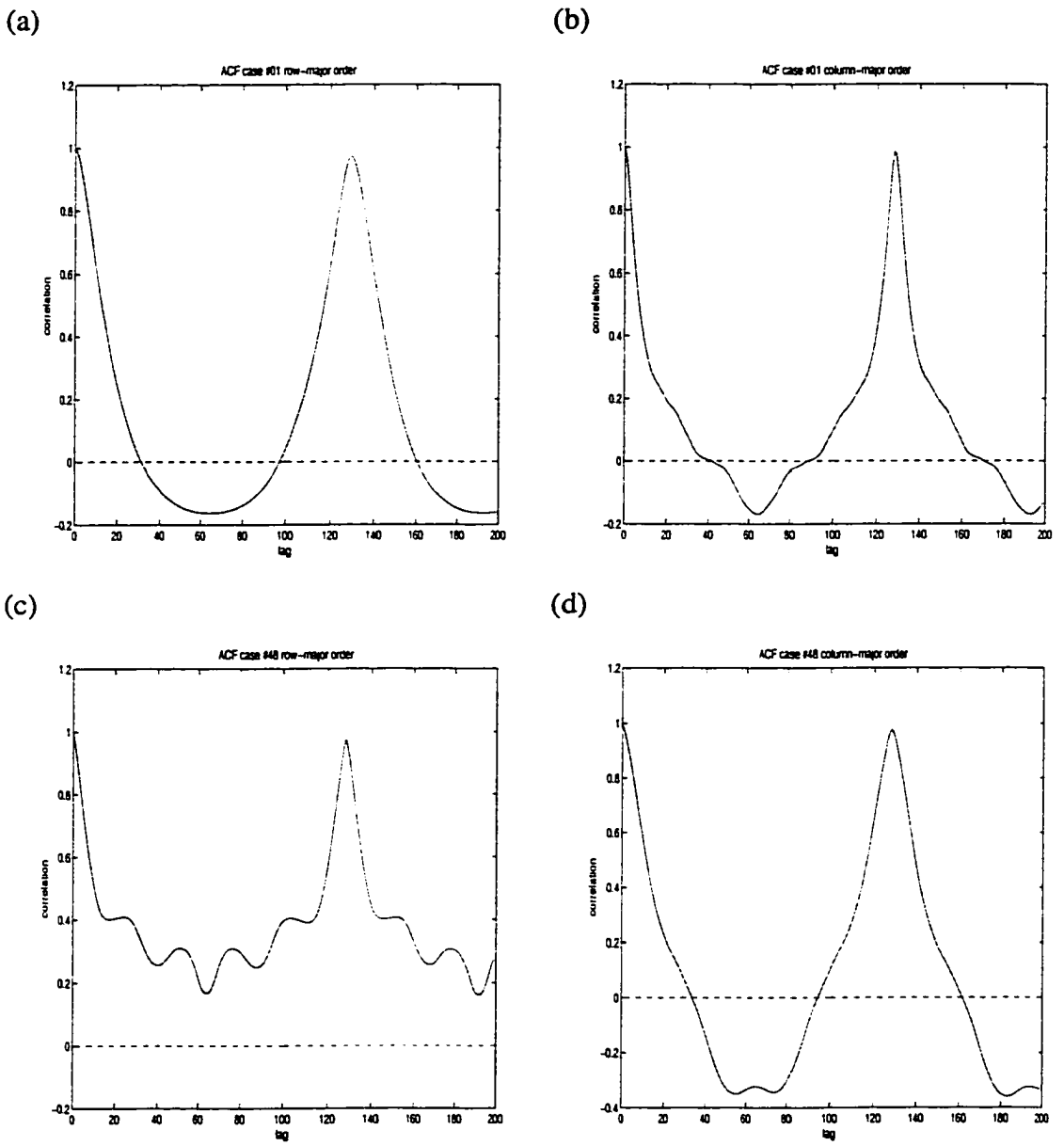


Figure 3-10: Autocorrelation function for the first 200 lags of case #1 stored in vector form in row-major order (a), case #1 in column-major order (b), case #48 in row-major order (c), and case #48 in column-major order (d).

1.0 at a lag of 128. This is due to the one-dimensional storage of the two-dimensional spa-

tial data, because the grid dimensions are 128×128, points that are separated by a lag of 128 in the vector series are physically separated by 1 grid point on the original 2-D analysis. Figure 3-10b shows how the ACF changes for this case when the two-dimensional grid is stored in vector form in column-major order. The ACF crosses zero at a lag of approximately 40 in this instance. A quite different ACF is obtained from case #48 in row-major order (figure 3-10c) where the ACF remains positive throughout the plot. When stored in column-major order (figure 3-10d) a plot similar to case #1 in row-major order is obtained, with a zero crossing near lag=35. Note that in all of these examples, the autocorrelation remains non-zero for large values of lag, therefore a proper value for the number of non-zero autocorrelations seems difficult to determine. On the other hand, Newey and West (1987) show that eq. 3.5 provides a consistent estimate of the covariance matrix even if the number of non-zero autocorrelations is not known or is not even finite, as long as q grows as a fractional power of sample size ($q < T^{1/4}$). This provides an upper bound to q (for the target data set, T=16384 therefore an upper bound for q is $16384^{1/4} \sim 11$).

In order to determine the sensitivity of the results to variations in q, several different values (q ranging from 0 to 10) were tested in estimating the gamma parameters (see Tables 3-2 through 3-6). In addition, the number of moments included in the GMM estimation varied from two to four (two moments will produce identical results to the traditional method of moments technique for any value of q). One can see that the values of α and β vary only slightly when different numbers of moments are used, and when different values for q are used. To further illustrate this, figure 3-11 shows examples of histograms from two different cases along with theoretical distribution plots for various values of q

and numbers of moments in the GMM estimation. As can be seen in this figure, the differences in gamma distribution probability distribution functions when two, three, or four moments are used (figures 3-11a and b) are quite small. In addition, almost imperceptible differences in gamma distribution plots are found for different values of q (=1, 5, 10) when three moments are used in the GMM (figures 3-11c and d) or when four moments are used (figures 3-11e and f).

3.5.4Classification results

The GMM estimates of α and β are now used as trial attributes in an automated classification (as in section 3.4) in order to find groups of similar rainfall events. The data matrix that becomes input to the cluster analysis is a $n \times m$ (n=48, m=2) matrix containing the raw, unnormalized (α,β) attributes for each of the 48 cases in the target data set. Figure 3-12 shows a dendrogram of results from the Ward's method on the target data set for the 48 cases using α, β estimated by GMM using 2 moments (see Table 3-2). In the dendrogram, there appears to be four main clusters which are separated at the breakpoint (on the y-axis) of ~3 (value indicates the square root of the increase in the sum of the squared errors caused by merging two clusters). The cases found within these four clusters are listed in Table 3-7. This result shows the cluster analysis successfully produces clusters whose members fall into the subjectively determined convective/non-convective classes. For example, clusters 1, 2, and 4 are unanimously populated by convective-type events (both linear and cellular). Cluster 3 is dominated by non-convective events, with 1 exception (case 5). Case 31 is an outlier in this example. For the 2-class case, there is only 1 "mis-classified" event out of 47, resulting in a 97.8% classification accuracy. This is much higher than the baseline experiment, which only produces 83.7% correct cases and

Table 3-2: Results of GMM parameter estimation for gamma distribution for two, three, and four moments with q=0.

	2-moments		3-moments q=0		4-moments q=0	
case	α	β	α	β	α	β
1	0.51	6.91	0.49	6.50	0.50	6.31
2	0.67	2.01	0.62	2.11	0.56	2.28
3	0.73	1.85	0.63	1.49	0.67	1.38
4	0.42	4.54	0.42	4.52	0.39	4.83
5	0.53	1.47	0.49	1.29	0.52	1.19
6	0.39	4.07	0.30	3.20	0.34	2.80
7	0.32	3.67	0.26	2.99	0.28	2.76
8	0.50	3.99	0.44	3.62	0.45	3.54
9	0.38	7.50	0.37	7.55	0.40	6.87
10	0.50	2.00	0.49	1.94	0.57	1.58
11	0.48	4.16	0.44	3.64	0.48	3.32
12	0.40	3.32	0.39	3.16	0.44	2.62
13	0.43	2.98	0.42	2.77	0.48	2.33
14	0.45	3.28	0.51	2.80	0.57	2.42
15	0.27	7.25	0.24	6.73	0.26	6.08
16	0.51	5.09	0.52	5.03	0.61	3.96
17	0.52	2.17	0.52	2.17	0.59	1.83
18	0.17	8.34	0.16	8.18	0.17	7.33
19	0.62	2.33	0.59	2.44	0.55	2.58
20	0.36	4.44	0.33	4.33	0.35	4.03
21	0.68	3.08	0.67	3.13	0.65	3.21
22	0.25	5.18	0.19	4.46	0.20	4.21
23	0.39	2.83	0.40	2.72	0.40	2.71
24	0.42	8.26	0.38	7.40	0.39	7.09
25	0.57	3.56	0.53	3.47	0.53	3.47
26	0.53	3.92	0.46	3.50	0.48	3.34
27	0.59	3.33	0.52	3.11	0.49	3.31
28	0.53	2.83	0.51	2.68	0.53	2.56
29	0.61	2.41	0.54	2.35	0.51	2.48
30	0.17	5.41	0.17	5.40	0.17	5.48
31	0.14	12.18	0.14	11.99	0.15	10.75
32	0.26	6.08	0.31	5.13	0.32	4.94
33	0.62	2.04	0.57	2.03	0.54	2.18
34	0.22	6.41	0.22	6.44	0.28	4.44
35	0.49	3.22	0.49	3.18	0.48	3.25
36	0.54	3.00	0.52	2.62	0.58	2.24
37	0.31	5.19	0.30	5.17	0.35	4.17
38	0.47	1.25	0.50	1.17	0.53	1.10
39	0.53	0.90	0.52	0.92	0.52	0.92
40	0.54	0.67	0.52	0.62	0.58	0.54
41	0.77	0.47	0.76	0.47	0.83	0.42
42	0.57	0.59	0.59	0.57	0.67	0.48
43	0.64	0.61	0.61	0.63	0.62	0.62
44	0.45	1.18	0.71	0.73	0.81	0.64
45	1.65	0.77	1.65	0.70	1.61	0.72
46	0.51	1.11	0.59	0.96	0.66	0.83
47	0.59	0.71	0.66	0.63	0.71	0.56
48	1.16	0.92	1.11	0.94	1.05	1.00

Table 3-3: Results of GMM parameter estimation for gamma distribution for three moments with q=1, 2, and 3.

	3-moments q=1		3-moments q=2		3-moments q=3	
case	α	β	α	β	α	β
1	0.48	6.46	0.48	6.43	0.48	6.40
2	0.62	2.11	0.62	2.10	0.62	2.09
3	0.63	1.47	0.63	1.46	0.62	1.45
4	0.42	4.52	0.42	4.52	0.42	4.52
5	0.48	1.28	0.48	1.28	0.48	1.28
6	0.30	3.16	0.30	3.13	0.29	3.10
7	0.26	2.97	0.25	2.94	0.25	2.92
8	0.44	3.57	0.43	3.52	0.43	3.49
9	0.37	7.54	0.37	7.53	0.37	7.51
10	0.49	1.92	0.49	1.91	0.49	1.90
11	0.44	3.60	0.44	3.57	0.44	3.55
12	0.39	3.14	0.39	3.11	0.39	3.10
13	0.42	2.73	0.42	2.70	0.42	2.68
14	0.52	2.75	0.53	2.71	0.53	2.68
15	0.24	6.63	0.24	6.58	0.24	6.54
16	0.52	5.03	0.52	5.04	0.52	5.05
17	0.52	2.17	0.52	2.17	0.52	2.17
18	0.16	8.13	0.16	8.07	0.16	8.02
19	0.59	2.44	0.59	2.44	0.59	2.44
20	0.33	4.32	0.33	4.31	0.32	4.30
21	0.67	3.13	0.67	3.12	0.67	3.12
22	0.19	4.31	0.19	4.28	0.18	4.28
23	0.40	2.72	0.40	2.71	0.40	2.71
24	0.38	7.37	0.37	7.36	0.37	7.35
25	0.52	3.45	0.52	3.42	0.52	3.39
26	0.46	3.49	0.46	3.48	0.46	3.46
27	0.52	3.06	0.52	3.02	0.52	2.98
28	0.51	2.65	0.51	2.63	0.51	2.61
29	0.54	2.33	0.54	2.31	0.53	2.29
30	0.17	5.40	0.17	5.40	0.17	5.40
31	0.14	11.94	0.14	11.87	0.14	11.83
32	0.31	5.05	0.31	4.99	0.31	4.96
33	0.57	2.01	0.57	2.00	0.57	1.98
34	0.22	6.46	0.22	6.47	0.22	6.47
35	0.49	3.18	0.49	3.18	0.49	3.18
36	0.52	2.61	0.52	2.60	0.52	2.59
37	0.30	5.14	0.30	5.13	0.30	5.11
38	0.50	1.16	0.51	1.15	0.51	1.15
39	0.52	0.92	0.52	0.92	0.52	0.92
40	0.51	0.62	0.51	0.62	0.51	0.62
41	0.76	0.46	0.77	0.46	0.77	0.46
42	0.59	0.57	0.59	0.57	0.59	0.58
43	0.62	0.62	0.62	0.61	0.62	0.61
44	0.73	0.71	0.74	0.70	0.75	0.69
45	1.65	0.70	1.64	0.70	1.64	0.69
46	0.59	0.94	0.60	0.93	0.60	0.92
47	0.67	0.62	0.67	0.61	0.67	0.61
48	1.11	0.94	1.11	0.93	1.11	0.93

Table 3-4: Results of GMM parameter estimation for gamma distribution for three moments with q=4, 5, and 10.

	3-moments q=4		3-moments q=5		3-moments q=10	
case	α	β	α	β	α	β
1	0.48	6.37	0.48	6.34	0.48	6.23
2	0.62	2.08	0.62	2.07	0.61	2.04
3	0.62	1.44	0.62	1.44	0.60	1.42
4	0.42	4.52	0.42	4.52	0.42	4.51
5	0.48	1.27	0.47	1.27	0.46	1.26
6	0.29	3.08	0.29	3.06	0.27	3.05
7	0.25	2.90	0.24	2.88	0.23	2.86
8	0.42	3.48	0.42	3.47	0.40	3.49
9	0.37	7.50	0.37	7.48	0.37	7.43
10	0.49	1.89	0.49	1.88	0.49	1.87
11	0.43	3.54	0.43	3.53	0.42	3.48
12	0.39	3.08	0.39	3.07	0.39	3.04
13	0.42	2.66	0.42	2.65	0.43	2.61
14	0.53	2.65	0.54	2.62	0.55	2.51
15	0.24	6.53	0.23	6.53	0.22	6.57
16	0.52	5.06	0.52	5.07	0.51	5.13
17	0.52	2.17	0.52	2.17	0.52	2.17
18	0.16	7.96	0.16	7.92	0.16	7.80
19	0.59	2.44	0.59	2.43	0.59	2.41
20	0.32	4.29	0.32	4.28	0.32	4.23
21	0.67	3.12	0.67	3.12	0.67	3.11
22	0.18	4.30	0.18	4.31	0.17	4.38
23	0.40	2.71	0.40	2.70	0.40	2.71
24	0.37	7.35	0.36	7.35	0.35	7.37
25	0.52	3.37	0.52	3.35	0.51	3.29
26	0.45	3.45	0.45	3.44	0.44	3.40
27	0.51	2.96	0.51	2.93	0.50	2.88
28	0.51	2.59	0.51	2.57	0.51	2.53
29	0.53	2.27	0.53	2.26	0.52	2.21
30	0.17	5.40	0.17	5.40	0.17	5.40
31	0.14	11.81	0.14	11.79	0.13	11.76
32	0.31	4.94	0.31	4.93	0.31	4.93
33	0.56	1.97	0.56	1.96	0.55	1.92
34	0.22	6.48	0.22	6.48	0.22	6.48
35	0.49	3.18	0.49	3.18	0.49	3.18
36	0.52	2.58	0.52	2.58	0.51	2.55
37	0.30	5.10	0.30	5.08	0.30	5.04
38	0.51	1.14	0.51	1.14	0.51	1.14
39	0.52	0.91	0.52	0.91	0.52	0.90
40	0.51	0.62	0.51	0.62	0.50	0.61
41	0.77	0.46	0.77	0.46	0.77	0.45
42	0.59	0.58	0.59	0.59	0.58	0.60
43	0.62	0.60	0.62	0.60	0.62	0.60
44	0.75	0.69	0.76	0.68	0.76	0.68
45	1.64	0.69	1.64	0.69	1.61	0.68
46	0.61	0.91	0.61	0.91	0.62	0.89
47	0.68	0.61	0.68	0.60	0.69	0.59
48	1.11	0.92	1.11	0.92	1.11	0.90

Table 3-5: Results of GMM parameter estimation for gamma distribution for four moments with q=1, 2, and 3.

	4-moments q=1		4-moments q=2		4-moments q=3	
case	α	β	α	β	α	β
1	0.50	6.25	0.50	6.20	0.50	6.16
2	0.55	2.29	0.55	2.28	0.55	2.27
3	0.67	1.35	0.67	1.33	0.67	1.32
4	0.39	4.89	0.39	4.93	0.38	4.96
5	0.52	1.18	0.52	1.17	0.51	1.17
6	0.34	2.72	0.34	2.65	0.34	2.60
7	0.28	2.71	0.28	2.66	0.27	2.61
8	0.45	3.45	0.44	3.39	0.44	3.36
9	0.40	6.79	0.40	6.73	0.40	6.68
10	0.58	1.54	0.58	1.51	0.58	1.49
11	0.48	3.25	0.47	3.21	0.47	3.18
12	0.45	2.55	0.45	2.49	0.46	2.45
13	0.48	2.28	0.48	2.24	0.49	2.21
14	0.58	2.40	0.58	2.38	0.58	2.38
15	0.26	5.87	0.26	5.75	0.26	5.66
16	0.61	3.93	0.62	3.90	0.62	3.87
17	0.59	1.80	0.60	1.78	0.60	1.76
18	0.17	7.21	0.17	7.08	0.17	6.97
19	0.54	2.61	0.54	2.64	0.53	2.65
20	0.35	3.99	0.35	3.95	0.35	3.92
21	0.65	3.22	0.65	3.22	0.65	3.22
22	0.20	4.02	0.20	3.92	0.20	3.86
23	0.40	2.72	0.40	2.72	0.40	2.72
24	0.39	7.03	0.39	6.98	0.39	6.93
25	0.53	3.41	0.53	3.36	0.53	3.32
26	0.48	3.31	0.48	3.29	0.48	3.26
27	0.49	3.24	0.49	3.18	0.49	3.13
28	0.54	2.51	0.54	2.47	0.54	2.44
29	0.51	2.46	0.51	2.43	0.51	2.40
30	0.17	5.48	0.17	5.47	0.17	5.45
31	0.15	10.53	0.15	10.43	0.15	10.38
32	0.32	4.91	0.32	4.89	0.32	4.87
33	0.53	2.16	0.53	2.14	0.53	2.12
34	0.29	4.28	0.29	4.19	0.29	4.14
35	0.48	3.23	0.48	3.20	0.48	3.18
36	0.58	2.22	0.58	2.20	0.58	2.19
37	0.36	4.06	0.36	3.98	0.36	3.91
38	0.53	1.10	0.53	1.09	0.53	1.09
39	0.52	0.91	0.53	0.90	0.53	0.89
40	0.58	0.53	0.58	0.52	0.58	0.52
41	0.84	0.41	0.85	0.40	0.86	0.39
42	0.68	0.47	0.68	0.47	0.69	0.46
43	0.63	0.60	0.63	0.60	0.63	0.59
44	0.83	0.62	0.84	0.61	0.84	0.61
45	1.61	0.72	1.61	0.72	1.61	0.71
46	0.66	0.81	0.67	0.80	0.67	0.80
47	0.72	0.56	0.72	0.56	0.72	0.55
48	1.05	0.99	1.05	0.99	1.05	0.98

Table 3-6: Results of GMM parameter estimation for gamma distribution for four moments with q=4, 5, and 10.

	4-moments q=4		4-moments q=5		4-moments q=10	
case	α	β	α	β	α	β
1	0.50	6.12	0.50	6.09	0.49	6.01
2	0.55	2.26	0.55	2.24	0.54	2.19
3	0.67	1.31	0.67	1.30	0.65	1.28
4	0.38	4.99	0.38	5.01	0.37	5.10
5	0.51	1.16	0.51	1.15	0.50	1.13
6	0.34	2.55	0.33	2.52	0.32	2.42
7	0.27	2.58	0.27	2.55	0.26	2.46
8	0.43	3.35	0.43	3.34	0.41	3.33
9	0.40	6.64	0.40	6.61	0.40	6.57
10	0.58	1.48	0.59	1.47	0.59	1.45
11	0.47	3.16	0.47	3.14	0.46	3.08
12	0.46	2.41	0.46	2.38	0.46	2.29
13	0.49	2.19	0.49	2.17	0.49	2.13
14	0.58	2.37	0.58	2.36	0.58	2.32
15	0.26	5.62	0.26	5.58	0.25	5.56
16	0.62	3.84	0.62	3.82	0.63	3.73
17	0.61	1.74	0.61	1.72	0.62	1.68
18	0.17	6.88	0.17	6.82	0.17	6.63
19	0.53	2.66	0.53	2.65	0.52	2.62
20	0.35	3.89	0.35	3.86	0.34	3.79
21	0.65	3.22	0.65	3.22	0.65	3.21
22	0.20	3.82	0.20	3.79	0.20	3.71
23	0.40	2.72	0.40	2.72	0.40	2.73
24	0.38	6.89	0.38	6.86	0.37	6.74
25	0.53	3.29	0.53	3.27	0.53	3.20
26	0.48	3.24	0.48	3.22	0.47	3.15
27	0.49	3.10	0.49	3.07	0.48	3.00
28	0.54	2.42	0.54	2.40	0.54	2.36
29	0.51	2.38	0.51	2.36	0.50	2.30
30	0.17	5.44	0.17	5.43	0.17	5.38
31	0.15	10.34	0.15	10.30	0.15	10.21
32	0.32	4.87	0.32	4.87	0.32	4.88
33	0.53	2.10	0.53	2.09	0.51	2.06
34	0.29	4.11	0.29	4.10	0.29	4.11
35	0.48	3.17	0.48	3.15	0.48	3.10
36	0.58	2.17	0.58	2.16	0.58	2.10
37	0.36	3.86	0.37	3.82	0.37	3.71
38	0.53	1.09	0.53	1.09	0.53	1.09
39	0.53	0.89	0.53	0.88	0.53	0.87
40	0.58	0.52	0.58	0.51	0.57	0.51
41	0.87	0.39	0.87	0.39	0.88	0.38
42	0.69	0.46	0.69	0.45	0.70	0.44
43	0.64	0.59	0.64	0.58	0.63	0.58
44	0.85	0.61	0.85	0.60	0.86	0.60
45	1.60	0.71	1.60	0.71	1.57	0.70
46	0.68	0.79	0.68	0.78	0.69	0.76
47	0.72	0.55	0.72	0.55	0.73	0.54
48	1.05	0.97	1.06	0.97	1.06	0.95

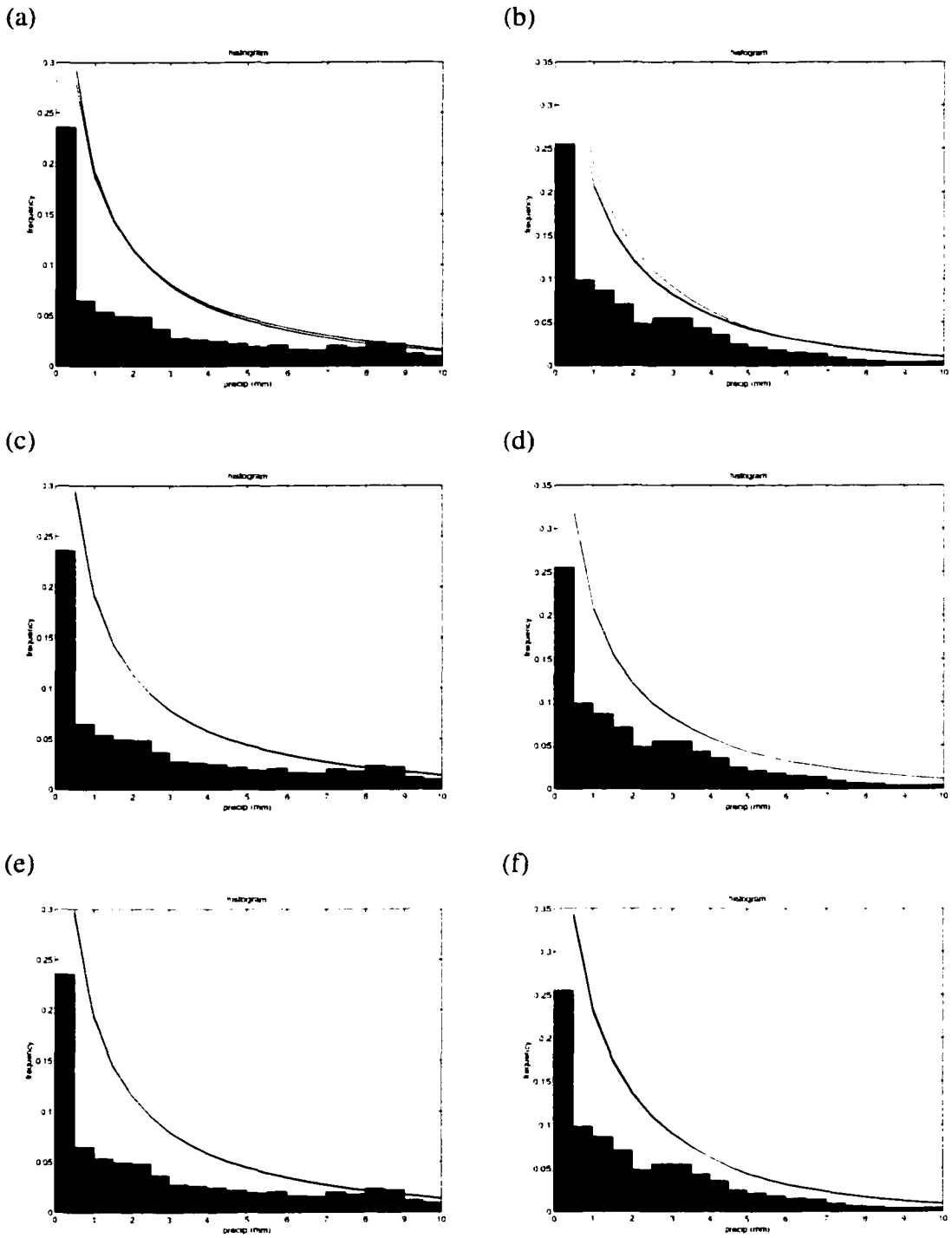


Figure 3-11: Sample histograms and theoretical distributions fit using GMM. Left column is for case #1, right column is for case #16. Gamma distributions for plots in top row (a, b) are for two (blue), three (green), and four (red) moments using q=0. For plots in the middle row (c, d) three moments are used in GMM with q=1 (blue), q=5 (green) and q=10 (red). Plots in the bottom row (e, f) are as in middle row, except for four moments in the GMM estimation.

Table 3-7: Cluster membership for the 2-moment experiment

Cluster # (# of members)	cases
1 (10)	4, 6, 8, 11, 16, 20, 22, 26, 30, 37
2 (18)	2, 3, 7, 10, 12, 13, 14, 17, 19, 21, 23, 25, 27, 28, 29, 33, 35, 36
3 (12)	5, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48
4 (7)	1, 9, 15, 18, 24, 32, 34

also had a higher number of outliers. Since this is a two-dimensional problem, we can easily visualize the clustering by plotting each case on the α–β plane, with the clusters indicated by a contour around the cases involved. (Figure 3-13). In this figure, there is a threshold value of β (~1.5) that cleanly separates the three "convective" clusters from the "non-convective" cluster.

Now we examine how well the cluster analysis classifies the cases into three classes (linear, cellular, stratiform). Returning to the 2-moment experiment (Figure 3-12, Table 3-7), cluster 1 contains five cases that were subjectively classified as linear and five that were subjectively classified as cellular precipitation events. Cluster 2 is also split among the linear and cellular precipitation events with eight linear cases and ten cellular events. Cluster 3 contains mainly stratiform (11) events, with one linear event included. Cluster 4 contains four linear events and three cellular events. Overall, this experiment placed 30 of the 47 cases into correct dominant classes, for a 63.8% percent correct. This result is worse than the baseline experiment which correctly classified 72.3% of the cases into the dominant class for each cluster. These results show that the cluster analysis did not produce groups within the convective class with a clear preference for a particular sub-class (linear, cellular) in this experiment.

These results were consistent with those found by using three and four moments in

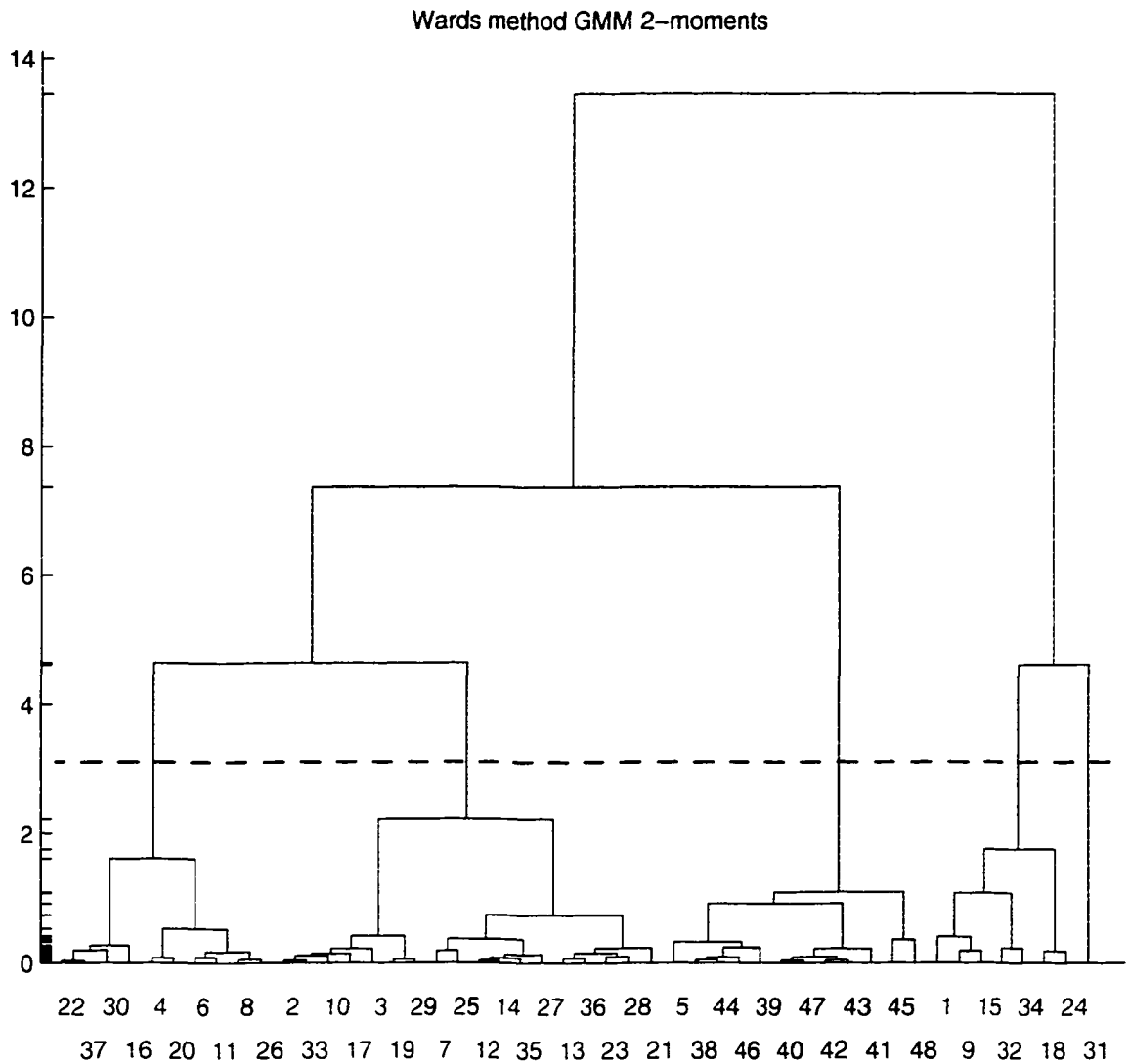


Figure 3-12:Dendrogram produced by Ward's method with target data set using 2-moment GMM [α, β] as attributes. Each object is color-coded by its subjective classification, linear events are green, cellular events are red, stratiform events are blue. Dashed line indicates subjectively determined cut-level for this experiment.

the GMM, and by increasing q from 0 to 10, as summarized in Table 3-8. The automated classification shows some sensitivity to the choice of moments used in the parameter estimation. However, it does not appear to be very sensitive to the choice of q, since the estimated α and β values vary only slightly q changes (see Tables 3-2 through 3-6). For the three-moment GMM experiment, the cluster analysis produced identical main clusters for

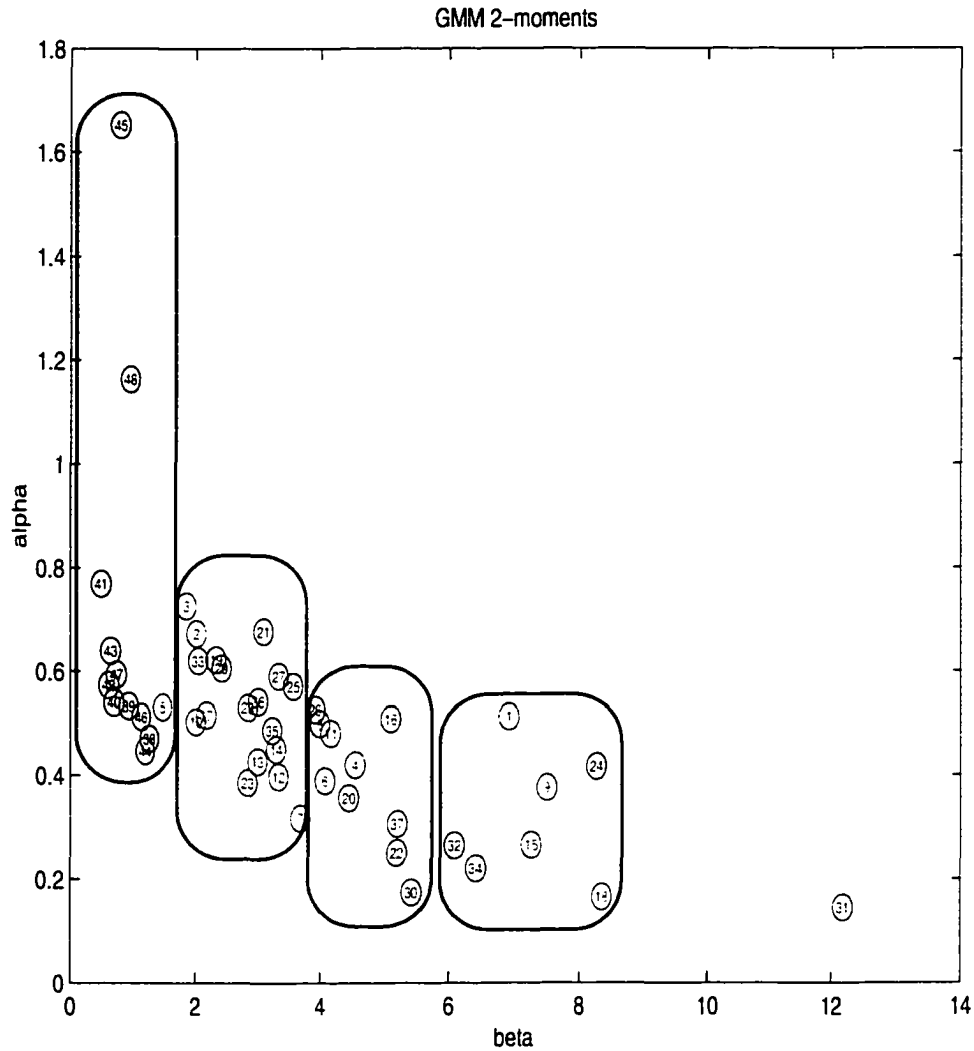


Figure 3-13: Distribution of objects in target data set in [α, β] space for 2-moment GMM. Each object is color-coded by its subjective classification, linear events are green, cellular events are red, stratiform events are blue. Contours indicate clusters found in figure 3-12.

Table 3-8: Summary of results of automated classification using attributes from GMM.

experiments	2-class percent correct	3-class percent correct
2-moments	97.8%	63.8%
3-moments; q=0, 1, 2, 3, 4, 5, 10	95.7%	66.0%
4-moments; q=0	91.5%	68.1%
4-moments; q=1, 2, 3, 4, 5, 10	91.5%	66.0%

all values of q from 0 to 10 (Table 3-9). The dendrogram for the three-moment q=1 exper-

Table 3-9: Cluster membership for the 3-moment experiments, all values of q (0-5, 10).

Cluster # (# of members)	cases
1 (21)	2, 6, 7, 8, 10, 11, 12, 13, 14, 17, 19, 21, 23, 25, 26, 27, 28, 29, 33, 35, 36
2 (13)	3, 5, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48
3 (6)	1, 9, 15, 18, 24, 34
4 (7)	4, 16, 20, 22, 30, 32, 37

iment is provided in figure 3-14a; other experiments using other values of q produced similar looking plots. The percent correct for these experiments was slightly lower than the two-moment experiment for the two-class (convective/non-convective) case (95.7% or 45 of 47 correct), and slightly higher for the three-class case (66.0% or 31 of 47 correct). On the other hand, the four-moment GMM experiments did show some sensitivity to variations in q, although this did not have a large impact on the overall percent correct. For example, one case (#30) that was subjectively classified as cellular switched from a cellular dominant cluster in the q=0 experiment (figure 3-14b) to a linear dominant cluster for the q=1 (figure 3-14c). This caused the percent correct for the three-class (linear, cellular, stratiform) situation to drop from 68.1% to 66.0%. The membership of the first two clusters changed from the q=1 to q=2 experiments (figures 3-14c-d), then remained the same for all values of q greater than 2 (not shown). However, this did not affect the overall percent correct. The percent correct for the two-class case remained the same for all values of q in the four-moment experiments (91.5% or 43 of 47 correct). Among all of these experiments, the GMM estimation using three moments (first, second, and third) showed the

least amount of sensitivity to variations in q, and provided percent correct values greater than 95% for the two-class case, and greater than 65% in the three-class case. None of the two-moment or four-moment experiments met these levels of performance for both degrees of classification hierarchy.

3.6 Summary

The automated classification algorithm using attributes produced by analysis of the observed histograms successfully separated the target data set into convective and non-convective classes. However, looking at the next level of classification hierarchy, the main clusters contained a fairly even split of linear and cellular events within the parent convective class. This should not come as a surprise since the attributes used here (α, β) only contain information about the overall distribution of rainfall within the object. They do not provide any information on the organization or relative position of rainfall values within each object. It is therefore reasonable to expect that additional attributes are required in order to increase the degrees of freedom and allow the classification system to identify finer and more specific classes of events. Work of this type will be performed in the next chapter.

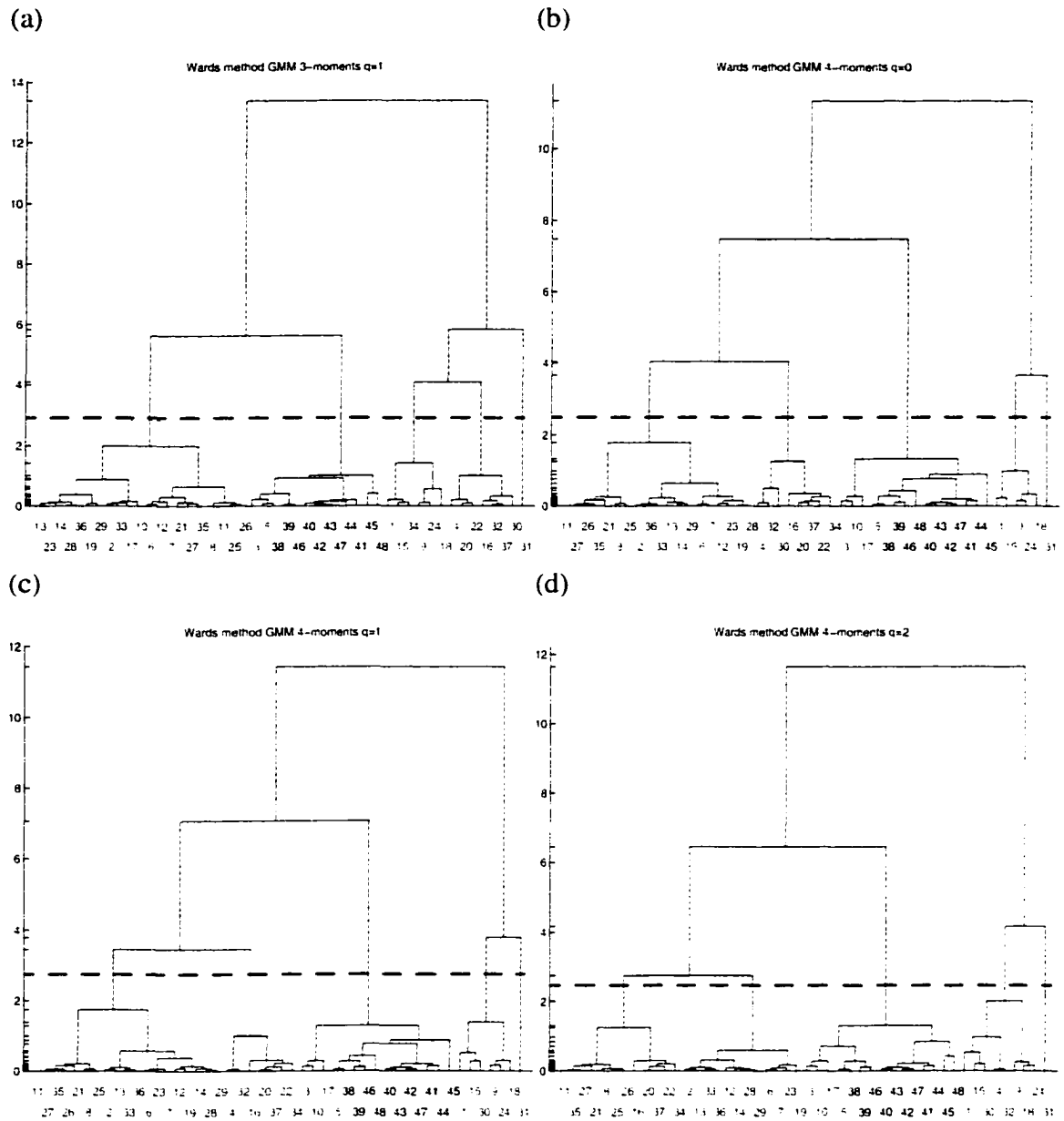


Figure 3-14: Dendrograms produced by Ward's method with target data set using GMM [α, β] as attributes. The three-moment, q=1 experiment (a), four-moment q=0 (b), q=1 (c), and q=2 (d) experiments are shown. Each object is color-coded by its subjective classification, linear events are green, cellular events are red, stratiform events are blue. Dashed lines indicate subjectively determined cut-levels for each experiment.

Chapter 4

Spatial analysis

4.1 Introduction

The specific purpose of this research is to determine those attributes which are most useful in an automated rainfall pattern classification system. A preliminary target data set has been collected to test various trial attributes. This data set consists of hourly accumulated rainfall analyses from an operational analysis system. A set of 48 separate precipitation events which occurred at various times and locations across the United States were selected for inclusion in this data set. Cases were included based upon the existence of "typical" rainfall patterns that are often found; in particular, linear, cellular, and stratiform precipitation events. Note that the linear and cellular classes are sub-classes of the parent convective class, while the stratiform class is a sub-class of the non-convective class in the overall classification hierarchy. Each case was subjectively classified into these main classes to allow for validation of various automated classification experiments.

The first step in this multi-faceted process involved a "baseline" automated classification using the raw values of rainfall at each point in space as attributes; in other words, no data reduction was performed. This classification experiment resulted in percent correct values of 72.3% for the three-class (linear, cellular, stratiform) and 83.7% for the two-class (convective, non-convective) case. Next, the dimension of the data was reduced by analyzing the "bulk" global distribution of rainfall values across each object, using the parameters of the gamma distribution fit to the observed histogram using the generalized method of moments technique. The automated classification experiments using these

attributes were able to successfully separate the convective events from the non-convective events with over 95% accuracy. However, these attributes proved to be less successful in further separating the convective cases into linear/cellular classes, as the percent correct for the three-class case dropped to approximately 65%. This was likely due to the fact that these attributes are only able to describe the overall distribution of rainfall, and not how the rainfall amounts were organized spatially.

In order to obtain summary information on the spatial continuity of rainfall, statistical measures that are a function of the *location* as well as the amount of rainfall are required. There is a long history of research using geostatistical tools to examine the characteristics of spatial radar/rainfall data. For example, Kessler and Russo (1963) and Kessler (1966) computed the two-dimensional auto-correlation of radar reflectivity. Kessler and Russo (1963) noted how the ellipticity of the auto-correlation 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. In this chapter, results of automated classification experiments using similar attributes related to the linear organization of the spatial field are presented. Using attributes of this type that summarize the geostatistical characteristics of the rainfall pattern, the classification system is able to separate the linear and cellular events, with the percent correct for the three-class case increasing to over 90%. Much of the work described in this chapter was included in Baldwin and Lakshmivarahan (2003).

4.2 Geostatistics

In the previous chapter, results from an automated classification using attributes related to the overall distribution of rainfall across an object were presented. These results

showed that useful information on the intensity of rainfall within each object was obtained through the use of these attributes. However, it was determined that these attributes do not provide information on the spatial continuity and variability of the rainfall within an object. For example, identical histograms could be obtained from events that are either randomly unorganized or spatially continuous, since the distribution ignores information on the *location* of rainfall amounts. In order 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 are several measures of spatial variability and continuity to choose from (Isaaks and Srivastava 1989; Deutsch and Journel 1988), three were used in this work: two-dimensional plots of the *semivariogram, correlogram, and covariance.* All three measure aspects of the spatial field as a function of a two-dimensional separation vector h (see figure 2-4). All possible pairs of values that are separated by h within an object will be used to compute the various statistics. The semivariogram $\gamma(h)$ is defined as half of the average squared difference between the pairs of all values separated by h (eq. 4.1). The covariance $C(h)$ is the traditional covariance (eq. 4.2) between all possible pairs of "tail" and "head" values separated by h. The correlogram $\rho(h)$ is also known as the auto-correlation, which is the correlation between all possible pairs of "tail" and "head" values separated by h (eq. 4.3).

$$\gamma(h) = \left(\frac{1}{2N(h)}\right) \sum_{N(h)} (w_t - w_h)^2 \tag{4.1}$$

$$C(h) = \left(\frac{1}{N(h)}\right) \sum_{N(h)} (w_t w_h - m_t m_h) \tag{4.2}$$

$$\rho(h) = \frac{C(h)}{\sigma_t \sigma_h} \tag{4.3}$$

Here m_t and m_h are the means of the tail and head values, respectively, and σ_t and σ_h are the standard deviations of the tail and head values, respectively. N(h) is the total number of possible pairs of tail and head values for a given separation vector. A more detailed summary of geostatistical measures is provided in section 2.10.

4.3 Synthetic data

The type of information that can be obtained about the continuity and variability of rainfall patterns from these geostatistics will be illustrated via simulated examples. In these examples, adapted from Baldwin et al (2001), *synthetic* precipitation fields are generated using an elliptical shape function (modified from Williamson 1981).

$$r = \left([(x-x_o)\cos\gamma + (y-y_o)\sin\gamma]^2 + \varepsilon[(y-y_o)\cos\gamma - (x-x_o)\sin\gamma]^2\right)^{1/2} \qquad (4.4)$$

$$p(x, y) = \frac{A}{4}\left(1 + \cos\left(\pi\frac{r}{R}\right)\right)^2 \quad r < R$$

$$p(x, y) = 0 \quad r \geq R \qquad (4.5)$$

An individual elliptical "blob" of precipitation is determined by the amplitude A multiplied by the given radial function (eq. 4.5), where the remaining function is equal to one at the center of the blob (r=0) and zero where the radius r is greater than the size parameter R. The center of the blob is given by (x_o, y_o), the orientation of the ellipse (angle between the major axis and the x-axis) is γ, and the ratio of the semi-major and semi-minor axes is $\sqrt{\varepsilon}$. Therefore, each rainfall blob is determined by six parameters (A, R, x_o, y_o, γ, ε), and the entire rainfall field is determined by summing the rainfall contribution from a number of individual elliptical features.

An example of a single rainfall feature, defined by A=10, R=60, x_o=64, y_o=64,

γ=30°, ε=3, is shown in figure 4-1a. Corresponding plots of the semivariogram, covariance, and correlogram for this event (figure 4-1 b-d) provide fairly consistent information;

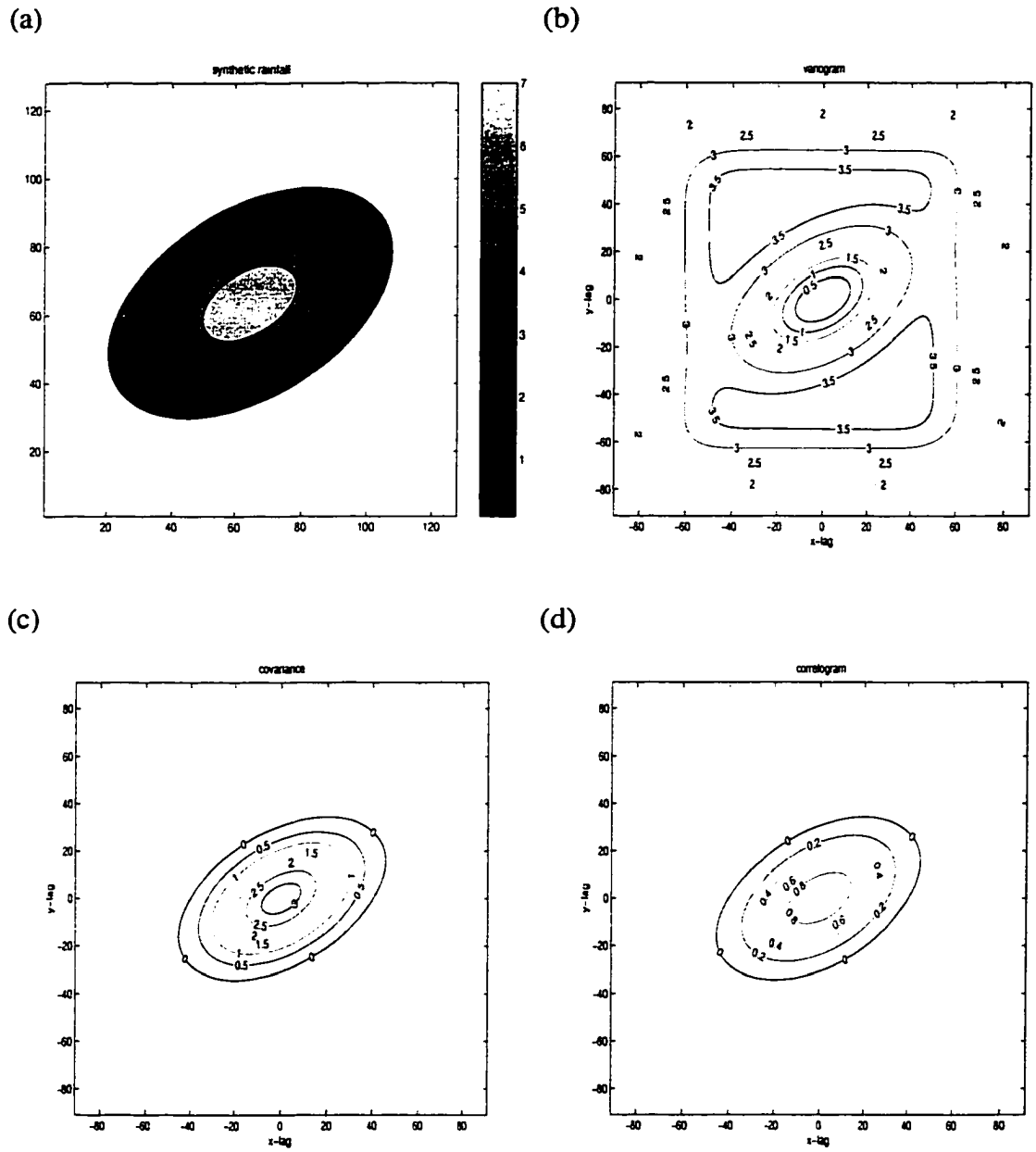


Figure 4-1: Synthetic rainfall field (a) of arbitrary units. Variogram (b), covariance (c), and correlogram (d) plots corresponding to given synthetic rainfall field, lags indicated on axes are in terms of arbitrary unit grid spacing from the original field.

rainfall values are similar over a relatively large distance in the direction along the major

axis of the rainfall ellipse and similar to other values over a relatively short distance in other directions. The semivariogram (figure 4-1b) provides information on the average squared difference between head/tail values, therefore the statistic equals zero at the origin (h=(0,0)) and its magnitude increases as h moves further from the origin. The covariance (figure 4-1c) plot works in the opposite sense, indicating how pairs of values simultaneously vary from their means, the value at the origin is the variance of the overall field. The correlogram (figure 4-1d) operates in a similar fashion to the covariance plot, except the value at the origin is normalized to 1.0.

To confirm that the ellipse produced by equation 4.5 using these parameters embodies the characteristics that were outlined in the previous paragraph, the orientation and ellipticity of the rainfall field itself will be analyzed using image processing algorithms. First, the rainfall feature is considered an object, and all contiguous grid points with rainfall greater than 0.1 are given the same object label using the connected component labeling algorithm outlined in section 2.12. Next, the edge of this connected region is found using the edge detection algorithm also outlined in section 2.12. These processes equate to locating the 0.1 contour on the rainfall plot (the first contour in figure 4-1a). Once this is determined, the largest distance from the center of the object to this edge is found, and this distance is assumed to be the length of the semi-major axis (a) of the ellipse. In this case, a=47.5. The shortest distance from the center of the region to the edge is found next, and this is assumed to be the length of the semi-minor axis (b=25.8). As discussed previously, $\sqrt{\varepsilon}$ in equation 4.4 represents the ratio of the semi-major and semi-minor axes, since ε=3, this was specified as $\sqrt{3} = 1.73$. In this case, this ratio was measured using the discrete a, b obtained via the image processing routines as 47.5/25.8=1.84, approxi-

mately 6% larger than the exact value. Since the discrete semi-major axis has been determined, the angle between it and the x-axis can be found, and in this case that angle is 30.3°, resulting in an error of ~1% when compared to the exact value. These results confirm that the characteristics of the ellipse obtained by measurements using image processing routines agree with the exact values specified by the derivation of equation 4.5 within a small margin of error due to the discretization of the ellipse onto a regular grid.

Now it will be determined whether comparable information can be obtained via a similar analysis of geostatistical measures, rather than from direct analysis of the rainfall field itself. Since the correlogram is normalized, its values will not depend on the units or overall magnitude of the field. In addition, as shown above, the three geostatistical measures produce similar qualitative information. For these reasons, the correlogram (figure 4-1d) will be selected for more detailed analysis. In this case, various correlation contours will be analyzed using the same image processing routines that were previously used directly on the rainfall field. The results for correlation thresholds of 0.0, 0.2, 0.4, 0.6, and 0.8 are given in table 4-1. The estimates of $\sqrt{\varepsilon}$ provided by a/b, for all contours except 0.8, are nearly identical to the value measured by direct analysis of the rainfall field (=1.83), and are within a few percent of the exact value (=1.73). For the 0.8 contour the error is slightly larger (~15%), since this region is relatively small, the grid discretization has more of an impact on the result. All estimates of γ are within 10% of the exact value. These results confirm that *indirect* analysis of the rainfall field via summary geostatistical measures is able to characterize the orientation and ellipticity of the original field, to a reasonable degree of accuracy.

Next, the effect of multiple rainfall features upon the correlogram will be illustrated.

Table 4-1: Measurements of lengths of semi-major (a) and semi-minor (b) axes, their ratio (a/b), and the angle (γ) between the semi-major axis and the x-axis of the correlogram in figure 4-1d.

correlogram contour	a	b	a/b	$\gamma(°)$
0.0	49.2	26.8	1.84	29.2
0.2	37.6	20.6	1.83	28.6
0.4	29.7	16.1	1.84	32.6
0.6	22.5	12.0	1.87	32.3
0.8	15.3	7.6	2.00	31.6

Two identical features (figure 4-2, defined by A=10, R=25, γ=45°, ε=3, at (x_o, y_o) = (40,40) and (88,88)) produce three maxima in the correlogram. The central maximum represents the "within-blob" correlation contributed by both features. The other maxima contain contributions from the "between-blob" correlations. The separation vector corresponding to the location of these other correlation maxima is equal to the separation between the two features (h=(48,48) and (-48,-48)). Three evenly spaced identical features (same characteristics as in figure 4-2, except for an additional feature at (x_o, y_o) = (64,64)) that are organized along a line result in multiple maxima (figure 4-3). Again, the central maximum mainly represents contributions from within-blob correlation, and other maxima represent various between-blob contributions. Some correlation contour values (such as 0.2) extend in a continuous fashion along the axis of orientation for a considerable distance, indicating a relatively high degree of linear spatial organization of the individual features. When three identical blobs are placed in a disorganized fashion (figure 4-4, same characteristics as two feature case except with additional feature at (x_o, y_o) = (88,40)) multiple correlation maxima are produced. None of the correlation contours

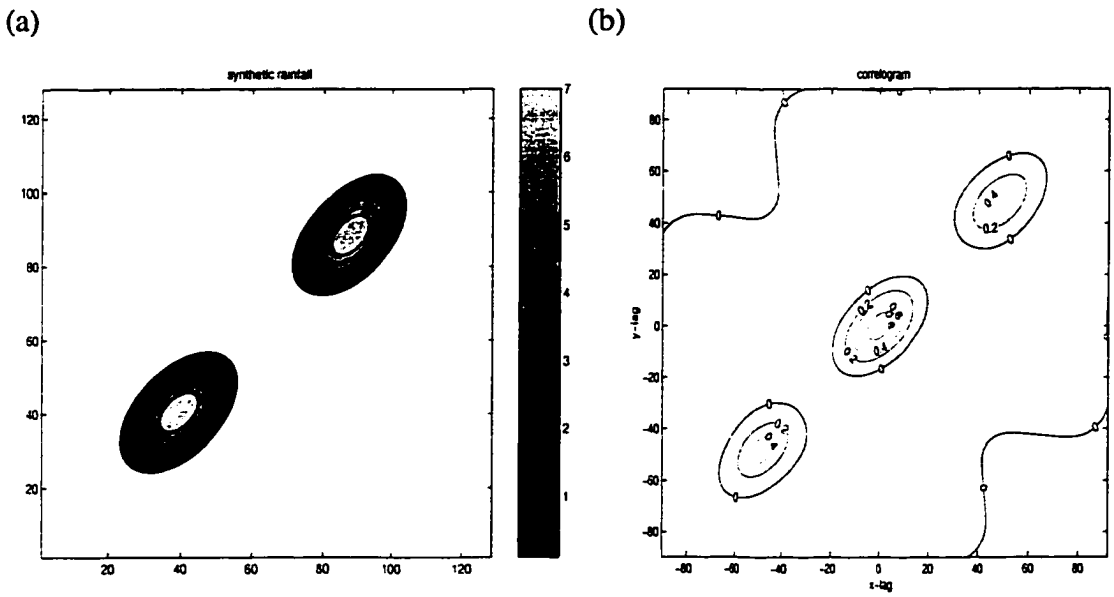


Figure 4-2: Synthetic rainfall field (a) of arbitrary units. Correlogram (b) corresponding to given synthetic rainfall field, lags indicated on axes are in terms of arbitrary unit grid spacing from the original field.

extend continuously over a large distance, indicating the complexity of the field. To be exact, there are seven correlation maxima produced from three distinct rainfall features. If the three rainfall features are labelled as A, B, and C, the corresponding correlation maxima will represent the within-blob correlation for all features, and contributions from the between-blob correlation from features $A \rightarrow B$, $A \rightarrow C$, $B \rightarrow C$, $B \rightarrow A$, $C \rightarrow A$, and $C \rightarrow B$. In general, given N distinct rainfall features, one should expect to find 2N+1 maxima on the corresponding correlogram. In cases where features are evenly spaced the number of maxima will be reduced. For example, if the separation vector between features A and B is identical to the vector between features B and C, the contributions of between-blob correlations from $A \rightarrow B$ and $B \rightarrow C$ will occur at the same place on the correlogram. When three identical features are aligned closely together, (figure 4-5, same characteristics as three feature cases, with (x_o, y_o) = (52,52), (64,64), and (76,76)) the rain-

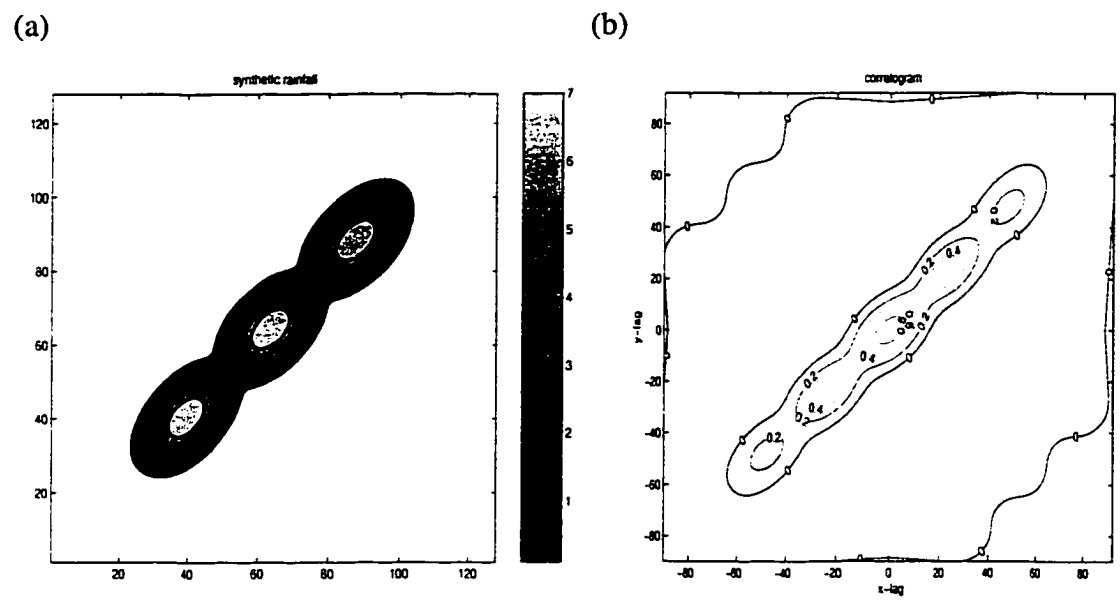


Figure 4-3: Synthetic rainfall field (a) of arbitrary units. Correlogram (b) corresponding to given synthetic rainfall field, lags indicated on axes are in terms of arbitrary unit grid spacing from the original field.

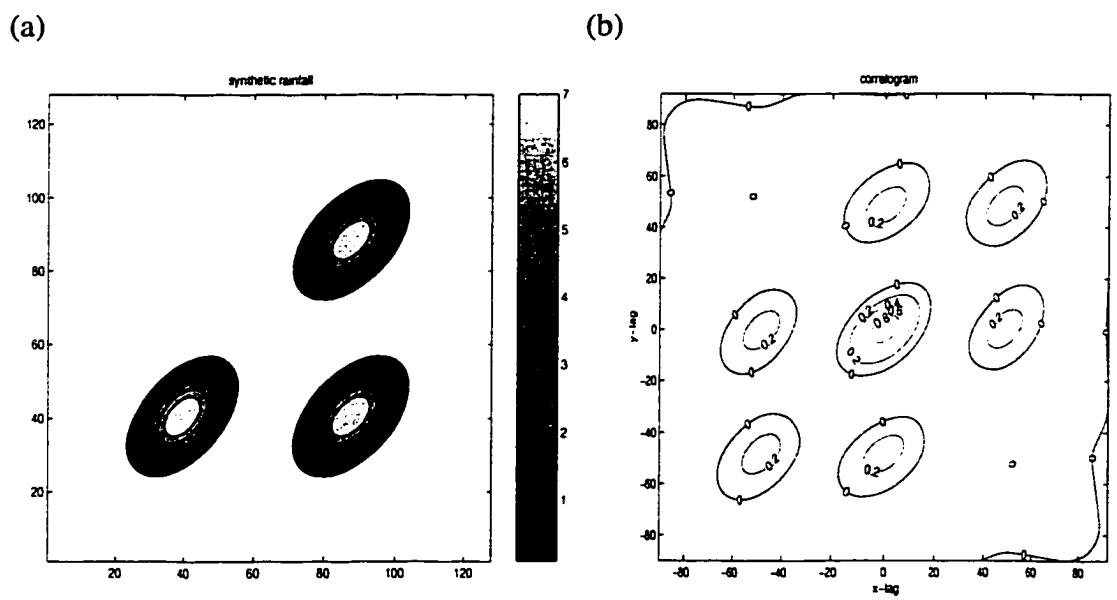


Figure 4-4: Synthetic rainfall field (a) of arbitrary units. Correlogram (b) corresponding to given synthetic rainfall field, lags indicated on axes are in terms of arbitrary unit grid spacing from the original field.

fall from each individual feature overlaps, resulting in a single continuous "line" feature.

The correlogram in this case is highly elliptical, over several values of correlation, indicat-

ing a large degree of continuity along the line.

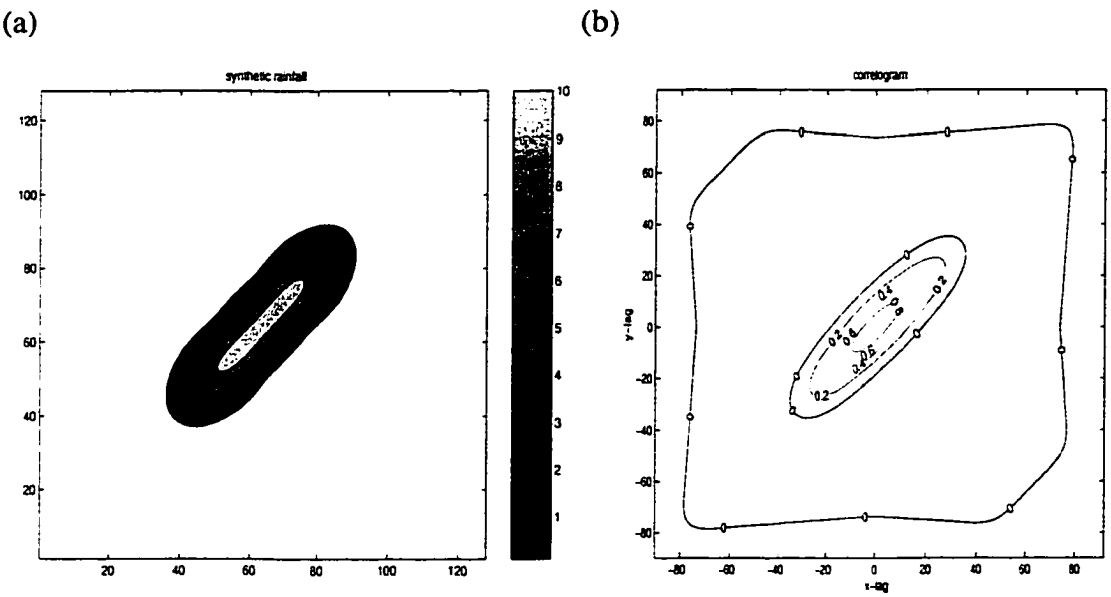


Figure 4-5: Synthetic rainfall field (a) of arbitrary units. Correlogram (b) corresponding to given synthetic rainfall field, lags indicated on axes are in terms of arbitrary unit grid spacing from the original field.

Overall, these results show the potential of using several summary measures from the correlogram to characterize the rainfall field. The number of maxima could be counted to approximate the number of discrete "features" in the original field. However, as previously discussed, this could be problematic if some features are evenly spaced along a line. The ellipticity and areal coverage of various correlation contours from the central maxima (connected to the origin) seem to represent the degree of alignment in the original field. These aspects of the correlogram will be estimated by fitting an ellipse to various correlation contours by measuring the lengths of approximate semi-major and semi-minor axes. The ellipticity of the correlation contour will be approximated by computing the ratio between the semi-major and semi-minor axes. The product between them will be used as an approximation to the area of an ellipse. These attributes will be tested

for their effectiveness in further refining the automated classification system in the next section.

4.4 Target data set results

For each event in the target data set (see section 3.2 for a complete description of the target data set), a correlogram was computed using GSLIB, a freely available library of software packages for geostatistics developed at Stanford University (Deutsch and Journel 1988). Plots of the target data set correlograms are provided in figures 4-6 through 4-9. The largest separation vector in either direction that was included in these analyses was 64 grid boxes in length, therefore any correlation beyond 64 Δx (~250km) was ignored. Inspection of these plots reveals that correlation contours for cases that were subjectively classified as linear (cases 1-18) appear quite elliptical, while contours for cases subjectively classified as cellular (19-37) often appear more circular. Therefore, one might expect that summary measures of the ellipticity of various correlation contours will help the automated classification system to better discriminate between these two classes of convective precipitation.

To determine whether or not this is the case, the area and ellipticity of various contour values (0.2, 0.4, 0.6, and 0.8) in each correlogram from the target data set were analyzed. Using the same image processing algorithms that were described in the previous section, the lengths of the approximate semi-major (a) and semi-minor (b) axes of an ellipse fit to each correlation contour were determined. Briefly summarizing this process, the correlation contour is considered an object, and all contiguous grid points with correlation greater than the contour value that also include the origin are given the same object label using connected component labeling algorithm outlined in section 2.12. Next, the

edge of this connected region is found using the edge detection algorithm also outlined in section 2.12. These processes equate to locating the specified contour for the central maximum on the correlogram. Contours related to secondary maxima are not analyzed by this procedure. Once this is established, the largest distance from the origin to this edge is found, and this distance is assumed to be the length of the semi-major axis (a) of the ellipse. If a contour is not closed, the edge of the correlogram domain becomes the edge of the connected region. In this case, a becomes the largest distance where the correlation is greater than the contour threshold value, given that the point is located within the region connected to the origin. The shortest distance from the origin to the edge is found next, and this is assumed to be the length of the semi-minor axis (b). As discussed previously, the ratio of the semi-major and semi-minor axes (a/b) is a summary measure of the ellipticity or eccentricity of the ellipse. For a circle, this ratio will be equal to 1.0. The ratio will increase as the ellipticity of the contour increases. The product of the two axis lengths (ab) will also be used as a summary measure approximating the area of the ellipse ($= \pi ab$).

The results of this analysis are provided in tables 4-2 and 4-3. To determine the usefulness of these summary measures as attributes in an automated classification system, the attributes corresponding to the 0.6 correlation contour were selected for initial analysis. This contour was chosen because it was closed on all of the correlograms provided from the target data set (figures 4-6 through 4-9). As in the analysis described in chapter 3, the automated classification system uses a hierarchical cluster analysis algorithm, specifically Ward's method. Validation of the automated classification will be performed in the same way as described in section 3.4. The use of hierarchical cluster analysis as a classification

tool requires a subjective decision in order to determine the number of clusters. In this work, this subjective decision will be made as objectively as possible. The dendrogram will be cut into four or five clusters containing the majority of objects in the target data set. Outliers will not be classified, and no more than six outlier cases will be allowed. Every attempt will be made to cut the dendrogram at a point where there is substantial separation between the intra-cluster and inter-cluster variance. Each major cluster will be considered a separate class of objects. The definition of each class will be determined by the highest percentage of subjectively classified cases detected for that particular cluster. The percentage of objects that are correctly classified by their membership in the dominant subjective class for each cluster will be used as the metric for determining the skill of the automated classification.

The purpose of this research is to determine which attributes are most useful in an automated rainfall pattern classification system. In order to build upon the system that was successfully able to separate convective and non-convective events (section 3.5), estimates of *ab* and *a/b* will be added to the histogram-related attributes to determine if geostatistics-related attributes further refine the classification. It was determined in the last chapter that the classification using gamma distribution parameters (α,β) from the generalized method of moments estimation using three moments produced the best results overall. In addition, since those results were not sensitive to q, an arbitrary choice of q can be made. Therefore, the histogram-related attributes that will be used for the remainder of this chapter will be from the GMM using three moments and q=1 (table 3-3).

Now that the set of attributes and the classification algorithm have been selected, the question now becomes whether the attributes require normalization and what combination

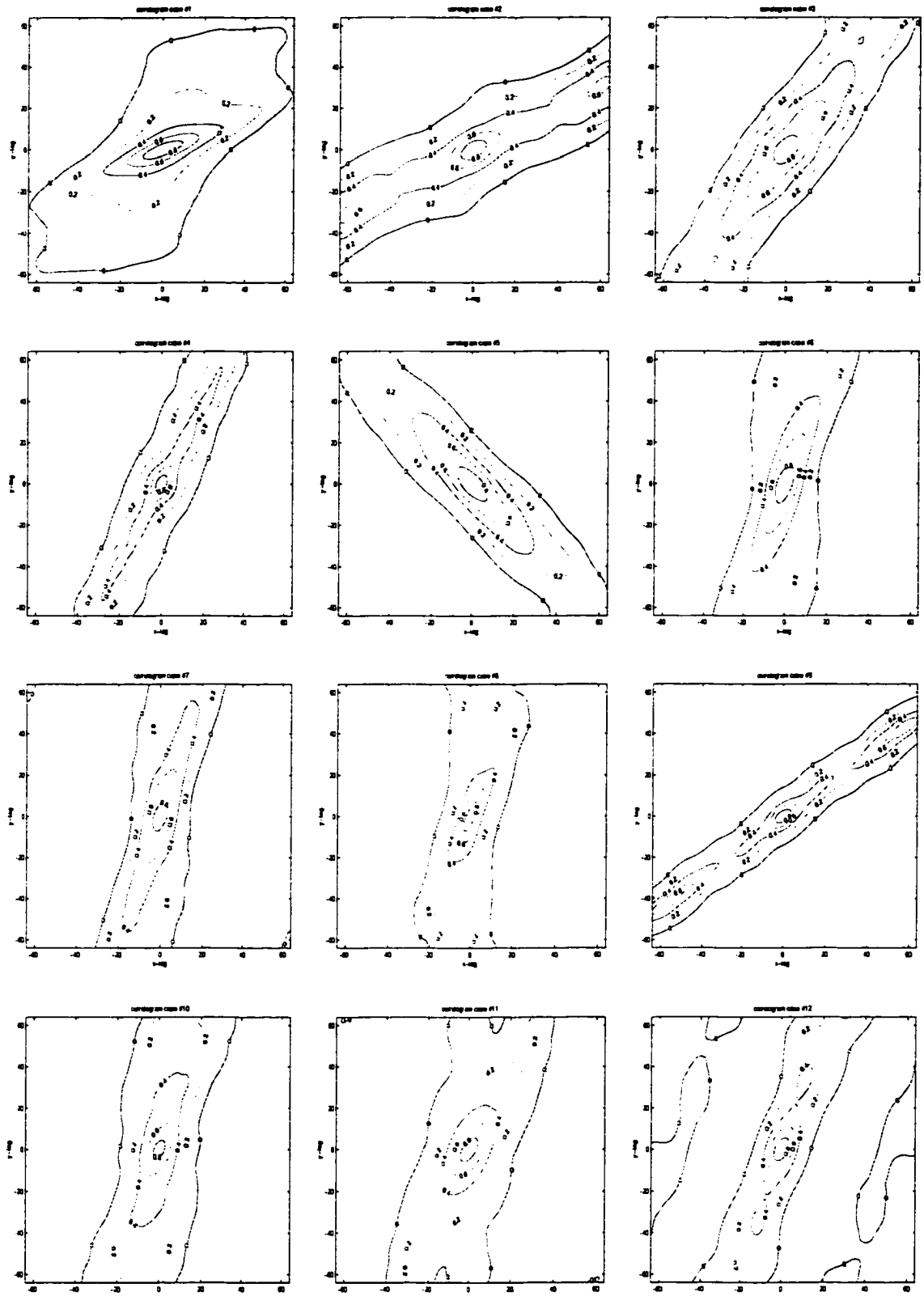


Figure 4-6: Correlogram plots corresponding to rainfall cases 1 through 12 of the target data set. Lags indicated on axes are in terms of 4km grid boxes from the original analysis.

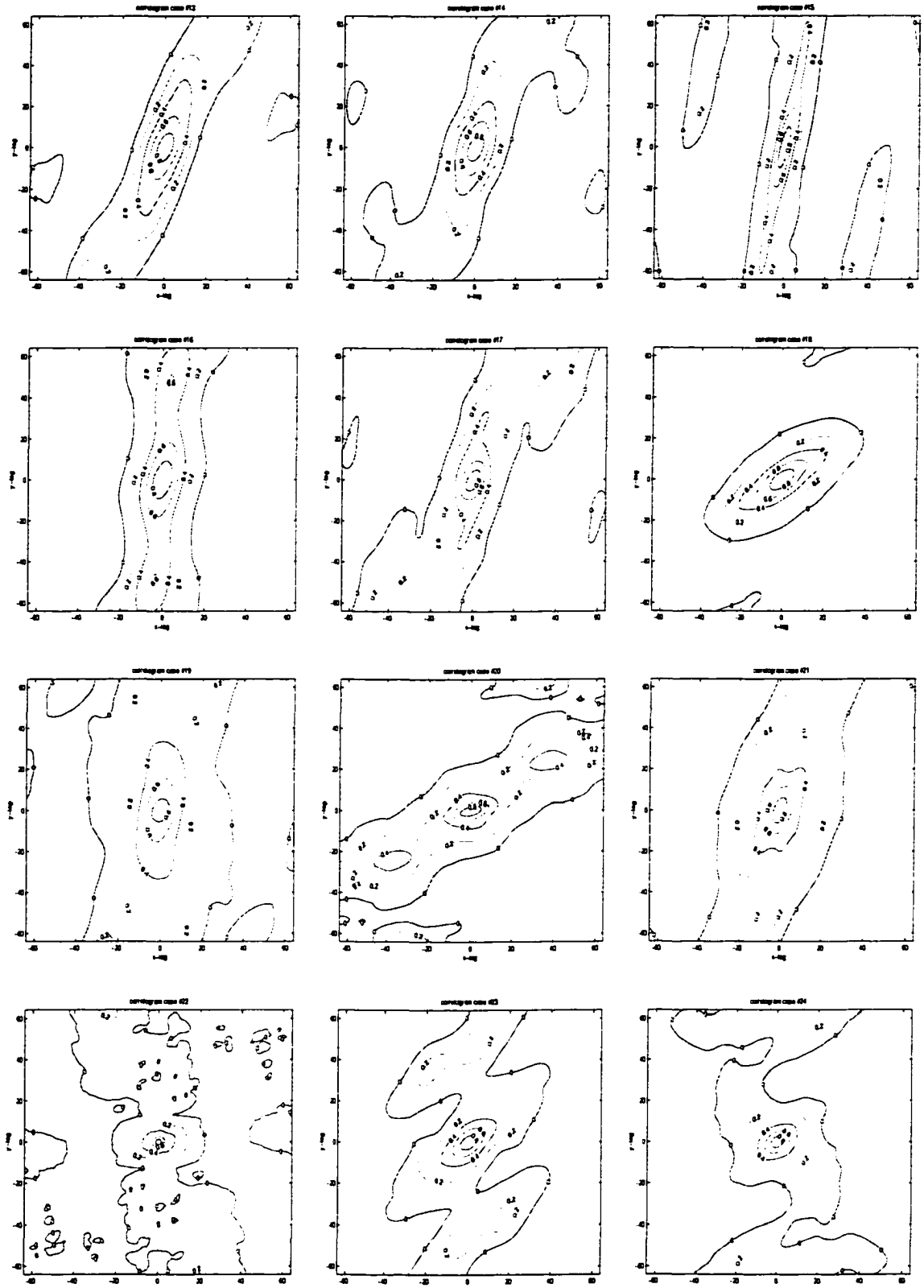


Figure 4-7: As in figure 4-6, except for rainfall cases 13 through 24 of the target data set.

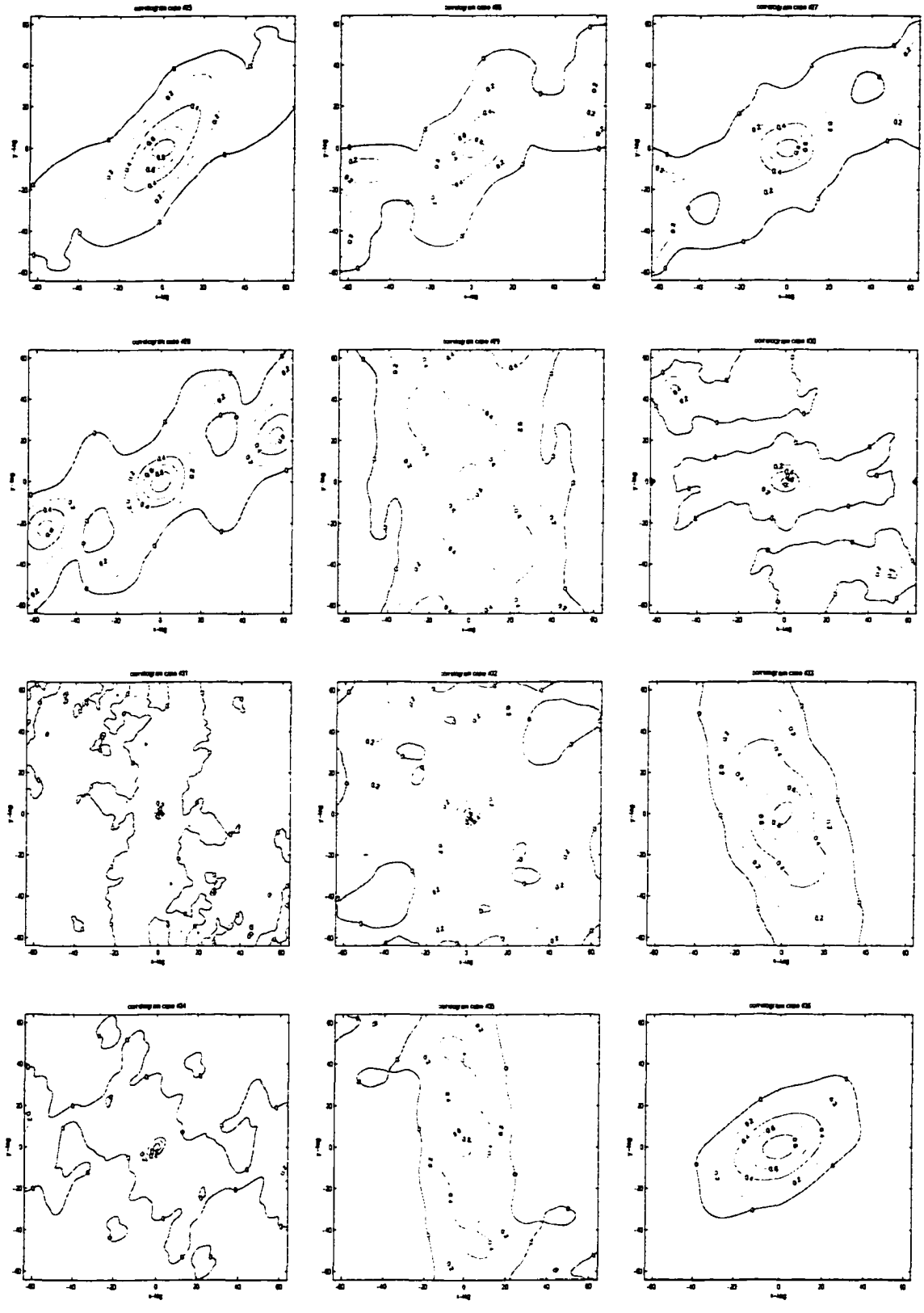


Figure 4-8: As in figure 4-6, except for rainfall cases 25 through 36 of the target data set.

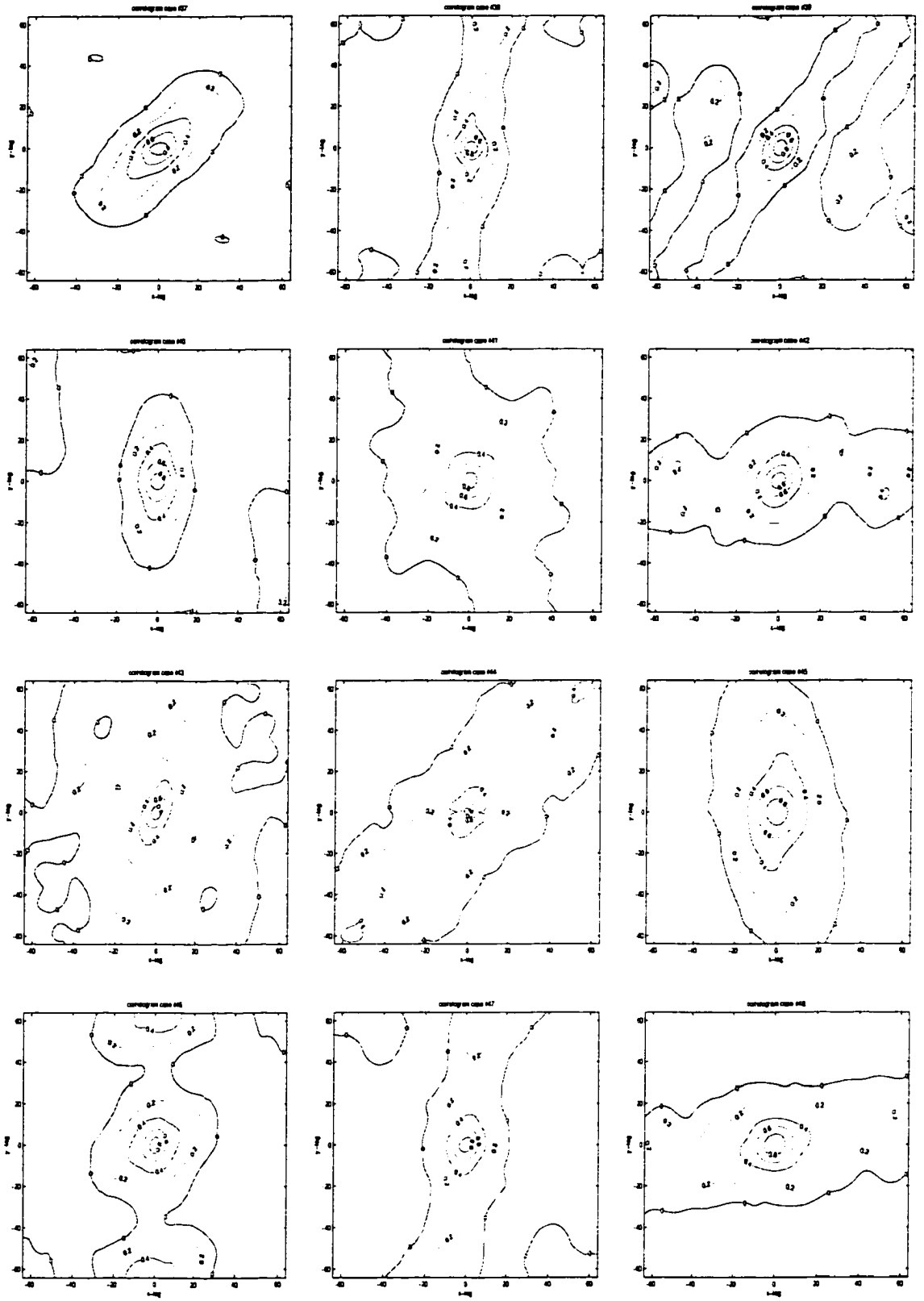


Figure 4-9: As in figure 4-6, except for rainfall cases 37 through 48 of the target data set.

Table 4-2: Characteristics of ellipses fit to the 0.2 and 0.4 correlation contours. The ratio (*a/b*) and product (*ab*) of the semi-major and semi-minor axes along with the angle between the semi-major axis and the x-axis (θ) in degrees are provided.

	0.2 contour			0.4 contour		
case	*a/b*	*ab*	θ (deg)	*a/b*	*ab*	θ (deg)
1	4.09	661.9	22.6	4.25	212.7	21.4
2	5.86	1054.5	36.7	8.86	646.9	33.7
3	5.66	1274.4	47.9	4.56	583.9	54.5
4	7.32	709.8	60.9	9.90	396.0	63.4
5	5.35	776.0	138.1	5.12	368.4	128.5
6	6.40	742.5	66.0	5.86	339.6	70.3
7	6.78	677.8	68.4	8.28	405.8	75.0
8	8.57	497.0	74.9	5.70	114.0	64.4
9	12.27	552.3	40.1	7.03	126.6	39.6
10	5.72	744.0	74.9	4.78	306.2	70.1
11	5.55	888.1	63.8	2.55	204.4	66.8
12	6.97	557.2	71.3	3.44	154.9	72.3
13	4.59	477.7	73.9	4.70	249.0	74.7
14	3.99	462.6	77.9	3.03	175.8	72.3
15	11.15	379.0	75.7	15.50	247.9	79.8
16	5.44	789.0	74.1	7.49	546.4	80.1
17	4.67	541.6	72.6	4.63	245.6	78.0
18	3.13	353.8	32.7	3.57	178.5	33.7
19	4.58	897.9	100.8	3.66	300.1	84.8
20	5.55	988.5	31.8	1.87	99.0	17.1
21	3.11	1370.3	74.9	2.52	206.7	61.2
22	1.70	115.4	180.0	1.70	34.1	23.2
23	2.96	251.6	28.4	2.47	101.2	34.7
24	1.42	204.4	49.8	1.61	72.6	33.7
25	3.08	554.6	57.8	2.77	235.1	48.2
26	2.61	443.5	65.7	2.09	186.1	59.5
27	1.94	328.1	56.3	1.36	120.8	51.3
28	1.57	254.6	53.1	1.48	125.4	54.0
29	3.13	2003.7	127.3	3.73	1207.2	110.1
30	1.74	85.2	9.5	1.52	24.3	170.5
31	1.90	15.2	111.8	2.24	2.2	116.6
32	2.27	240.1	59.0	2.69	48.4	52.1
33	2.75	1059.7	100.7	2.65	519.5	104.0
34	1.47	49.8	110.6	2.53	12.6	45.0
35	4.12	964.2	91.8	5.01	580.9	100.7
36	2.05	544.1	38.9	1.78	258.5	27.8
37	3.23	469.0	48.1	2.54	155.0	40.9
38	2.54	276.4	79.1	2.22	115.6	86.4
39	2.04	181.2	51.3	1.67	75.0	63.4
40	2.24	332.1	81.6	2.03	164.2	80.5
41	1.96	575.1	53.5	1.42	171.8	50.2
42	1.52	293.1	58.6	1.49	119.3	77.0
43	4.73	345.6	81.5	2.77	94.2	68.2
44	4.86	1215.0	55.1	2.68	67.1	63.4
45	2.63	951.3	90.0	2.17	368.8	81.9
46	1.64	433.5	34.3	1.41	137.2	21.0
47	2.10	373.8	88.0	1.72	123.5	74.1
48	3.71	1042.7	175.4	1.46	200.0	20.6

Table 4-3: Characteristics of ellipses fit to the 0.6 and 0.8 correlation contours. The ratio (*a/b*) and product (*ab*) of the semi-major and semi-minor axes along with the angle between the semi-major axis and the x-axis (θ) in degrees are provided.

	0.6 contour			0.8 contour		
case	*a/b*	*ab*	θ (deg)	*a/b*	*ab*	θ (deg)
1	5.00	65.0	19.4	4.74	19.0	18.4
2	2.58	74.8	30.3	2.37	19.0	26.6
3	4.04	202.2	53.5	2.00	26.0	56.3
4	2.48	32.2	63.4	3.16	6.3	63.4
5	4.14	165.5	133.5	2.77	36.1	126.9
6	5.06	126.5	71.6	3.35	26.8	71.6
7	5.16	103.2	72.3	3.64	14.6	74.1
8	5.06	40.5	65.2	7.28	7.3	74.1
9	2.76	22.1	39.8	3.00	6.0	45.0
10	2.83	56.6	71.6	3.16	6.3	63.4
11	2.49	72.2	63.4	2.61	13.0	59.0
12	2.77	47.0	74.7	2.92	11.7	59.0
13	3.10	77.6	75.1	3.26	16.3	74.1
14	2.62	68.0	77.0	2.83	14.1	71.6
15	8.68	43.4	78.1	10.20	10.2	78.7
16	3.04	109.5	80.5	2.92	29.2	77.5
17	2.30	39.1	71.6	2.69	10.8	68.2
18	3.40	61.2	33.7	2.86	14.3	51.3
19	2.62	89.2	78.7	1.90	15.2	68.2
20	3.07	27.7	12.5	5.10	5.1	11.3
21	2.70	54.0	65.6	3.16	12.6	71.6
22	2.12	4.2	180.0	1.00	1.0	180.0
23	2.09	35.5	35.5	2.24	8.9	26.6
24	1.86	24.2	26.6	2.24	4.5	18.4
25	2.16	69.1	55.0	2.06	16.5	31.0
26	1.50	60.0	18.4	1.70	17.0	21.8
27	1.30	52.2	14.0	1.41	14.1	153.4
28	1.49	59.7	58.0	1.80	14.4	78.7
29	1.84	222.5	81.5	1.65	41.2	104.0
30	1.58	6.3	161.6	1.41	1.4	135.0
31	1.00	1.0	45.0	1.00	1.0	0.0
32	2.69	10.8	68.2	2.83	2.8	45.0
33	1.62	110.0	103.0	2.06	16.5	59.0
34	3.61	3.6	56.3	1.00	1.0	45.0
35	1.56	62.3	66.0	1.58	12.6	63.4
36	1.72	112.0	30.3	1.77	30.0	15.9
37	1.98	49.5	45.0	2.24	8.9	26.6
38	1.26	25.3	135.0	1.58	6.3	161.6
39	1.63	32.6	74.1	1.61	8.1	56.3
40	1.63	55.3	71.6	1.84	9.2	76.0
41	1.33	54.3	45.0	1.50	12.0	45.0
42	1.52	38.1	66.8	1.61	8.1	56.3
43	2.11	27.5	66.8	2.24	4.5	71.6
44	2.24	8.9	63.4	1.41	1.4	135.0
45	1.42	103.9	80.5	1.48	25.1	99.5
46	1.70	34.1	66.8	2.92	5.8	76.0
47	1.52	38.1	66.8	1.80	7.2	56.3
48	1.30	65.2	49.4	1.41	18.4	78.7

of these attributes will produce the best classification results. Some sort of normalization will likely be necessary, since the range of values of *ab* is two to three orders of magnitude higher than the other attributes (see table 4-3). The question of normalizing the attributes prior to the cluster analysis will be investigated by using the raw attributes, normalizing each attribute vector to produce zero mean and unit variance, and normalizing each attribute by dividing by its maximum value. For each of these types of normalization, different combinations of subsets of the four attributes (α, β, *ab*, *a/b*) were used. This includes the six possible combinations of two of the four attributes, plus the four possible combinations of three of the four attributes, and all four attributes, resulting in 11 different experiments for each type of normalization. As in the previous chapter, the percent correct was considered for the three-class (linear, cellular, stratiform) as well the two-class cases (combining the linear and cellular classes into a parent "convective" class).

To illustrate how the validation of the automated classification operates, one example will be analyzed in detail. Figure 4-10 shows the dendrogram resulting from Ward's cluster analysis method using unnormalized *a/b*, α, and β as attributes. In this example, the tree was cut (at a level of square root of the increase in the sum of the squared error of approximately 4) to produce five clusters with two outlier cases. Cluster 1 contains all 11 cases that were subjectively classified as stratiform precipitation events. Cluster 2 is split among the linear and cellular precipitation events with seven linear cases and 11 cellular events. Cluster 3 contains only five linear events. Cluster 4 contains three linear events and two cellular events. Finally, cluster 5 contains two linear events and five cellular cases. Overall, this experiment placed 35 of the 46 cases into correct dominant classes, for a 76.1% percent correct for the three-class case. This result is improved over the experi-

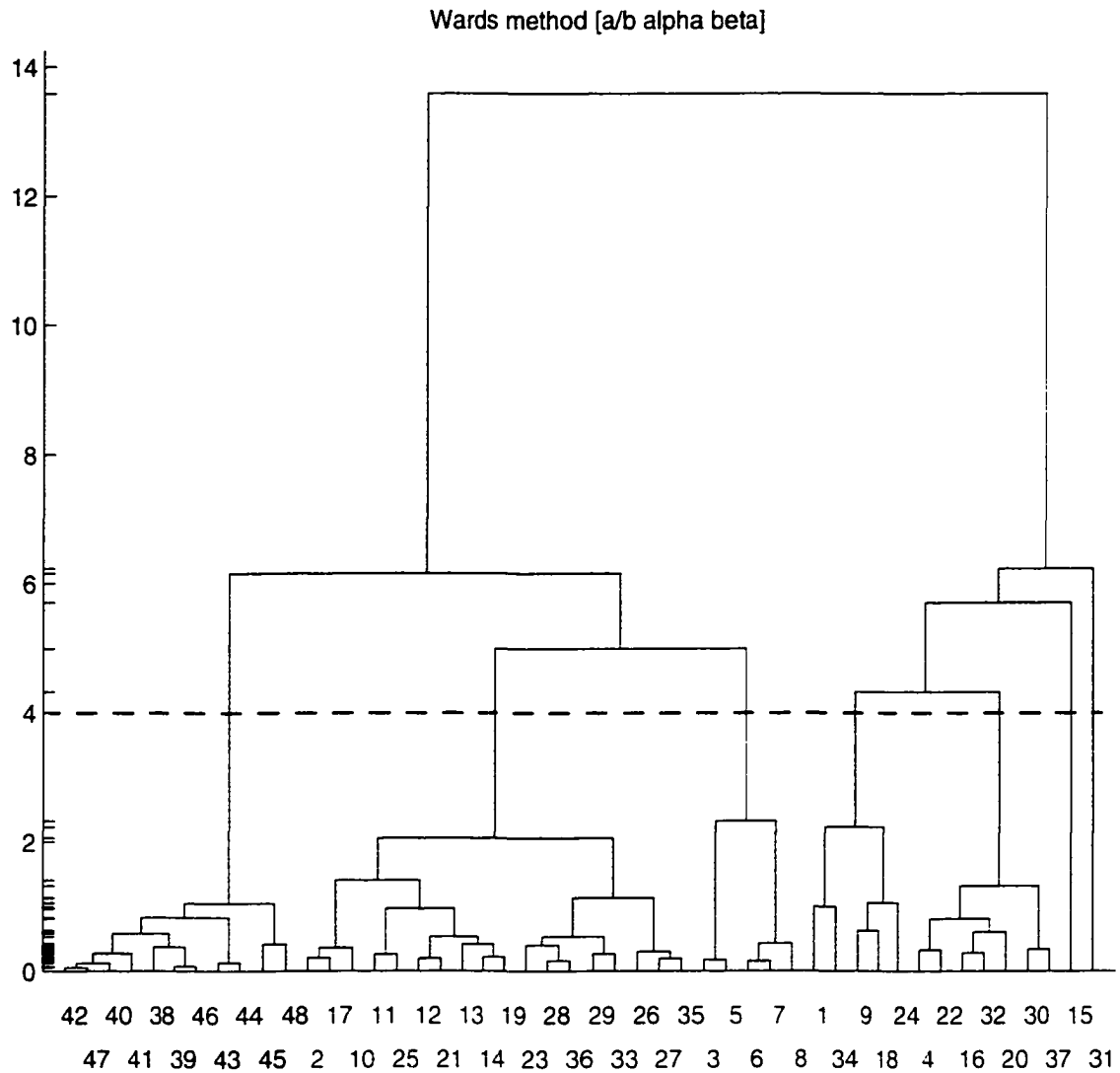


Figure 4-10:Dendrogram produced by Ward's method with target data set using *a/b* for 0.6 contour and 3-moment q=1 GMM [α, β] as attributes. Each object is color-coded by its subjective classification, linear events are green, cellular events are red, stratiform events are blue. Dashed line indicates subjectively determined cut-level for this experiment.

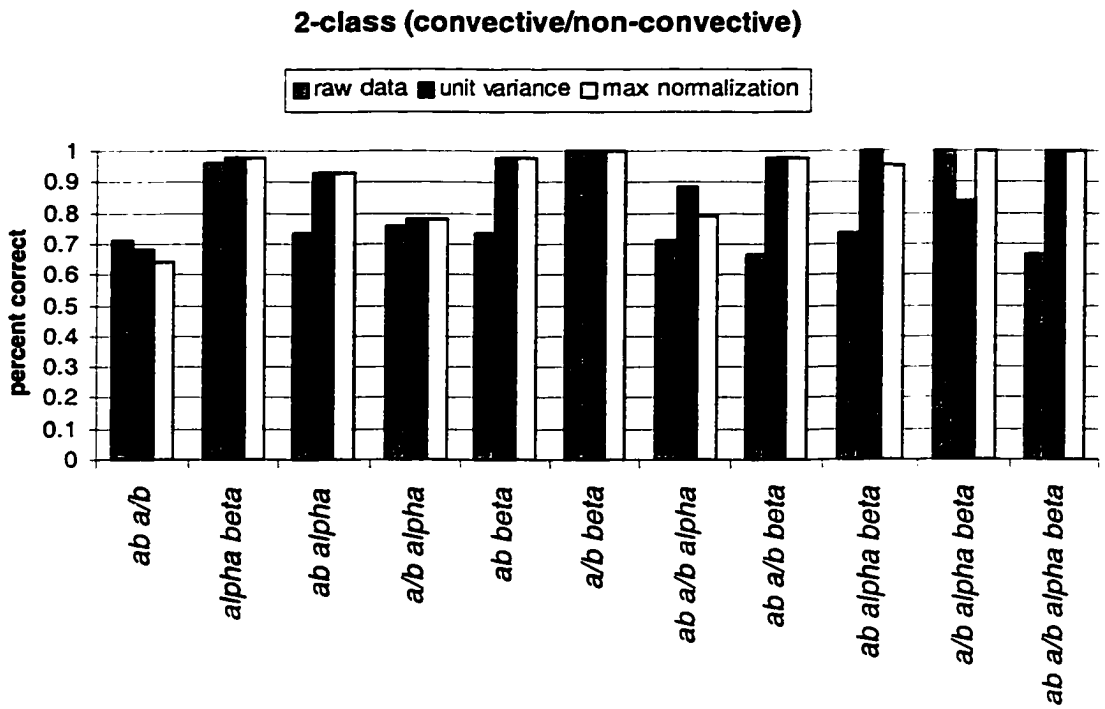
ment using only α and β as attributes, which scored 66% correct in this case, showing that the addition of information on the ellipticity of the correlogram helps to refine the classification system. On the other hand, this result is only slightly better than the baseline experiment (raw rainfall values) which correctly classified 72.3% of the cases into the dominant

class for each cluster. Since each cluster is unanimously populated with either convective (clusters 2-4) or non-convective (cluster 1) events, the two-class case results in 100% correct classification. Obviously, a perfect score is better than the results of the baseline experiment (83.7%) and also shows improvement over the results using only α and β as attributes (95.7%).

This validation was repeated for the entire set of 11 experiments for the raw (unnormalized) attributes as well as attributes normalized to produce zero mean and unit variance, and attributes normalized by their maximum value. Figure 4-11 shows the percent correct results for all of the different combinations of attributes and normalization. It is clear from comparing the raw data results with those obtained after the attributes had been normalized that some sort of normalization is necessary. The results of any experiment that included unnormalized *ab* were much worse than those that did not include *ab*. This did not come as a surprise, since the range of *ab* values is several orders of magnitude higher than that of the other variables. There does not appear to be a clear preference for the type of normalization, as the results from both types of normalization were nearly identical for almost every combination of the various attributes.

Concentrating on the results from the normalized attribute experiments in the 2-class case (figure 4-11a), the best combination of two of the four attributes were *a/b* and β, producing 100% correct, for both types of normalization. The best combination of three of the four attributes for the unit variance normalization was α, β, and *ab*, which also produced 100% correct. However, the best combination of three attributes for the maximum normalization was α, β, and *a/b*, which also produced 100% correct. In addition, the experiment using all four attributes produced 100% correct. Most of the experiments pro-

(a)



(b)

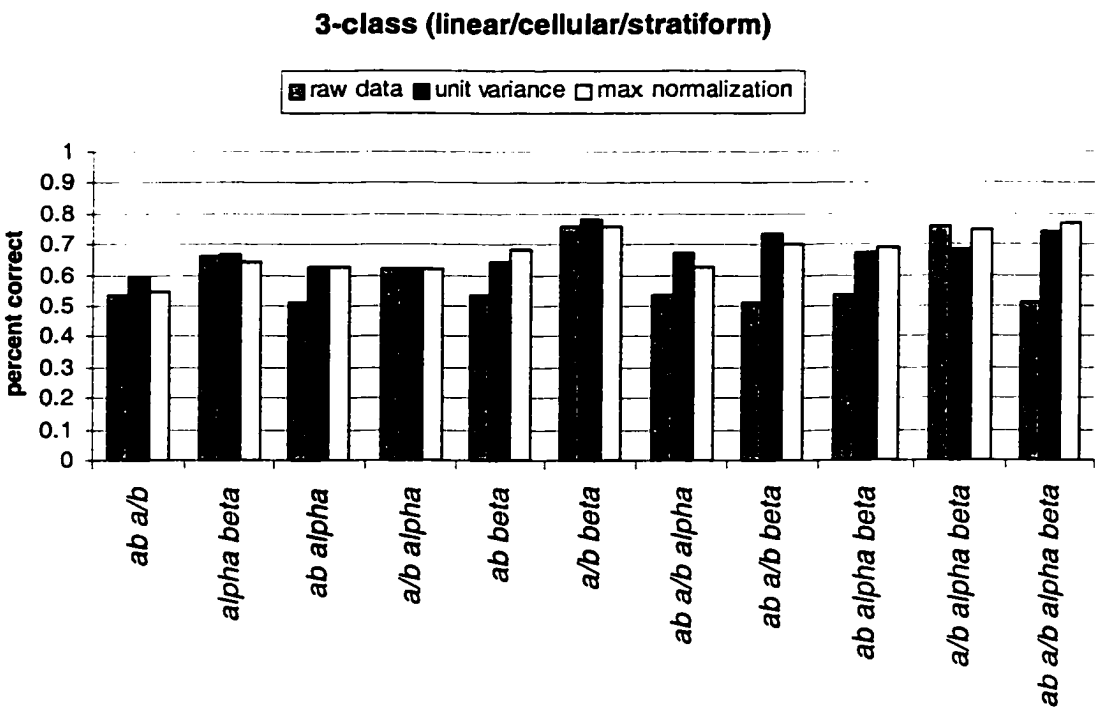


Figure 4-11: Percent correct results for 11 experiments in the two-class (a) and three-class (b) cases. Results using unnormalized attributes are in blue, attributes normalized to zero mean and unit variance are red, attributes normalized by their maximum are in white. The combination of attributes used in each experiment is indicated below each bar on the x-axis.

duced greater than 90% correct, which is consistent with results found in the previous chapter (section 3.5).

Figure 4-11b shows the percent correct results for all of the different combinations of normalized and raw attributes in the three-class case. The best combination of two of the four attributes, and the best combination overall was *a/b* and β, producing 78.2% correct in the unit variance normalization case, and this experiment was clearly superior to the other combinations of two attributes. For unit variance normalization, the best combination of three of the four attributes was *ab*, *a/b*, and β, which produced 73.3% correct. A similar level of performance was obtained in the maximum normalization case using α, β, and *a/b*, producing 75% correct. The experiment using all four attributes produced 74.4% correct using unit variance normalized attributes, and 76.7% correct when attributes were normalized by their maximum value. These results are similar to the best three of four and slightly less than the best two of four attribute experiments. These results are improved slightly when compared to the baseline experiment (raw rainfall values) which correctly classified 72.3% of cases.

In the three-class (linear, cellular, stratiform) case, the best results were obtained when only two out of four attributes were used (unit variance normalized *a/b* and β). This appears to be counter-intuitive, one might expect that additional information will always improve the automated classification. In fact, additional information will only improve the classification if it is consistent with the aspects of the data which were considered important in characterizing the various classes of objects. This sort of consistent information is defined by Romesburg (1984) as *essential*. This can be illustrated with a hypothetical example. In figure 4-12a, the data matrix contains 20 objects and 2 attributes,

therefore it can be easily visualized. For sake of argument, the subjective and automated classification schemes both agree that there are two clusters or classes of objects. In this case, the automated classification is perfect, therefore the attributes #1 and #2 used are considered essential. Next, in figure 4-12b, an additional attribute (#3) is added that does not include information that is consistent with the characteristics of the data set that were deemed important to the subjective analyst. Objects that belonged to the same class in the two-attribute case are now scattered and mixed due to the influence of the third, inessential attribute. The automated classification now disagrees with the subjective classification because many of the objects are now farther apart in attribute #3 space than they were in the space of attributes 1 and 2, and vice versa.

(a) (b)

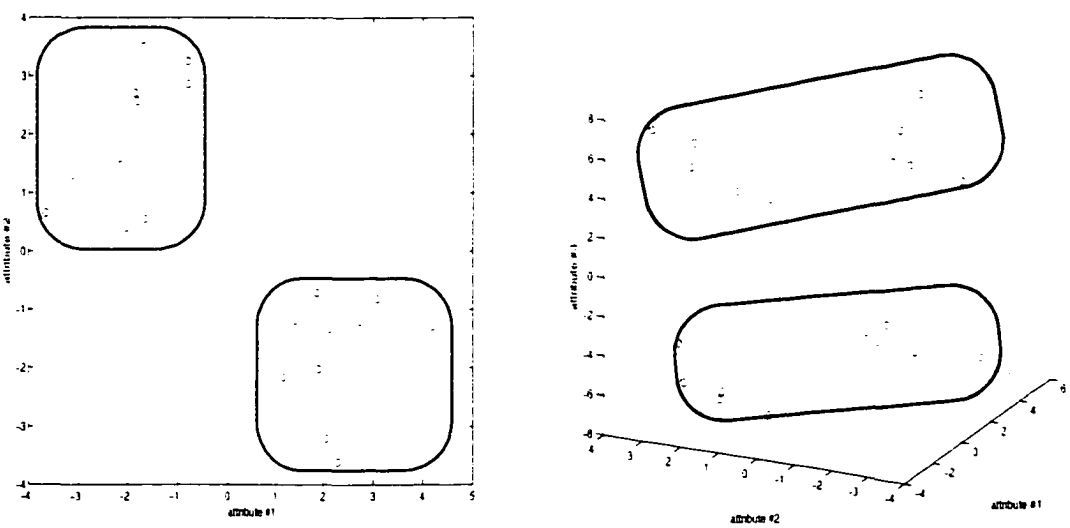


Figure 4-12: Hypothetical example illustrating the effect of the addition of an unimportant attribute to an automated classification. Left hand panel (a) shows the data matrix plotted in terms of attributes #1 and #2, similar colored objects are clustered together by the hypothetical automated classification system (also indicated by like-colored contours). Right hand panel (b) shows the same data matrix plotted in three dimensions, including attribute #3. Objects maintain the same colors as they have in two-attribute space plot (a). Black contours indicate results of hypothetical automated classification.

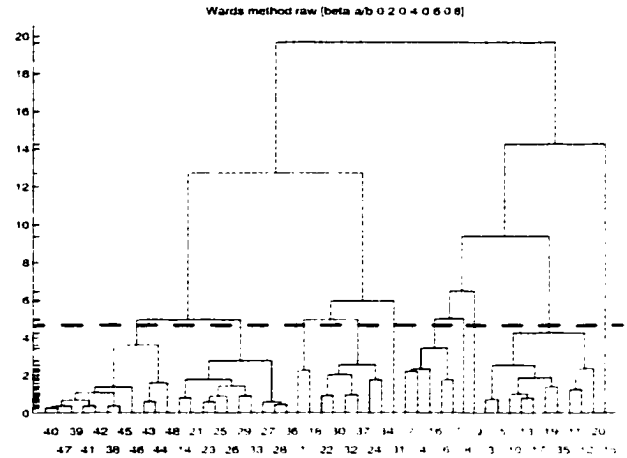
To this point, excluding the information on the angle between the semi-major axis and x-axis, ten attributes have been collected for each object in the data matrix: α, β, *a/b* for 0.2, 0.4, 0.6, 0.8 contours, and *ab* for the four contours. Information on the orientation of the rainfall pattern is excluded from this set since it is not expected to help in discriminating between classes, but instead should further describe the pattern once the general class has been determined. In fact, in order to determine the ideal selection of a combination of these ten attributes, thousands of experiments would be necessary. The exact number can be determined by turning to probability theory. The number of possible combinations of a subset of r objects selected from a set of n objects is defined as $\binom{n}{r} = \frac{n!}{(n-r)!r!}$. In this case, n=10 and r would vary from 1 to 10 since we would want to test every combination from of the set of 10 attributes. Therefore, the total number of experiments required to exhaustively determine the ideal set of attributes would be $= \sum_{i=1}^{10} \binom{10}{i} = 1023$. In addition, since three different types of normalization have been tested, this number would need to be multiplied by 3. Obviously, analysis of such a large number of experiments is beyond the scope of this work. However, results from the experiments already performed can be used as guidance in the selection of a proper set of attributes from the set of ten.

It appears that the addition of information on the elliptical nature of the 0.6 contour in the correlogram provides useful information that allows for further refinement of the automated classification system. The previous results show that *a/b* and β are the attributes with the most discriminating power. For example, the best results were obtained

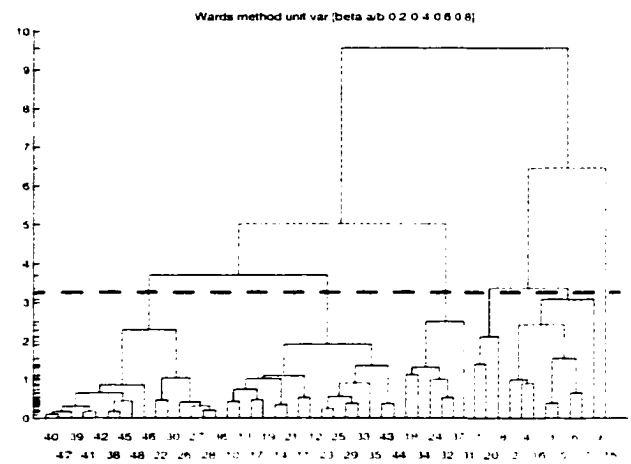
when these were used by themselves. When additional attributes were added, results degraded slightly. When only one of these were used in combination with other attributes, results were also degraded. Therefore, in order to further refine the classification, it seems reasonable to expand the number of attributes by including information on the ellipticity of other correlation contours in the correlogram.

Figure 4-13 shows three cluster analysis results using β from three-moment, q=1 GMM and a/b for 0.2, 0.4, 0.6, and 0.8 contours on the correlogram: raw attributes, attributes normalized to have zero mean and unit variance, and attributes standardized by dividing by their maximum values. Examining the results from the unnormalized attributes first (figure 4-13a), a cut has been made on the dendrogram separating the objects into five main clusters with six outliers (objects 1, 8, 9, 15 18, and 31). The first cluster is unanimously populated with all eleven stratiform events. The second cluster is dominated by nine cellular cases, with one linear case included. The third cluster is also cellular dominant, with six cases. The fourth cluster contains five linear cases, and the fifth cluster is split between seven linear cases and three cellular events. Validating these clusters, in the two-class case there are no mis-classified events, resulting in 100% correct classification. In the three-class case the clusters correctly placed 38 of 42 cases into the dominant class, for a 90.5% percent correct. Clearly, these validation results outperform the baseline experiment values as well as all other experimental results discussed previously. The automated classification using these five attributes have successfully separated the cellular, linear, and stratiform events with over a 90% accuracy rate.

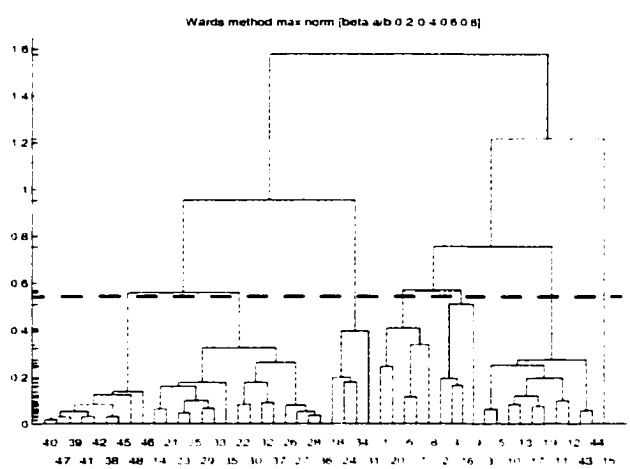
The results from using attributes normalized to have zero mean and unit variance (figure 4-13b) do not reach the same level of success, however. A cut is made on the den-



(a)



(b)



(c)

Figure 4-13: Dendrograms produced by Ward's method with target data set using [β, *a/b* 0.2 0.4 0.6 0.8 contours] attributes, unnormalized (a), normalized so each attribute has unit variance (b), and each attribute normalized by its maximum (c). Each object is color-coded by its subjective classification, linear events are green, cellular events are red, stratiform events are blue. Dashed line indicates subjectively determined cut-level for each experiment.

drogram in a similar fashion to create five clusters, this time with only one outlier (case #15). Notice in this case that the first cluster contains two different groups, one that includes nine stratiform events and the other which has six cellular cases. There was no way to cut the dendrogram such that these would be divided into two distinct groups while also producing four to five main clusters with six outliers or less. If one were to cut the dendrogram to produce six clusters, these two groups could be divided. Given this mixed stratiform/cellular cluster, the percent correct for the two-class case has dropped to 83.0%, and is 66.0% in the three-class case. Although these validation results are not as favorable as those obtained from unnormalized attributes, the automated classification is grouping events in a similar fashion to the subjective classification.

On the other hand, results from using the five attributes after normalization by their maximum values are consistent with the favorable results obtained without normalization (figure 4-13c). Again, a cut is made to produce five main clusters with five outliers (cases 2, 4, 9, 15, and 16). In this case, it is possible to separate the main stratiform cluster from the other clusters. Here, cluster 1 contains nine stratiform cases. Cluster 2 contains 14 cellular events and one linear case. Cluster 3 contains three cellular and one linear case. Cluster 4 contains four linear and one cellular case, and the last cluster contains seven linear, one cellular, and two stratiform events. Overall, this classification is able to correctly place 41 of 43 cases into their proper convective/non-convective classes, for a 95.4% percent correct. In addition, 37 of 43 cases are placed into their proper linear/cellular/stratiform classes, for a 86.1% correct validation.

It is interesting to look more closely at some cases that were placed in a class other than the one that was subjectively assigned in order to determine why they might be incor-

rectly classified. For example, in each classification experiment shown in figure 4-13, linear case #14 is incorrectly placed with a group of cellular events. Figure 4-14a shows the rainfall for this case, and figure 4-14b shows the rainfall plot for case #21, which is one of the similar cellular cases that case #14 is grouped with in these cluster analysis results. In fact, these two cases do appear to be quite similar. They are both somewhat organized, case #14 has a prominent circular-shaped feature and several other intense cells are organized more or less along a line. Case #21 is organized to a lesser degree along a line, which explains why it was subjectively classified as a cellular event. Another example is case #19, which was subjectively classified as cellular but is placed with a group of linear events in the automated classification experiments. Figure 4-14c shows the rainfall from case #19 while figure 4-14d shows the rainfall from case #13, which is one of the linear cases that case #19 is closely connected to on the dendrograms found in figure 4-13. Again, visually these cases appear to be quite similar. They are both organized to a certain degree along a line. Case #19 has a small intense cell on the western edge of the domain, which indicates the more cellular nature of this case. Case #13 also has a small cell on the easter edge of its domain, but this cell is closer to the organizational line than the offset western cell in case #19. In both of these examples, the "mis-classified" events have some characteristics of the both linear and cellular classes. For example, these cases appear to be more similar to each other than they do other "classical" linear (such as case #6, figure 3-1) or cellular (such as case #34, figure 3-3) events. They appear to be on the fuzzy boundary which separates these classes, and might be better classified as "hybrid" events. One could imagine a less strict validation system where such events were not penalized if they were classified as either linear or cellular. In such a case, it would not be unreason-

able to expect the percent correct to climb well above the 90% threshold obtained using these five raw attributes with "strict" scoring standards. In any event, this level of success demonstrates that useful attributes for an automated rainfall pattern classification system have been discovered.

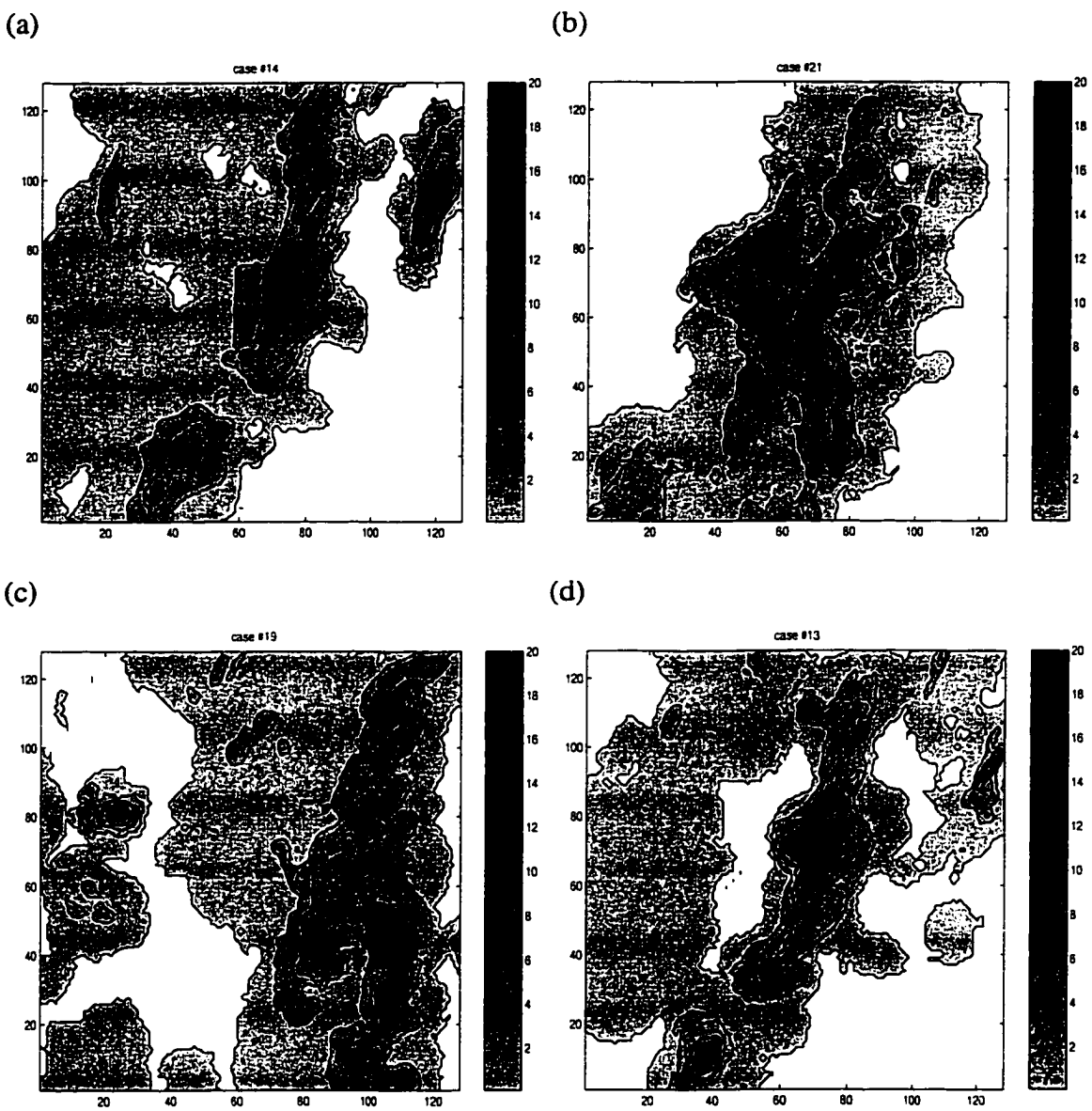


Figure 4-14: Cases #14 (a), #21 (b), #19 (c), and #13 (d) of the target data set. Colorbar on the side of each image indicates rainfall amounts in mm.

4.5 Summary

The automated classification algorithm using attributes produced by summary measures of the bulk statistical and geostatistical properties of rainfall patterns has successfully separated the target data set into linear, cellular, and stratiform classes. In order to confirm that these attributes are producing the desired effect of characterizing the spatial rainfall patterns in a way that is consistent with subjective impressions, the data matrix will be visualized in terms of the five attributes that produced the best classification results. Since visualization of a five-dimensional data set is somewhat difficult, the data matrix will be projected onto a two-dimensional plane that accounts for the largest possible fraction of the total variance contained in the data. Principal component analysis (PCA) of the Grammian matrix allows one to represent a data set in terms of a basis that is uncorrelated. Section 2.11 describes PCA in more detail. In this case, the correlation matrix (table 4-4) was used for the Grammian. Given that many of the attributes are highly correlated, especially the various *a/b* values, it is not surprising to find that the first two components explain 84.6% of the total variance. Therefore, a plot of the objects projected onto the directions of the first two principal components will be a fairly good two-dimensional representation of the five-dimensional data set. Figure 4-15 shows such a plot of the target data set in terms of the first two principal component scores (projected in the directions represented by the eigenvectors, normalized by dividing by the square root of the eigenvalues). Here, stratiform cases are grouped together in one part of the transformed "attribute space", while linear and cellular events are distributed in other sections of the plane. Cases on the fuzzy boundary between linear and cellular classes (such as #14 and #19) have some of the characteristics of both classes. What is important is not

whether each decision made by the automated classification agreed perfectly with those made by the subjective analyst. Instead, what is important is that the summary statistical measures are able to describe the degree of intensity and the degree of linear organization found within each object. This allows for meaningful comparisons of different objects. This plot shows that the goal of discovering attributes that characterize the intensity and degree of alignment of rainfall patterns has been met.

In this work, the determination of useful attributes that characterize important aspects of rainfall patterns has been built upon a relatively small target data set, comparing the results of automated classification experiments with a subjective classification. The target data set is a small, somewhat random sample of all possible rainfall objects that occur in nature. In order to get a better picture of the true distribution of rainfall objects in this attribute space, the characteristics of a large data set covering the entire year of 2002 will be analyzed in the next chapter, using a completely automated rainfall object detection and analysis system.

Table 4-4: Correlation matrix for the data matrix containing [β *a/b* 0.2 0.4 0.6 0.8] **attributes.**

	β	0.2 a/b	0.4 a/b	0.6 a/b	0.8 a/b
β	1.00	0.16	0.23	0.26	0.23
0.2 a/b	0.16	1.00	0.80	0.68	0.68
0.4 a/b	0.23	0.80	1.00	0.75	0.69
0.6 a/b	0.26	0.68	0.75	1.00	0.84
0.8 a/b	0.23	0.68	0.69	0.84	1.00

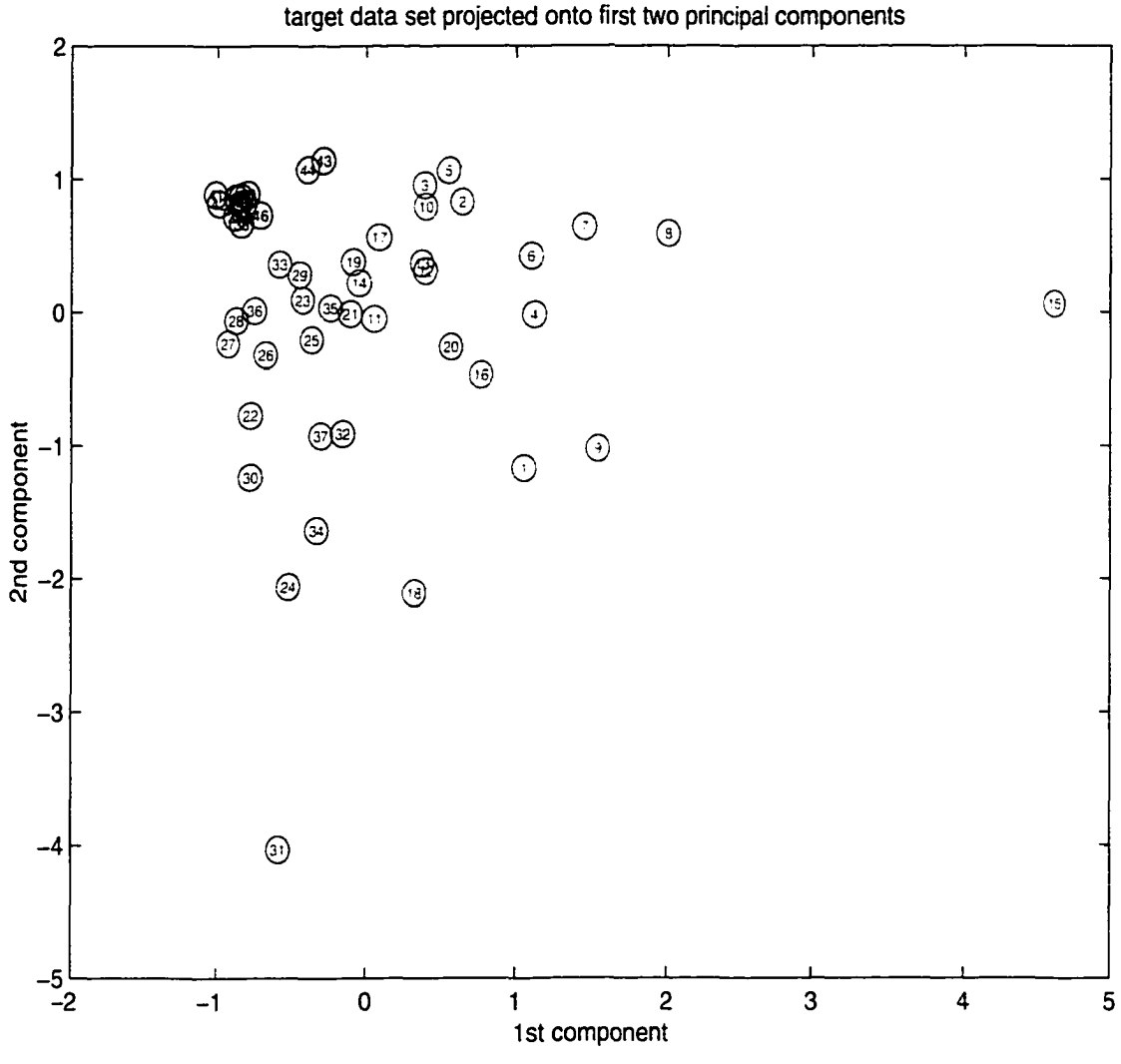


Figure 4-15: Projection of the objects in the target data set onto the first two principal component directions, normalized by the square root of the eigenvalues (PCA scores). Each object is color-coded by its subjective classification, linear events are green, cellular events are red, stratiform events are blue.

Chapter 5

Automated rainfall object classification system

5.1 Introduction

The overall goal of this work is to develop an automated rainfall pattern classification system. To accomplish this task, the discovery of a set of attributes that allow for automated characterization of rainfall patterns was required. To this point, the determination of useful attributes has been based upon a relatively small target data set, comparing the results of automated classification experiments with a subjective classification. The first step in this process involved a "baseline" automated classification using the raw values of rainfall at each point in space as attributes. Next, the dimension of the data was reduced by analyzing the "bulk" global distribution of rainfall values across each object, using the parameters of the gamma distribution fit to the observed histogram using the generalized method of moments technique. When the set of attributes was expanded to include those that summarize the geostatistical characteristics of the rainfall pattern, the classification system was able to separate the convective and non-convective events, and further separate the convective cases into linear and cellular events with over 90% accuracy. Therefore, it was concluded that a useful set of attributes had been obtained.

The target data set is a small sample of all possible rainfall objects that could occur in nature. In order to get a more complete picture of the population, the characteristics of a much larger data set must be analyzed. A completely automated rainfall object classification system will be developed to accomplish this task. A large data set covering a full year (2002) will be analyzed using an automated rainfall object identification and analysis

system. The automated classification system will be developed using the best results from previous classification experiments with the target data set. Summary statistics of analyzed attributes from this year will be examined.

In order to validate the automated classification system, a random sample of rainfall events taken from 2002 data will be verified. The distribution of the random sample will be compared with the full year's distribution to ensure that this validation data set is representative of the population. Once this has been confirmed, the sample will be classified subjectively and objectively via the automated classification system. Comparison of these results will provide independent confirmation of the accuracy of this classification system.

5.2 Automated rainfall object identification and analysis system

The bulk of this research has focused on the determination of a useful set of attributes that allow for accurate classification of rainfall patterns. Although the classification system used automated cluster analysis algorithms, the object identification process described in section 3.2 was purely subjective, not automated. To allow for practical implementation of this research, development of a fully automated system for rainfall object identification and characterization is required.

For this task, the discipline of image processing provides a wide range of tools. Although the concepts of an image are not new to meteorologists, some definitions of common terms will be provided at the start, a more detailed discussion of image processing is provided in section 2.12. An image is a representation of values onto a set of spatial coordinates. A pixel is an element of the image, represented by its location and value $(x,y,f(x,y))$. An object is a connected set of pixels, often a representation within the image of an entity in the physical world. It is convenient to assume that an object represents a set

of pixels that have fairly uniform characteristics, although this might not always be the case. For example, the brightness of an illuminated sphere will change continuously over the surface of the sphere. Computer-based images consist of data stored as grids with a finite number of spatial coordinates and image values, known as digital images. A Cartesian coordinate system is often used, where x and y take on integer values with intervals $1 \le x \le M, 1 \le y \le N$ for an image sized $M \times N$. The origin is typically taken to be the lower left corner of the grid. Although most meteorological analyses consist of continuous variables represented on discrete grids, in this work rainfall analyses are treated as digital images.

The task of object location and identification involves locating a proper set of connected pixels within the image. Determining a proper set of connected pixels can be accomplished in several ways. As previously mentioned, an object can be thought of as a set of pixels possessing somewhat uniform characteristics. This idea leads to one class of object identification algorithms, region-growing or *agglomerative* routines. These are developed on the basis of objects appearing to be fairly uniform, therefore an object can be built by grouping or clustering together similar-looking pixels. On the other hand, an object distinguishes itself from the background or from other objects by a perceptible change in characteristics, for example, changes in color, texture, or shading. The location of this change is considered the edge of an object. This concept leads to another branch of object identification algorithms, those related to *edge detection*. Here, an object will be identified as the region outlined by an edge on the image. Ideally, for rainfall analyses, any objects that are identified will represent significantly different meteorological phenomena. The difficulties in obtaining such a system using some edge detection or object

identification algorithms will now be discussed.

5.2.1 Agglomerative methods

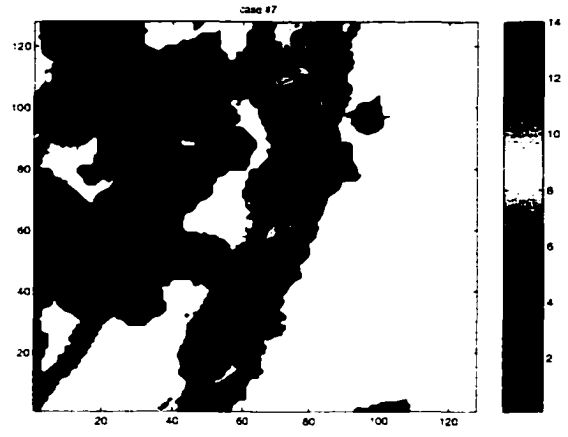
In general, agglomerative region-growing algorithms operate by clustering together pixels with similar characteristics. These routines perform a cluster analysis, as in section 2.3, where the data matrix is made up of pixels (the *objects* in cluster analysis terminology) and various attributes associated with them. A recent example of an agglomerative routine for processing weather-related images is provided by Lakshmanan (2001). Here, the *texture* of the image, represented by a vector of local statistical measures in the neighborhood of each pixel, becomes the attribute vector associated with each pixel. A set of K clusters are formed by minimizing a cost function, which is basically the sum of the Euclidean distances between each pixel and a cluster mean. In addition, the spatial location of each pixel is also included in the computation of the cost function. Therefore, clusters will contain pixels that are close together in terms of their texture and spatial location. This technique produces a hierarchy of objects over a range of spatial scales, where the number of clusters/objects is cut in half at each step in order to reach the next level of hierarchy. At some point, similar to the cluster analysis results that have been presented in previous chapters, a subjective decision as to the desired number of clusters or level of hierarchy must be made. This method does produce favorable results for weather radar and satellite images, and is currently being tested for its potential use in radar feature tracking algorithms at the National Severe Storms Laboratory (Lakshmanan 2003). However, it was not selected for this work since a subjective decision on the number of clusters or objects to keep for each image is required. While the selection of an acceptable threshold that would produce satisfactory results for any given rainfall image might be possible

to obtain, this would likely require a great deal of effort and tuning of the technique. This will be considered for future work.

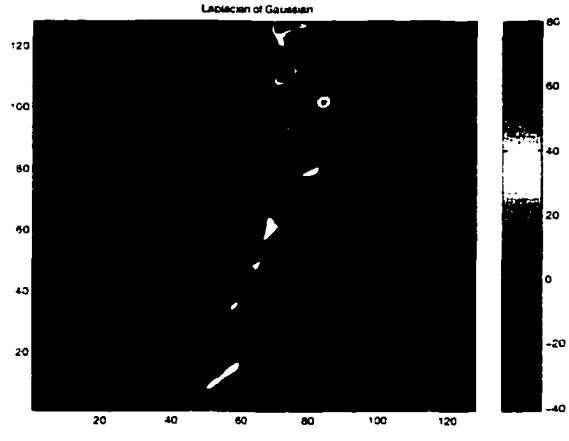
5.2.2Edge detection filters

Since the concept of object identification involves locating points where characteristics of the image are changing quickly, the idea of directly analyzing the gradient of intensity has been explored. A commonly used edge detection operator (Marr and Hildreth 1980) is based upon well-known property of the second-derivative in the vicinity of a edge, that is, edges are co-located with the zero-crossings of the second-derivative. While based upon calculus, the development of the operator was motivated by research into mammalian vision systems (Marr and Hildreth 1980), as the authors wished to make the algorithm as consistent with human vision processes as possible. The Marr-Hildreth operator combines the band-pass smoothing properties of the Gaussian with the edge-finding properties of the Laplacian (second-derivative); in fact, it is also known as the Laplacian of Gaussian (LOG) filter (Klette and Zamperoni 1996, see section 2.12 for more details). One might expect that rainfall may pose some problems for this type of algorithm. For example, the edge separating very light rain and no rain will correspond to a weak gradient; the zero-crossing of the second derivative in this instance will separate small values of the Laplacian. On the other hand, an edge separating heavy rain from lighter rain will correspond to strong gradients, and the zero-crossing of the second derivative will be found between much larger values of the Laplacian.

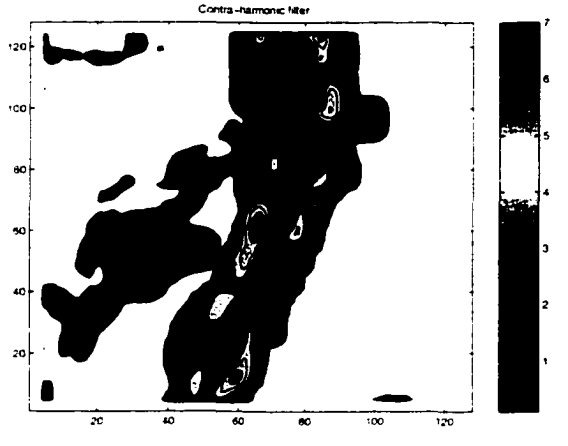
To determine the potential usefulness of the LOG filter for rainfall object edge detection, the LOG filter was applied to a rainfall analysis taken from the target data set (figure 5-1). The rainfall for this case consists of a region of light rainfall to the west of a linear



(a)



(b)



(c)

Figure 5-1: Examples of edge detection algorithms for a rainfall analysis (a). Rainfall values are in units of mm. Plot (b) shows results of Marr-Hildreth LOG filter for σ = 2.17 grid boxes. Plot (c) shows results of contra-harmonic filter for r=1.2. for a 9x9 window.

region of heavier rainfall (Figure 5-1a). The results of the LOG filter, with $\sigma = 2.17$ grid boxes, are shown in figure 5-1b. Again, the zero-crossing of this filter indicates the edges, with the zero contour separating the light blue and warmer colors from the region of darker blues. In this case, the edge detection algorithm highlights a separation between the region of heavier rain and the remainder of the image fairly well. However, for the indication of the edge of the rain/no-rain region found within the trailing stratiform area, the LOG filter does not produce satisfactory results.

Next, a second type of edge detection filter was tested. The contra-harmonic filter (Klette and Zamperoni 1996, see section 2.12 for more details) highlights edge pixels based upon the difference between estimates of the local maximum and minimum values within a window. These estimates are non-linear calculations of averages (contra-harmonic average) of the local pixel values. The basic idea behind this filter is that, in the vicinity of edges, large differences between local maximum and minimum values should exist. Once again, for rainfall one might expect that this sort of filter would have some trouble detecting the edge of a region of light rainfall, since the difference between the local maximum and minimum will be quite small. Figure 5-1c shows the results of the contra-harmonic filter with r=1.2 and a 9×9 grid point window (see section 2.12). In this case, edges are indicated by large values of the filter. Here, large values are indicated on either side of the linear area of heavier rainfall. Therefore, this filter does show promise in separating regions of heavier rain from the remainder of an image. However, there are small values of the filter throughout the region of light rain, even beyond the rain/no-rain contour. In this case, as expected, the edge detection algorithm does not provide satisfactory results in the determination of the rain/no-rain edge.

At this point it is not obvious how best to take advantage of the information provided by these edge detection or object identification algorithms. The agglomerative region-growing methods require a decision regarding the number of objects to locate within the image. The task of detecting an edge which separates zero rain from light rainfall proved to be problematic for both of the filters tested here. Perhaps these filters could be used to find edges separating heavy and light rainfall regions, and some other algorithm could be used to locate the edge separating the rain/no rain regions. In order to use the information provided by either of these edge detection filters to locate objects within the image, additional processing would be required. For example, in the case of the contra-harmonic filter, a threshold value would need to be selected that would indicate edge locations. In the example shown here, there are gaps between large values from the filter, so some sort of interpolation between highlighted edge points would likely be necessary in order to completely outline a rainfall object. For the LOG filter, a decision on the spatial scales one wishes to smooth through the choice of σ is also required. Marr and Hildreth (1980) recommend using the filter over a range of scales, as one will find certain edges across a wide range of scales and other edges will only be found at smaller scales. They provide a set of rules for combining the information from different applications of the LOG filter to create a description of the image that they call the *raw primal sketch*. Again in this case, some sort of interpolation between edge segments found by this filter would be needed to completely identify a rainfall object. One can easily envision a very complicated, heuristically based scheme for locating objects using either of these types of edge detection algorithms. There is no reason to expect that, even after a considerable amount of effort is expended, such a system will produce ideal results. Therefore, a more natural choice of edge detec-

tion for rainfall will be made.

5.2.3 Binary edge detection

Since the object identification and edge detection algorithms discussed previously did not provide a clear solution to the object location and identification problem, a simpler method of converting the image to binary through the use of a simple threshold was chosen. For a binary image, consisting of only light or dark pixels, the edge detection task is fairly simple (see, for example, the algorithm provided in equation 2.44). Here, edge detection for a dark object involves finding those dark pixels which are located immediately next to light pixels. In order to convert a greyscale image into a binary image, a threshold must be selected. Once this has been accomplished, a straightforward binary edge detection algorithm can be applied to the converted image.

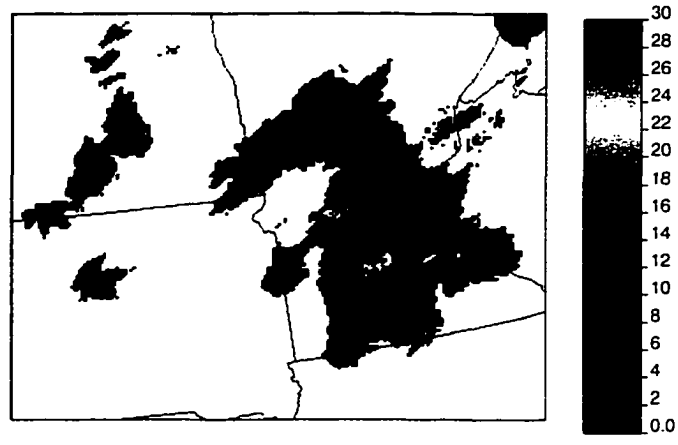
The problem now becomes one of selecting an appropriate threshold. As discussed in Davies (1996), the choice of threshold will depend on the practical problem that the user wishes to solve. For example, if the task is optical character recognition, the objects are typically dark shaded letters and numbers that have been printed onto a lighter shaded background. If the grey values represented by the dark characters and the light background are internally consistent, the proper choice of threshold can be obtained by analyzing the histogram of pixel intensity values. A pixel value located in a valley on the histogram separating the dark and light peaks makes an ideal threshold. However, there are often major difficulties with such an approach. For example, the background brightness values may vary substantially across an image, such as what one might find in a poor quality facsimile or Xerox copy. In this case, an adaptive thresholding technique might be beneficial. This involves determining a local threshold based upon the pixel values in a

neighborhood surrounding each location.

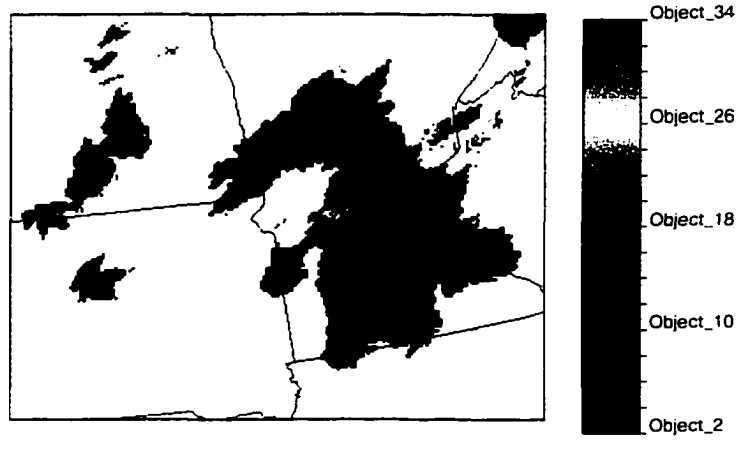
Considering the case of rainfall data, as well as other types of images, an ideal threshold is often difficult to find. The histogram may have multiple minima due to detail or noise in the image. The pixel intensity histogram may be highly skewed, making it difficult to determine a relevant threshold within the tail of the distribution. A quite simple and natural choice of threshold for rainfall data is one that separates measurable rain from no rain. Rainfall objects in this case will be connected regions of rainfall greater than the rain/no-rain threshold. This method for object identification has been selected for the automated object classification system for the remainder of this work. This simple method certainly avoids many of the problems related to developing a more complex object identification technique. This also solves the problem that the various edge detection filters had in locating the light-rain/no-rain edge. On the other hand, this method will likely often produce large objects that contain several different types of rainfall events combined into a single object. For example, a line of convection connected to a trailing region of stratiform rainfall would be considered a single object when a rain/no-rain threshold is used. The selection of rainfall object identification technique for the automated system has been made based on the desire to include all measurable rainfall while also allowing for relatively simple implementation and understanding of the procedure. A detailed explanation of this method will be provided by an example image in the next section.

5.2.4 Example

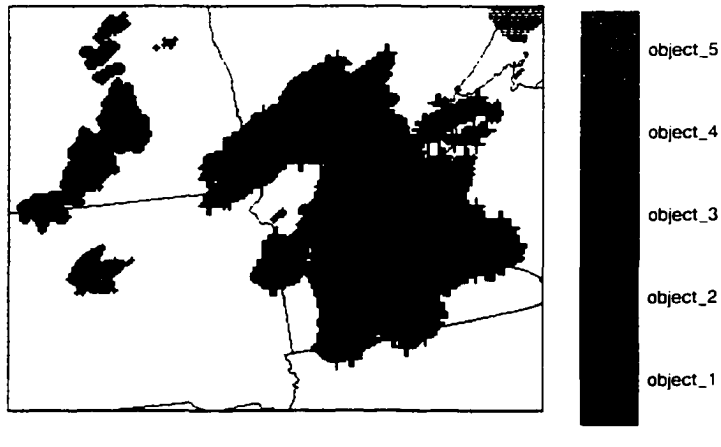
The operation of the rainfall object identification algorithm used for the remainder of this work will be illustrated via an example. In figure 5-2a, a small subset of the Stage IV



(a)



(b)



(c)

Figure 5-2: Steps of the rainfall object identification process. Top panel (a) shows 1h rainfall valid 23 UTC 28 July 2002. Middle panel (b) shows initial connected region labelling. Lower panel (c) shows final rainfall object labelling.

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 scattered light rain appear in Wisconsin just east of the main heavy rainfall region. As previously mentioned, a simple threshold (=0.05mm) was used to convert the rainfall image into a binary image (using the threshold algorithm given in equation 2.41). The connected component labeling algorithm (section 2.12) is 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. The result of this labeling is shown in figure 5-2b. Note that each contiguous region of rainfall is plotted with a different color, each connected region is assigned an integer label value, indicated by the colorbar on the edge of the figure. As discussed in section 3.5.2, since the atmosphere is diffusive, it is physically sensible to expect that a significant fraction of locations receiving non-zero precipitation over a region will receive less than a measurable amount of rain (also known as a "trace" amount). The size of the area receiving trace amounts of precipitation was specified to be 15% of the total area receiving detectable precipitation, executed in the same way as previously described for the target data set analysis. The areal extent of each object was increased by a integer number of pixels in each direction such that the object's area was increased by as close to 15% as possible. In addition, it is not unusual to find small "gaps" between nearby regions of measurable rain, such as those found between the large contiguous rainfall region and the scattered pixels of light rain just to the east. Therefore, the definition of "connected" pixels was expanded so that pixels that were within 5 points (~20km) of one another were considered connected, and therefore given

the same label value. Figure 5-2c displays the final result of this process. At this point 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, object_4 being a small region of scattered light rain in eastern North Dakota.

Now that the object identification process is complete, the rainfall objects will be analyzed so that statistical characteristics can be extracted. The analysis process for the large object in the previous example (object_1, figure 5-2c) will be examined in detail in order to illustrate the process. The parameters of the gamma distribution (α, β) will be fit to the histogram of rainfall amounts for each object using the generalized method of moments (three-moment q=1 GMM, sections 2.9.3, 3.5.3). For this example, the resulting parameters of the gamma distribution were estimated to be α=0.30 and β=6.95. Figure 5-3 shows the observed distribution of rainfall amounts along with the theoretical probability distribution using these parameters. This distribution is strongly skewed, resulting in a low value of α, and has a fairly thick tail, indicating the presence of heavier rainfall, and reflected in the relatively high value of β.

A correlogram will be computed for each object as well. The correlogram represents the auto-correlation for each object individually. Only the rainfall contained within each unique object is included in the calculation. In other words, the rainfall from object_2 will

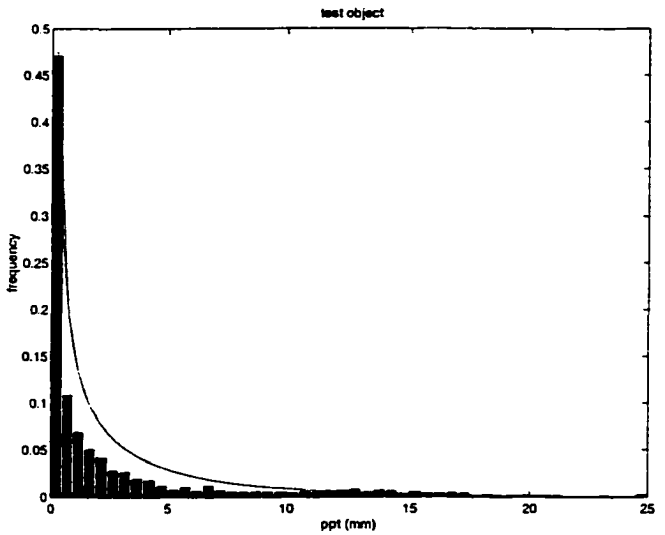


Figure 5-3: Histogram of rainfall amounts (mm) for example rainfall object (object_1, figure 5-2c). Curve indicates gamma distribution fit to the histogram by GMM, resulting in α=0.30 and β=6.95.

not affect the correlogram for object_1. In order to limit the computation time needed to compute each correlogram for very large objects, the maximum lag that can be analyzed was set to 181 pixels (approximately the length of $\boldsymbol{h}$=[128,128]). This was found to be greater than the largest distance between the origin and the 0.2 contour on correlograms computed without this limit for all objects during the month of January, 2002. Since this is a cool season period, one would expect to find a high degree of organization in several large, synoptic-scale rainfall objects, therefore this maximum lag should be quite applicable to the rest of the year.

As discussed in section 4.3, various contours surrounding the origin will be analyzed using several image processing routines described in section 2.12, and features of ellipses fit to those contours will be extracted. This process is illustrated in figure 5-4. Since the correlogram (figure 5-4a) is symmetric, only the top half is shown in subsequent figures, since that is all that is required in order to estimate the lengths of the semi-major and semi-

minor axes of an ellipse. First, the correlogram is converted to a binary image (figure 5-4b), where pixels greater than the contour threshold are set to 1. Next, the connected component labeling algorithm is used to find the region of the binary image that is connected to the origin (figure 5-4c). The binary edge detection algorithm is used to locate the pixels that define the edge of the connected region (figure 5-4d). The largest distance from the origin to the edge is considered the length of the semi-major axis, *a*. The smallest distance from the origin to the edge is the length of the semi-minor axis, *b*. In this case, the edge pixel that is furthest from the origin is located at [x,y]=[-4,22], therefore *a*=22.36. The closest edge pixel is at [-14,7], therefore *b*=15.65. Once these lengths are found, as in section 4.4, the ratio and product of these axes will be taken (*ab*=350 and *a/b*=1.429 in this case), as well as the counterclockwise angle between the semi-major axis and the x-axis (= 100.3°). This angle is rotated to adjust for the polar-stereographic map projection, making the angle the clockwise angle between *a* and due East. For a polar-stereographic projection (Pearson 1990), the rotation angle is simply $\Delta\theta$ = central $\mathrm{lon}_{proj} - \mathrm{lon}_{obj}$, where the central longitude of the projection for the Stage IV analysis is 255E, and the longitude of the object is in degrees east (in this case, the angle of the semi-major axis after rotation = 89.5°). Table 5-1 lists these attribute values for the four complete objects represented in figure 5-2c.

In summary, for each object, the following set of attributes will be obtained and stored; date, time, location (center of mass x, y), average rainfall, size (number of pixels), α, β from GMM, *a/b*, *ab*, and θ for the 0.2, 0.4, 0.6, and 0.8 correlation contours. The system for classifying these objects based upon the attributes extracted via the methods outlined here will be described in the next section.

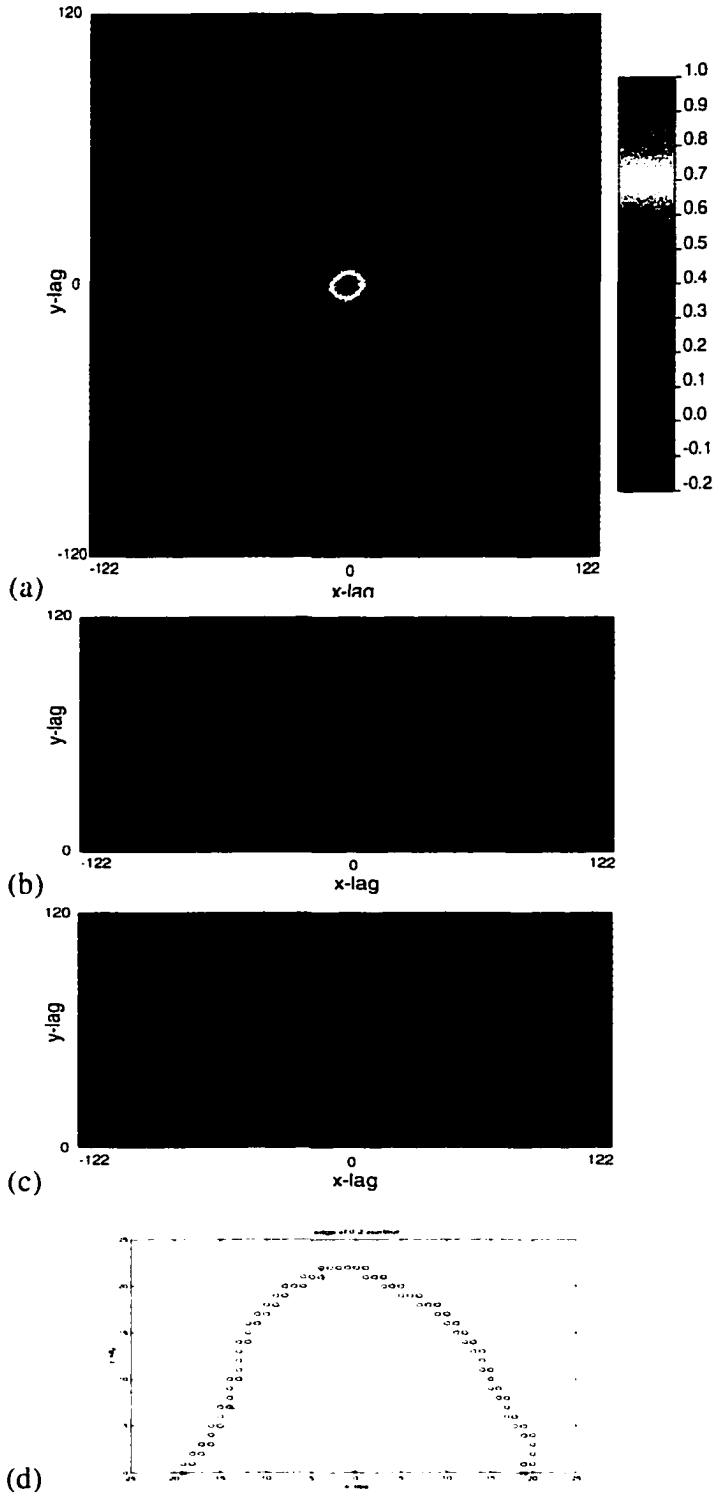


Figure 5-4: Correlogram 0.2 contour analysis process. Correlogram (a) from the example object (object_1, figure 5-2c), (b) 0.2 contour binary image, (c) 0.2 region connected to origin, (d) location of edge pixels for 0.2 connected region. Red line indicates furthest distance from origin to edge (semi-major axis, *a*), green line shows shortest distance (semi-minor axis, *b*).

Table 5-1: A sample of attributes extracted from the four objects found in figure 5-2c.

label	α	β	*a/b* 0.2	θ(deg)	size(pixels)
object_1	0.30	6.95	1.43	89.5	7644
object_2	0.16	4.67	6.32	7.6	275
object_3	0.23	5.05	2.85	58.6	1150
object_4	0.03	0.06	1.00	0.0	28

5.3 Automated rainfall object classification system

Now that a completely automated system for rainfall object identification and analysis has been developed, an automated system for classifying those rainfall objects is required. In the previous chapters, a hierarchical cluster analysis algorithm was used as a classification tool. When using such an algorithm, a subjective decision is required in order to split the data set into a fixed number of groups or clusters. The clusters that were found to be in closest agreement to the subjective classification in the previous chapter will be used to build the automated classification system here.

Figure 5-5 repeats the results (figure 4-13a) obtained from the hierarchical cluster analysis performed in section 4.4. This involved the use of raw, unnormalized attributes consisting of β, and *a/b* from the 0.2, 0.4, 0.6, and 0.8 correlogram contours. The analysis resulted in five major clusters and six outlier cases (cases 1, 8, 9, 12, 15, 18, and 31). The percent correct in the 3-class (stratiform, linear, cellular) case was 90.5%, and 100% in the 2-class (convective, non-convective) case. The automated classification system will place objects into one of these five clusters, depending on which is closest to the object in terms of Euclidean distance. The location of each of these clusters will be defined by the mean of the attribute vector for the members of each cluster. The cluster mean attribute vectors

are provided in table 5-2 .

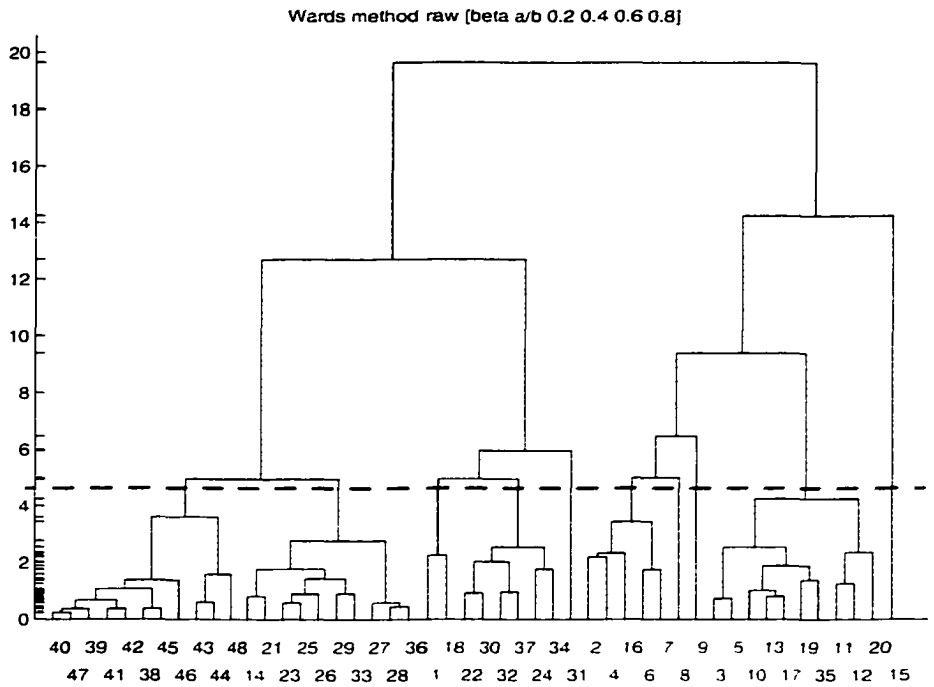


Figure 5-5: Dendrogram produced by Ward's method with target data set using raw [β, *a/b* 0.2 0.4 0.6 0.8 contours] attributes. Each object is color-coded by its subjective classification, linear events are green, cellular events are red, stratiform events are blue. Dashed line indicates subjectively determined cut-level.

Table 5-2: Cluster mean attribute vectors, from clusters denoted in figure 5-5.

	β	*a/b* 0.2	*a/b* 0.4	*a/b* 0.6	*a/b* 0.8
cluster 1	0.75	2.72	1.91	1.61	1.76
cluster 2	2.82	2.72	2.39	1.90	2.07
cluster 3	5.62	1.97	2.10	2.31	1.79
cluster 4	3.56	6.36	8.08	3.66	3.09
cluster 5	2.63	5.28	4.03	2.89	2.80

As shown in figure 5-5, cluster 1 is unanimously populated by stratiform events, therefore it is considered the stratiform cluster. The distinguishing feature of the stratiform cluster mean is the low value of β. The other convective clusters all have considerably higher mean β values. Cluster 2 is mainly a cellular cluster, with an additional linear

member. Therefore, it will be considered a cellular cluster, with perhaps some hybrid linear characteristics. Cluster 3 is unanimously populated with cellular events, therefore it is considered a cellular cluster also. Clusters 2 and 3 have relatively low mean values of *a/b*, indicating a characteristic lack of linear continuity. Cluster 3 has a considerably higher mean value of β than cluster 2, indicating a thicker-tailed distribution and therefore more frequent instances of heavy rainfall. Cluster 4 is entirely a linear cluster, and cluster 5 contains mainly linear events with a few cellular cases included. These will be considered linear clusters, although cluster 5 might be considered to be closer to the fuzzy boundary between linear and cellular, perhaps more precisely called a linear/hybrid class. Clusters 4 and 5 have relatively high mean values of *a/b*, indicating their linearly-organized character. Cluster 4 contains the highest mean values of *a/b*, therefore one might expect cases belonging to this class to be on the highly-organized end of the spectrum.

The automated classification system will proceed as follows. The raw values of the five attributes (β, *a/b* 0.2, 0.4, 0.6, and 0.8 contours) for a given object will be compared to the five cluster mean vectors described previously. The Euclidean distances (equation 2.3) between the object and each of the five cluster means will be computed. The object will be placed into the class represented by the nearest cluster mean in terms of the smallest Euclidean distance, in other words, the nearest-neighbor cluster.

This automated classification system has been built upon the cases found in the target data set. As previously mentioned, it is not clear at this point whether or not this target data set is a representative sample of the range of possible rainfall events found in the population. A large data set covering a full year (2002) will be analyzed using the automated rainfall object identification and analysis system. Summary statistics of analyzed

attributes from this year will be examined. In order to validate the automated classification system, an independent random sample of rainfall events from 2002 will be taken. The distribution of the random sample will be compared with the full year's distribution to ensure that this validation data set is representative of the population. Once this has been confirmed, the sample will be classified subjectively and objectively via the automated classification system. Comparison of these results will provide an independent confirmation of the accuracy of this classification system.

5.4 Validation of automated classification system

5.4.1Summary statistics

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 for the entire year of 2002. As discussed in section 3.2, the Stage IV analysis is a national mosaic of optimal estimates of hourly accumulated rainfall using radar and raingage data, which is available on a 4km x 4km mesh covering the contiguous 48 states. Unlike in the earlier chapters, the domain that was analyzed included the entire lower 48 states. Each hourly analysis was processed using the automated rainfall object identification and analysis system described in section 5.2. Out of a possible 8760 hours over the course of the year, 8679 hours, or 99.1% of the hours in the year were included in the data set that was obtained from NCEP. In total, 799014 objects, 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 figure 5-6. The histogram (figure 5-6a) clearly shows that the majority of objects are relatively small in size. This can be further illustrated by examining the cumulative distribution function (figure 5-6b), where the

probability of an object being larger than X pixels is plotted versus the size (X) of an object. Studies related to self-organized criticality (e.g., Bak et al 1987; Song et al 2002) have found that many naturally occurring phenomena (forest fires, avalanches, earthquakes) when plotted in a similar fashion (size vs. frequency) show power-law scaling properties, that is, their distributions follow a straight line on a log-log scale. In this instance, one might determine three separate regimes where such scaling power-law properties exist, one in the range of approximately 5-150 pixels, another between 150-2000 pixels, and the third for objects larger than 2000 pixels. In terms of physical dimensions for the transition points separating these three regimes, 150 pixels is approximately (50 km)^2 while 2000 pixels is approximately (200 km)^2. It is interesting to note that these length scales are quite close to those suggested by Orlanski (1975) for different regimes of mesoscale phenomena (meso-gamma scale features are of length scale 2-20km, meso-beta are between 20-200km, and the length scales of meso-alpha scale phenomena are between 200-2000km). Given these three size regimes, the objects can be grouped into three size-related categories, small (meso-γ) objects of size 150 pixels or less, medium (meso-β) sized objects greater than 150 pixels and less than or equal to 2000 pixels, and large (meso-α) objects of size greater than 2000 pixels. For instance, in the 2002 data there are 524224 small objects (65.6% of the total, an average of 60/hr), 242914 medium size (30.4%, average 28/hr), and 31876 large (4%, 3.7/hour) objects. For reference, figure 5-2c provides examples of typical objects in each size regime, a large object over Minnesota (object_1: 7644 pixels), medium sized objects in North Dakota (object_3: 1150 pixels) and South Dakota (object_2: 275 pixels), and a small object in eastern North Dakota (object_4: 28 pixels).

(a) (b)

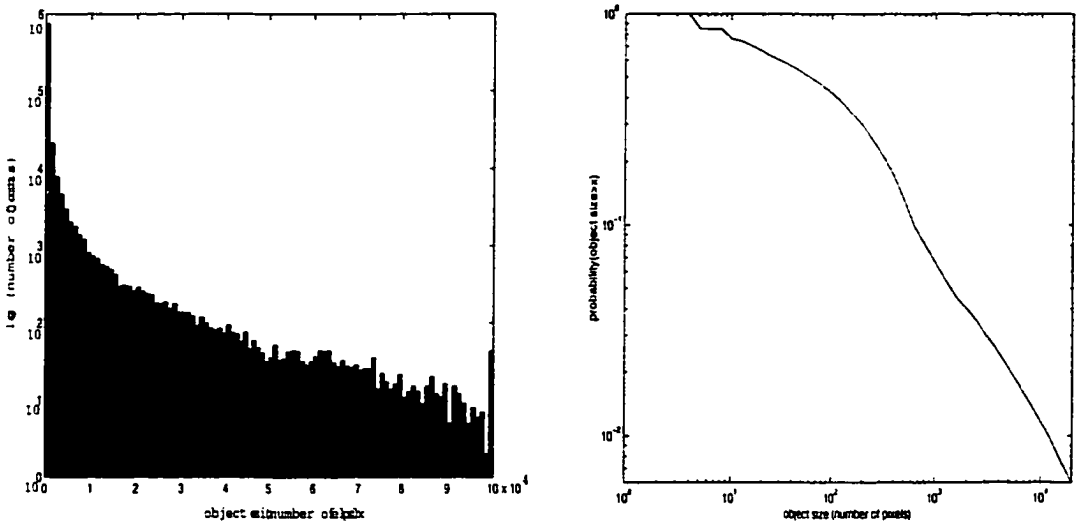


Figure 5-6: Distribution of 2002 rainfall objects by size. Left panel (a) shows histogram, right panel (b) shows probability of object size being larger than x pixels versus x, on a log-log scale.

The diurnal cycle of the 2002 rainfall objects is displayed in figure 5-7. Overall, the maximum number of rainfall objects occurs in the late afternoon (21 UTC) and the minimum occurs in the early morning (12 UTC). However, the times of these maxima and minima are a function of object size. For example, small-sized objects (figure 5-7b) are most frequent at 00 UTC and show a minimum in frequency at 16 UTC. The peak frequency for medium-sized objects (figure 5-7c) occurs earlier than for the overall distribution, at 19 UTC, while the minimum is in the early morning (12 UTC). For large objects (figure 5-7d), the maximum occurs at 22 UTC while the minimum also occurs at 12 UTC. The monthly distribution of objects is shown in figure 5-8. The overall distribution shows a peak in July, and a minimum in February. Again, there are differences in the monthly distributions depending on the object sizes. The distribution of small-sized objects (figure 5-8b) is quite similar to the overall distribution, except that the minimum occurs in January instead of February. Medium-sized objects (figure 5-8c) show no clear peak or valley

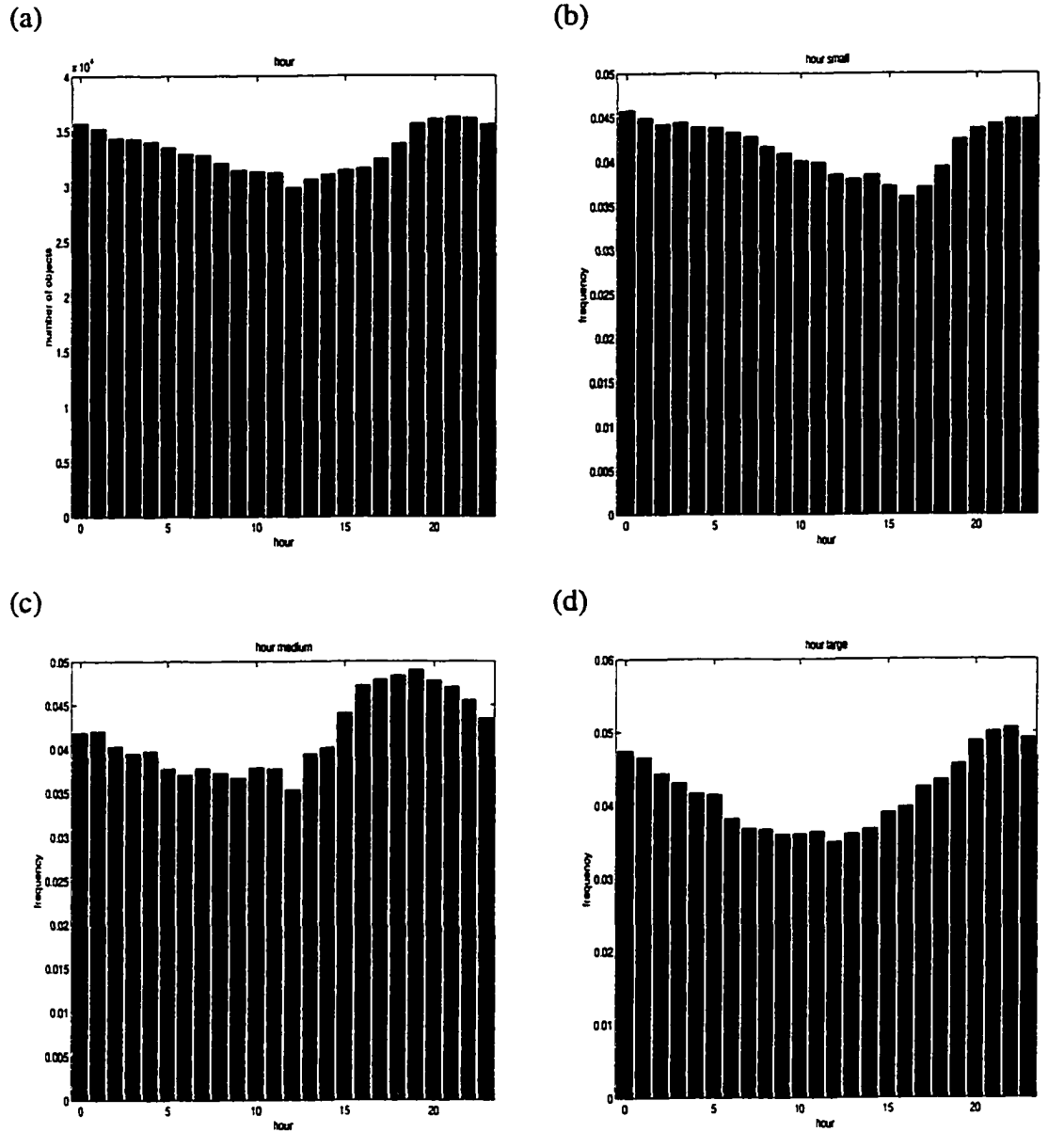


Figure 5-7: Hourly distribution of 2002 rainfall objects (UTC). Top left panel (a) shows number of objects for all sizes, top right panel (b) shows relative frequency of small objects, lower left panel (c) shows relative frequency of medium-sized objects, and lower right panel (d) shows relative frequency of large objects.

in their distribution, although the maximum does occur in July and the minimum in February as in the overall distribution. Finally, the monthly distribution of large objects (figure 5-8d) is consistent with the overall distribution. In general, these results are consistent with other studies of rainfall climatology (Geerts 1998) which show peak frequencies in

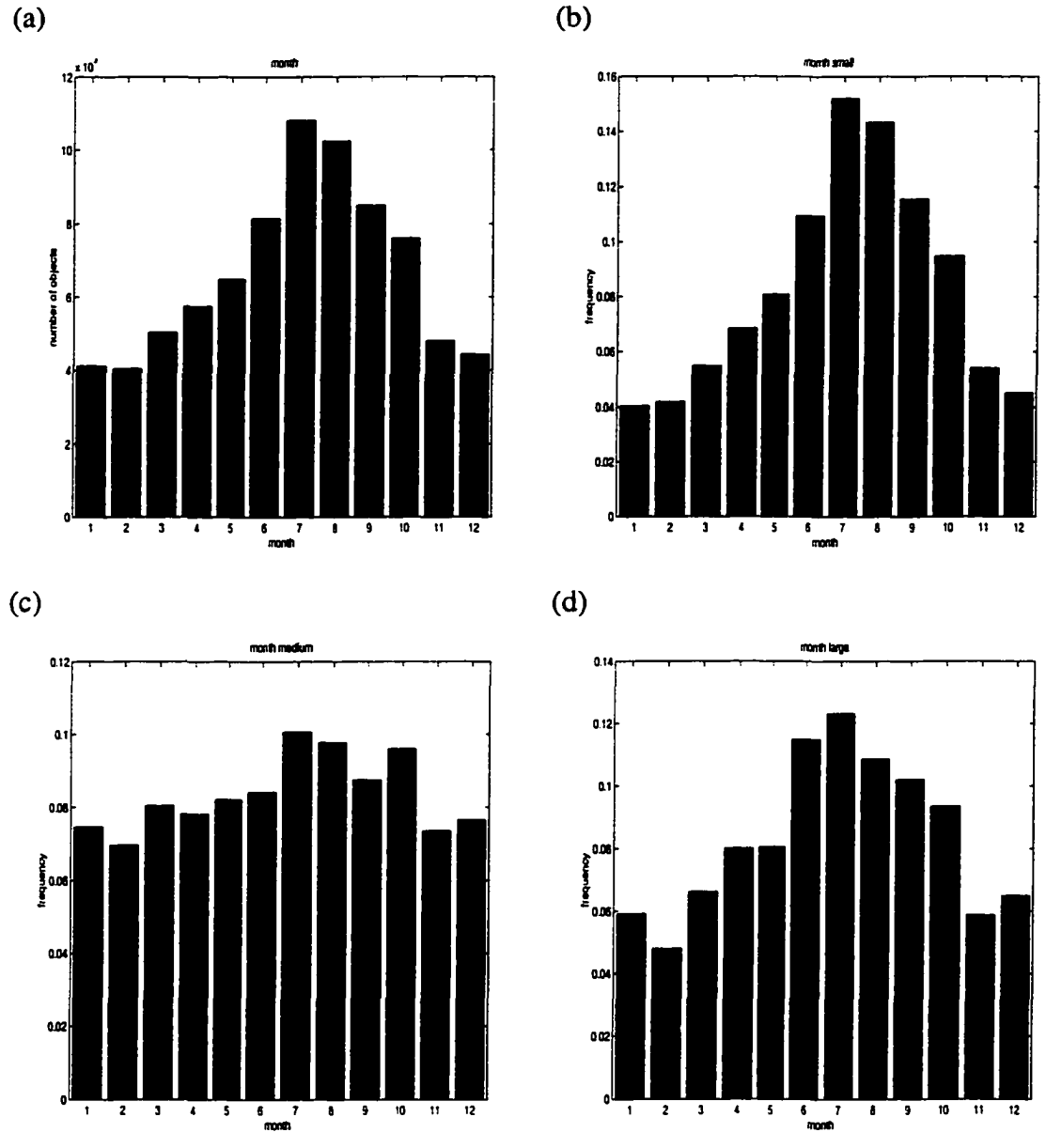


Figure 5-8: Monthly distribution of 2002 rainfall objects (1=January, 12=December). Top left panel (a) shows number of objects for all sizes, top right panel (b) shows relative frequency of small objects, lower left panel (c) shows relative frequency of medium-sized objects, and lower right panel (d) shows relative frequency of large objects.

the late afternoon and the warm season, etc.

The distribution of object center of mass locations (figure 5-9) will be examined next. To help to visualize this distribution, the object center of mass locations will be analyzed onto a grid with approximately 80km spacing. This will be done simply by sum-

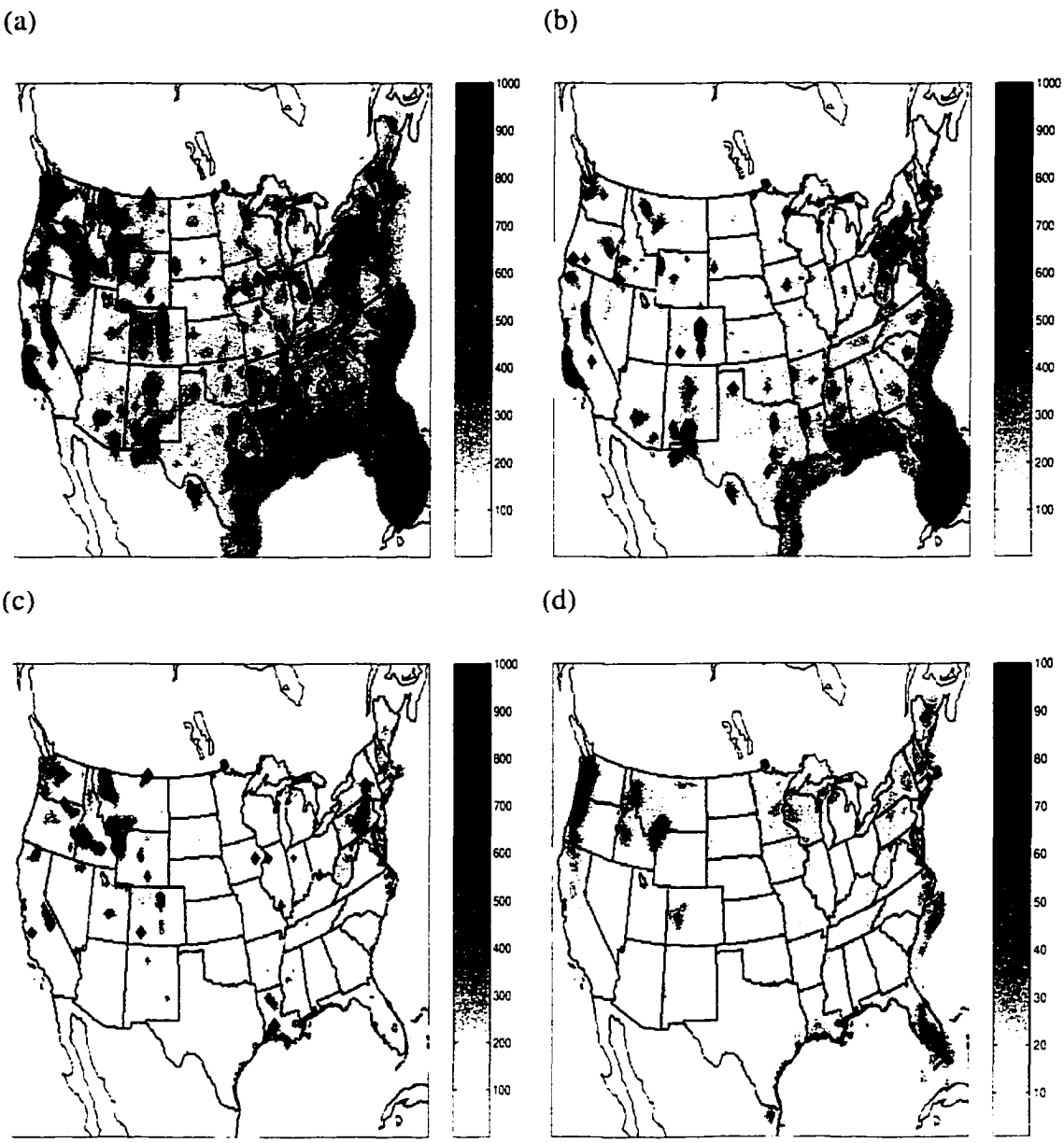


Figure 5-9: Spatial distribution of 2002 rainfall objects averaged onto 80km x 80km size grid boxes. Top left panel (a) shows number of objects for all sizes, top right panel (b) shows number of small objects, lower left panel (c) shows number of medium-sized objects, and lower right panel (d) shows number of large objects (note thresholds for colorbar are order of magnitude smaller in this panel).

ming all objects found within non-overlapping sets of 17×17 grid points from the original grid (Stage IV grid spacing is 4.7625km, 17×4.7625 km = 80.96km) to determine the number of objects on each 80km grid box. For comparison against climatology,

a plot of the mean number of days with measurable precipitation from a 30 year period (1961-1990) across the U.S. is shown in figure 5-10a (NOAA 2002). While this statistic is not the same as the number of hourly rainfall objects as identified here, it is similar enough for comparison purposes. The distribution of the objects of any size (figure 5-9a) shows several maxima. Maxima located across the western U.S. likely correspond to the higher terrain of the Rockies. It is not clear what fraction of these objects are actual rainfall and what fraction are anomalous objects caused by blockage of the radar beam by the high terrain. Similar maxima in daily rainfall frequency are found in the 30 year climatology (figure 5-10a). Another large area of high object frequency is located across South Florida. Again, this region shows relatively high daily rainfall frequency in the 30 year climatology (figure 5-10a). This is not surprising to anyone familiar with the weather of this region, as Burpee (1989) noted in the warm season, "a day without significant rainfall or radar echoes is rare" in South Florida. In general, there is a relatively high frequency of hourly rainfall objects across the southeastern U.S. In particular, the number of objects tends to increase as one moves away from land in this region. Since these areas are on the edge of radar coverage and there are no rain gages over water that can be included in the multi-sensor analysis, the spatial rainfall patterns will be dominated by the radar estimates, which generally contain more spatial detail. It is likely that this will result in a larger number of smaller objects than would be found over land in the vicinity of rain gage observations. Therefore, it appears that this feature may in part be a spurious effect of the analysis system. Another relatively large maximum in object frequency is located across Pennsylvania and West Virginia. This is a region of significant terrain variation, while not as large as in the western U.S., this could also be related to radar beam blockage or oro-

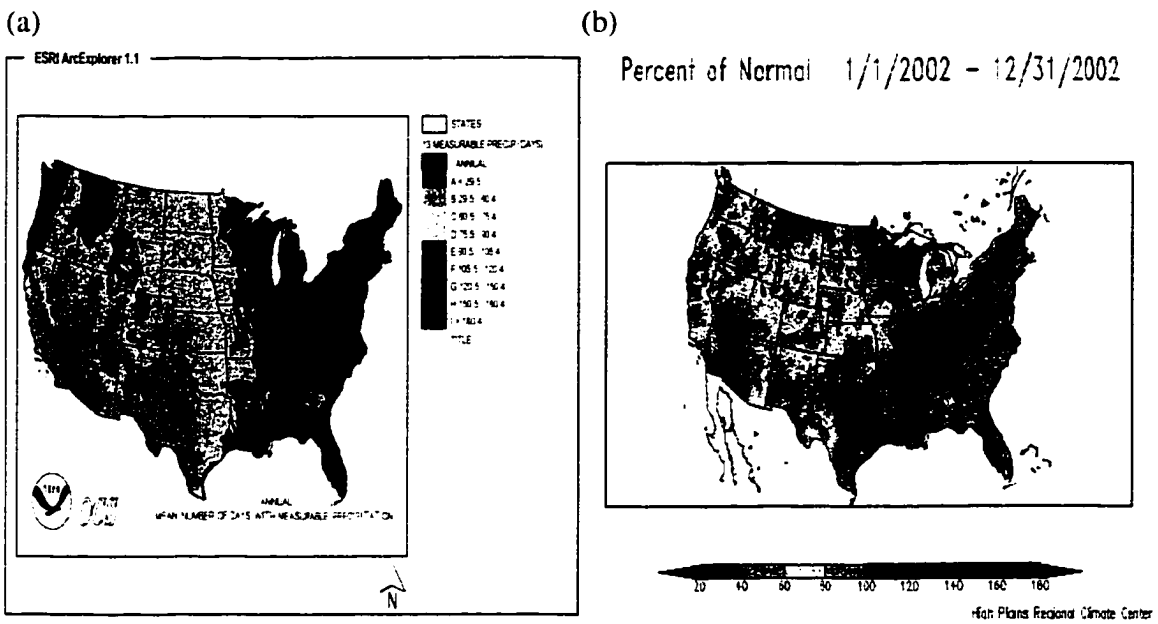


Figure 5-10: Left panel (a), annual mean number of days with measurable (>0.01 inch/ day) precipitation 1961-1990 (from NOAA 2002). Right panel (b), percent of normal annual precipitation for 2002 (from NOAA 2003).

graphically-forced rainfall. Regions of relatively few rainfall objects include southern regions of Nevada and California, and typical downslope regions just east of the Appalachians and the high Plains just east of the Rocky Mountains. Similar minima are found in the 30 year climatology (figure 5-10a).

The spatial distribution of the small (150 pixels or less) objects (figure 5-9b) is quite similar to the overall distribution, confirming that this size is the most frequent rainfall object observed in the 2002 data. As previously discussed, a large number of the objects found in the maxima over South Florida and off the southeastern U.S. are small in size. The distribution of medium-sized objects (151-2000 pixels, figure 5-9c) is considerably different from that of the small-sized objects. A large region of relatively high frequency for medium-sized objects is found across both the Rocky and Appalachian Mountains, as well as over the southeastern U.S. The distribution of large-sized objects (> 2000 pixels, figure 5-9d) shows a peak along the Cascade range of the western Rockies, along with a

maximum over the Florida peninsula. The distribution across the rest of the eastern half of the U.S. is fairly uniform. In some regions of southern Nevada and California, no large rainfall objects were found at any time during 2002. Recall that these object locations are the centers of mass for each object, and therefore a low frequency of occurrence does not indicate a complete lack of rainfall for that region. On the other hand, as shown in figure 5-10b, 2002 was abnormally dry in the southwestern U.S., so a sparsity of rainfall objects in this region is consistent with the NCDC 2002 climate assessment (NOAA 2003).

Turning now to the distributions of attributes associated with each object, a question

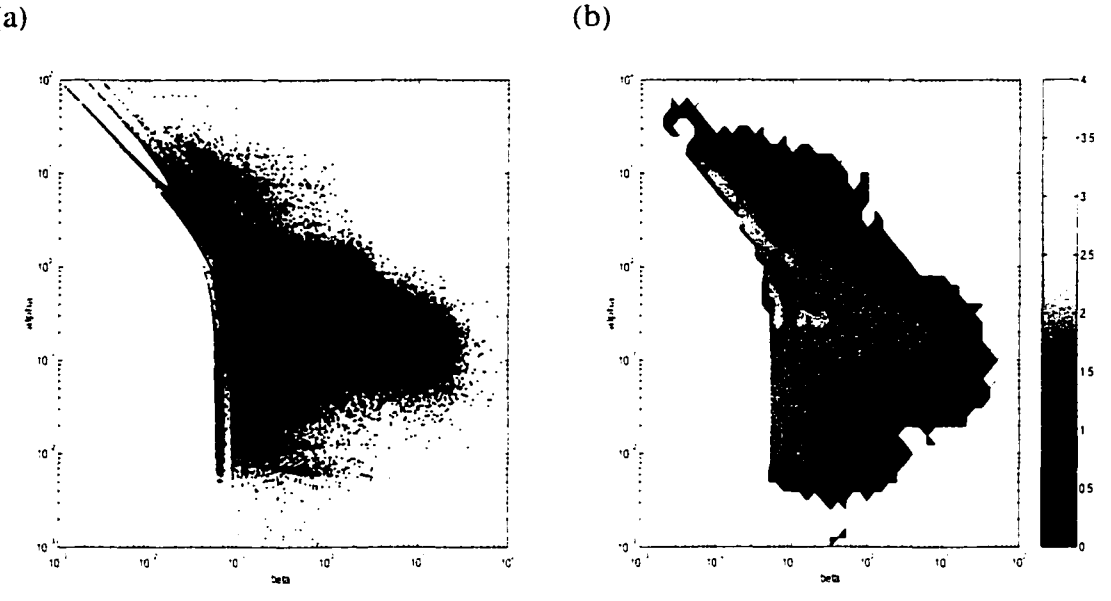


Figure 5-11: Distribution of 2002 rainfall objects in terms of attributes α and β. Left panel (a) shows scatter plot of objects for all sizes. Right panel (b) shows the density of objects (log(number of objects) per grid box) on a regularly spaced grid in log(α), log(β) space, consisting of 51 × 51 grid points. (note values for colorbar are in terms of log(number of objects).

as to how best to visualize such distributions arises. Since the objects of the target data set were originally plotted in α, β space (figure 3-13), a scatter plot of the 2002 data will also be made in this attribute space (figure 5-11a). This plot uses log-log axes to assist in the

visualization of the data. The reason why many of the objects tend to fall along some thin lines in α,β space for very small values of β is not clear, this is likely an computational artifact of the GMM algorithm. One can easily see that a large number of the objects are located in the range of α from 0.1 to 1.0 and β from 0.1 to 10. However, there is no way to assess the density of objects in this region since the dots representing each object in this region overlap to such a great extent. Therefore, it seems reasonable instead to analyze the density of objects in attribute space. Conceptually, this is accomplished by laying a regularly spaced grid in $\log(\alpha)$, $\log(\beta)$ space consisting of 51×51 grid boxes over the scatter plot, and objects found within each grid box are counted. The log(number of objects) is displayed in figure 5-11b. The density of objects in this space can now be visualized. One can now see where the various maxima lie in this space. Thousands of objects (per "grid box") have attribute values of α from 0.1 to 1.0 and β from 0.1 to 10.

Object density plots of this type can also be made for different size regimes (figure 5-12). Small objects account for nearly all of the objects in the 2002 data that have values of α less than 0.1, as well as a large number of the objects that have the tiny values of β ($<$ 0.05). In fact, nearly 90% of the small objects have values of β less than 0.5. Given these attributes, objects of this type must have a very strongly skewed histogram containing light values of rainfall. Similarly, medium-sized objects tend to have small values of β (nearly 80% of these objects have $\beta < 0.5$) over a wide range of α (majority of objects have $\alpha > 1.0$). As discussed in section 2.8, the gamma distribution with $\alpha > 1.0$ peaks at $x = \beta(\alpha-1)$ instead of at $x = 0$. These types of objects ($\alpha > 1$, small β) must have a "humped" distribution with a very thin tail, indicating mostly light rainfall values. Large-sized objects tend to have larger values of β (over 75% of these objects have $\beta > 0.5$) and

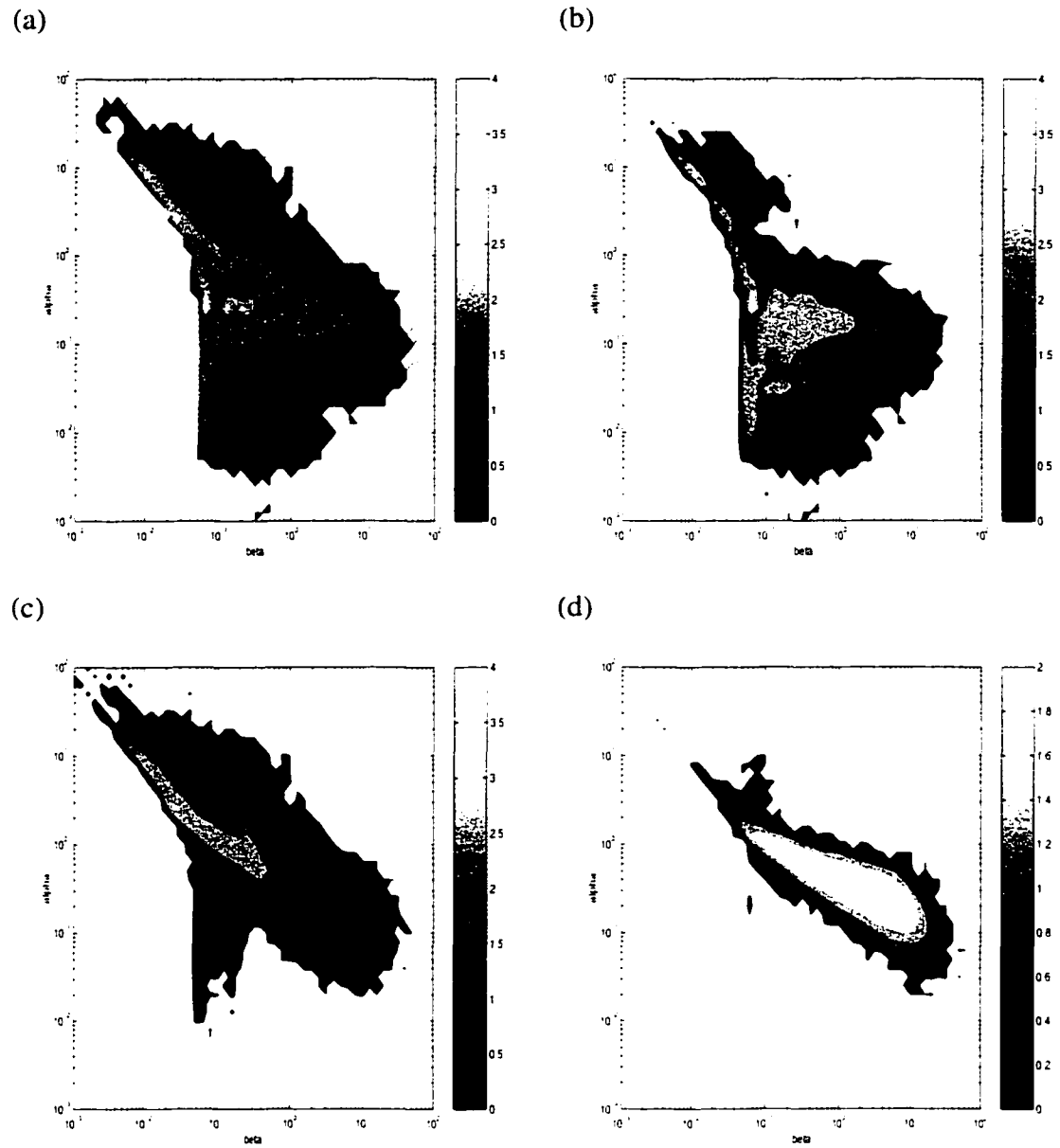


Figure 5-12: As in figure 5-11b, except for all objects (a), small-sized objects (b), medium-sized objects (c), and large objects (d). (note range of values in (d) are 2 orders of magnitude smaller than in other panels).

small values of α (over 90% of objects have $\alpha < 1.0$). One should expect objects of this size to have skewed histograms possibly with a thick tail, indicating the occurrence of heavier rainfall.

In order to determine which attributes to plot next, the correlation between attributes from the 2002 data will be examined. Focusing on the five attributes that are used in the

Table 5-3: Correlation matrix for 2002 data using [β, *a/b* 0.2, *a/b* 0.4, *a/b* 0.6, *a/b* 0.8] attributes.

	β	0.2 *a/b*	0.4 *a/b*	0.6 *a/b*	0.8 *a/b*
β	1.00	0.20	0.22	0.18	0.13
0.2 *a/b*	0.20	1.00	0.54	0.27	0.14
0.4 *a/b*	0.22	0.54	1.00	0.58	0.26
0.6 *a/b*	0.18	0.27	0.58	1.00	0.45
0.8 *a/b*	0.13	0.14	0.26	0.45	1.00

automated classification system, the correlation matrix was computed and is shown in table 5-3. As shown here, the various *a/b* attributes are somewhat correlated. The joint distribution of objects in a correlated space will not be very illuminating. One should expect objects in such a scatter plot to lie more or less along a line. In addition, *a/b* for the 0.4 contour is correlated with the other *a/b* attributes at a higher level than the other contour values. Therefore, it seems reasonable to assume that the information contained in all of the *a/b* attributes will be represented somewhat by the *a/b* 0.4 value alone. Table 5-3 shows that β is not highly correlated with any of the *a/b* attributes. Therefore, β and the 0.4 *a/b* attributes will be used to visualize the 2002 data next.

Figure 5-13 shows the density of objects distributed in this β, 0.4 *a/b* space. The density plot for objects of any size (figure 5-13a) shows that a great number of objects (over 75%) have *a/b* values equal to 1.0. For objects with *a/b* greater than 1.0, values of *a/b* in the 1-3 range occur much more often than larger values (> 5). Values of β for these objects are quite uniformly distributed between values of 0.01 and 10, although objects with *a/b* of 2 or less also appear to be more likely to have smaller (< 0.5) values of β. The density plot for small objects (figure 5-13b) appears to be in error at first glance, since the objects mostly lie along several constant values of *a/b*. However, this is in fact correct. For example, the minimum value that *b* (and *a*) can have is 0.0, corresponding to a correl-

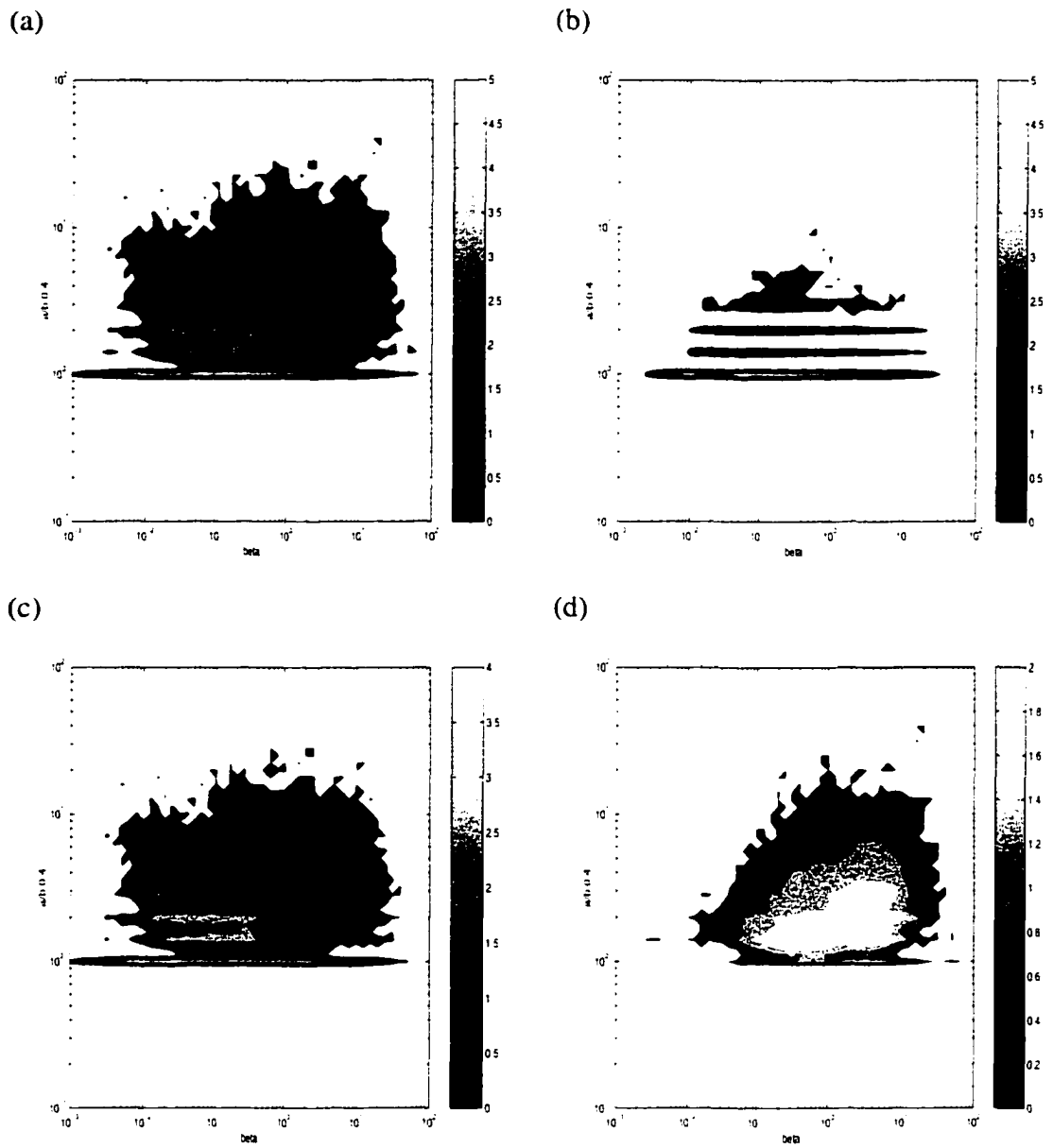


Figure 5-13: Density (log(number of objects) per grid box onto a regularly spaced grid in log(β), log(*a*/*b* 0.4 contour) space, consisting of 51×61 grid points) of 2002 rainfall objects. Plots of all objects (a), small-sized objects (b), medium-sized objects (c), and large objects (d). (note values for colorbar are in terms of log(number of objects and are different in each panel).

ogram with a correlation of 1.0 at the origin and values less than the contour level (0.4 in this case) for neighboring lags. Here, the 0.4 contour is assumed to pass through the origin, therefore $b = 0$. Since a/b is undefined for this situation, a value equal to 1.0 is assigned instead. The overwhelming majority (> 97%) of small objects have this charac-

teristic, as shown by the first high density region on figure 5-13b. In fact, over 85% of small objects have $a/b = 1$ and $\beta < 0.5$. Beyond this, the values of a and b (lengths of semi-major and semi-minor axes) are computed in terms of discrete grid boxes on the correlogram. For example, if the 0.4 correlation contour extends by one grid point in the x-direction and one grid point in the y-direction, then $a = \sqrt{2}$. If b in this instance is = 1, then $a/b = \sqrt{2}$. The second high density region up from the bottom of figure 5-13b represents objects with this characteristic. Medium-sized objects (figure 5-13c) also show relative maxima in population density at these discrete values of a/b, with a considerable fraction (~ 40%) of objects of this size having a/b attribute values = 1. The remaining medium-sized objects are distributed across a range of a/b from 1 to 10 and of β from 0.01 to 10. Large objects (figure 5-13d) also show relatively high density at discrete values of a/b, although not to the same degree as small and medium-sized objects. The majority of large objects are distributed across a range of a/b from 1 to 3 and of β from 0.1 to 10. It appears that objects in this size regime become more rare as with increasing values of a/b.

Now that a sample of the statistics related to the distribution of objects has been summarized, the next task involves selecting a random sample of these data to independently validate the automated classification system. This will be accomplished in the next section.

5.4.2 Validation sample

Again, the purpose of the previous section was to examine various statistics of the entire 2002 data prior to selecting a random sample of these data to validate the automated classification system. The task now becomes one of selecting a random sample of objects from the 2002 data. One could choose the sample from the entire data set, but since the

data set is dominated by objects of small size, one would expect the sample to be populated mostly by small-sized objects. As previously discussed, the attributes associated with these objects tend to be quite consistent, characterized by a/b=1 and very small values of α and β. A random sample of these consistent objects would be quite uninteresting. In addition, the statistics obtained from these objects are based upon a small number of pixels, therefore one would also expect that the statistical values to contain a relatively high level of uncertainty (for example, the standard error of the sample mean is proportional to $1/\sqrt{N}$). Medium-sized objects also show consistent characteristics, although to a lesser extent than the small objects. These objects are dominated by small β values and a large fraction have a/b = 1 as well. The target data set that the automated classification system was built upon consisted entirely of large objects (smallest object had just under 3250 pixels). Objects from the large-size class have attributes whose values vary across a wide range, including time of day, time of year, and location. Therefore, the validation sample will be taken entirely from the large size object regime.

A random sample of 100 objects was taken from the 31876 large objects. To confirm that this sample is representative of the entire population, the distributions of various attributes associated with these objects will be compared to the summary statistics provided in the previous section (figures 5-14 and 5-15). As shown in figures 5-14a and 5-14b, but for an anomalous spike in the early morning, the diurnal cycle of the validation sample is 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 (figures 5-14c, d) is also representative of the entire population, with relatively high frequency in the

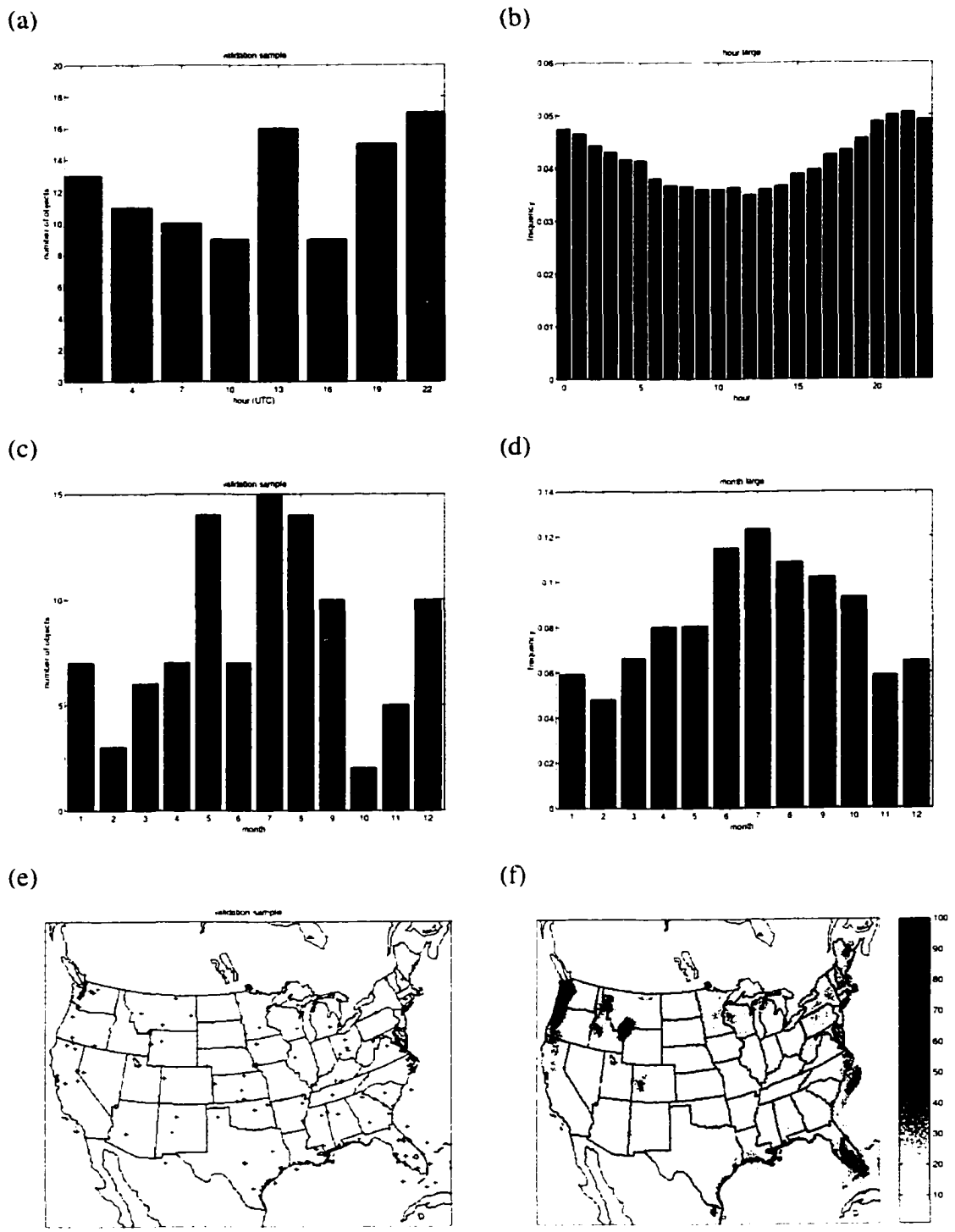


Figure 5-14: Comparison of characteristics of validation sample (left column) and large objects in 2002 data set (right column). As in figure 5-7, top row (a, b) shows distribution of objects as a function of time of day. As in figure 5-8, second row (c, d) displays distribution of objects by month. As in figure 5-9, last row (e, f) shows distribution of object center of mass locations.

warm season and low frequency in the cool season. The validation sample is also well distributed across the U.S. (figures 5-14e,f) 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 (figures 5-15a-d) also appear to be well representative of the entire 2002 population of

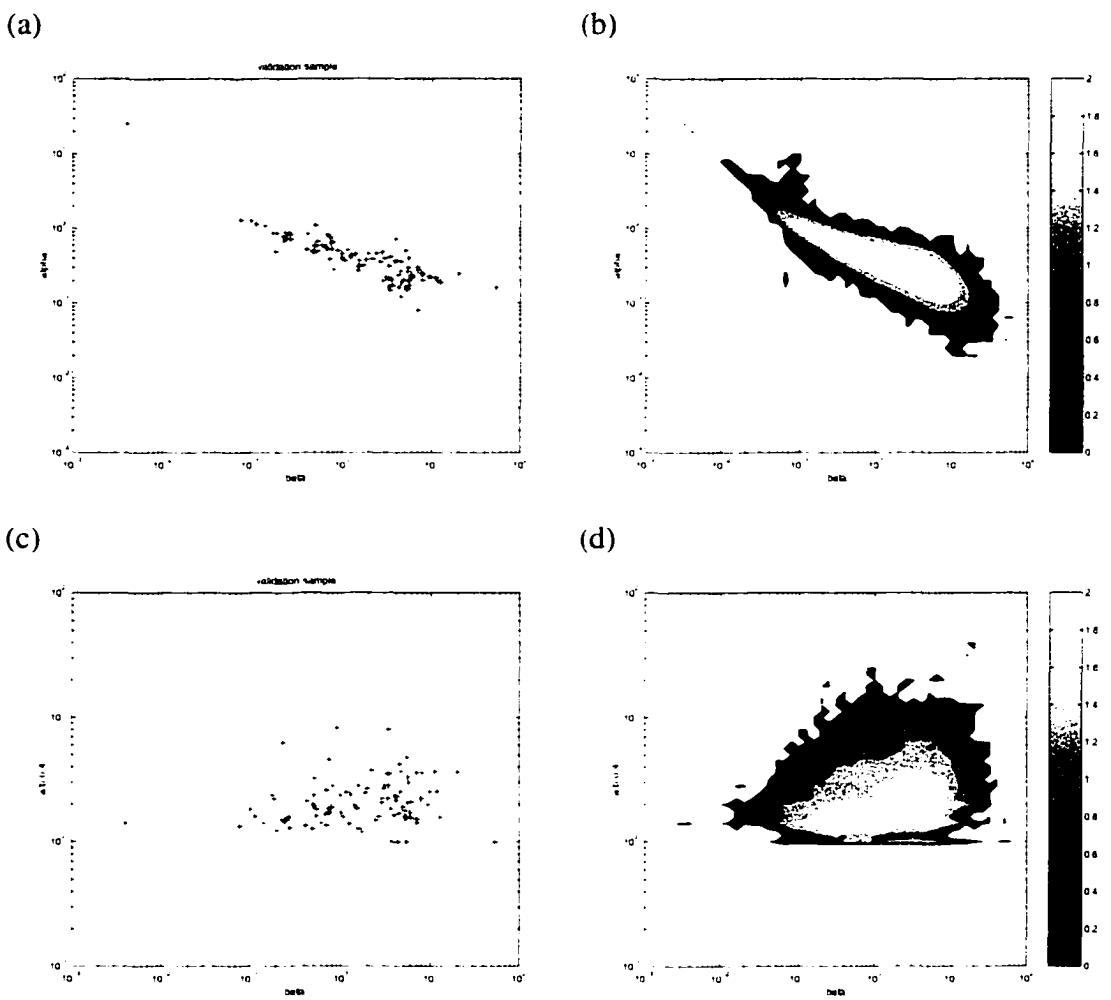


Figure 5-15: Comparison of characteristics of validation sample (left column) and large objects in 2002 data set (right column). As in figure 5-12, top row (a, b) shows object distribution density (log(number of objects)) in α, β plane. As in figure 5-13, second row (c, d) displays object distribution density (log(number of objects)) in β, 0.4 *a/b* plane.

large objects, with β values ranging from 0.1 to 10, α values ranging from 0.1 to 1, and *a/b* values ranging from 1 to 10. However, one object does appear to be an outlier, with a

very small β value, large α, and *a/b* slightly greater than 1. These results show that this sample is representative of the population and exhibits an interesting range of attribute values. This sample of 100 objects will be classified subjectively and objectively by the automated classification system. Comparison of these results will be presented in the next section.

5.4.3 Classification validation

In order to independently validate the automated classification system, a random, representative sample of 100 objects was pulled from the set of all large objects in the 2002 data. The large-sized objects show the greatest variability in attributes and should therefore pose the toughest classification challenge. In the previous section, the distributions of objects from the validation sample were compared with those from the entire set of large objects to confirm that this sample is representative of the population.

Each object from the sample was classified into five categories by the automated classification system (section 5.3). Class 1 is the stratiform class, classes 2 and 3 are cellular (class 2 is more of a cellular/hybrid class), and classes 4 and 5 are linear (class 5 is more of a linear/hybrid class). Figure 5-16a 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) shows that 46% of the objects were considered cellular by the automated system. Linear events were the most rare, combining classes 4 and 5 results 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 (figure 5-16b) further confirms that this sample is representative of the population (43% stratiform, 39% cellular, and 18% linear).

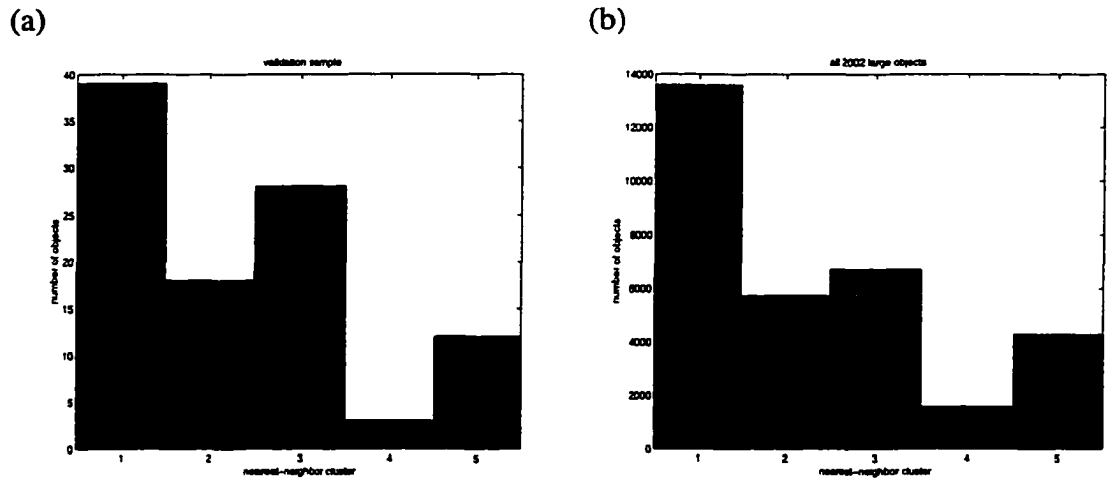


Figure 5-16: Distribution of objects by the automated classification system. Left panel (a) shows results from the validation sample, right panel (b) shows results for all large objects from 2002. Class 1 is the stratiform class, 2 is the cellular/hybrid class, 3 is the cellular class, 4 is the linear class, and 5 is the linear/hybrid class.

Each object from the validation sample was subjectively classified into three classes, stratiform, linear, and cellular. 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/non-convective classes (2-class case), and 85% of the objects were correctly classified in the 3-class case (stratiform, linear, cellular). To estimate the variability of these statistics, the validation results were resampled ("bootstrapping" Wilks 1995) with replacement 10000 times. The mean of the 2-class classification was 89.07% with a standard deviation of 3.13%. In this case, percent correct values varied from a minimum of 73% to a maximum of 99%. The mean of the 3-class classification was 85.06% with a standard deviation of 3.57%. Here, the percent correct values varied from a minimum of 68% to a maximum of 96%.

Figure 5-17a shows the distribution of objects color coded by their classification. The automated classification is indicated by the colored circle, the subjective classifica-

tion is the colored cross in the middle of each circle. One can visualize the incorrectly classified cases by the mismatched colors. Figure 5-17b shows the geographical distribution of cases in the validation sample. Different symbols are used to denote correctly and incorrectly classified cases. The incorrectly classified cases are scattered randomly across the contiguous U.S., indicating that the classification errors are independent of geographic location.

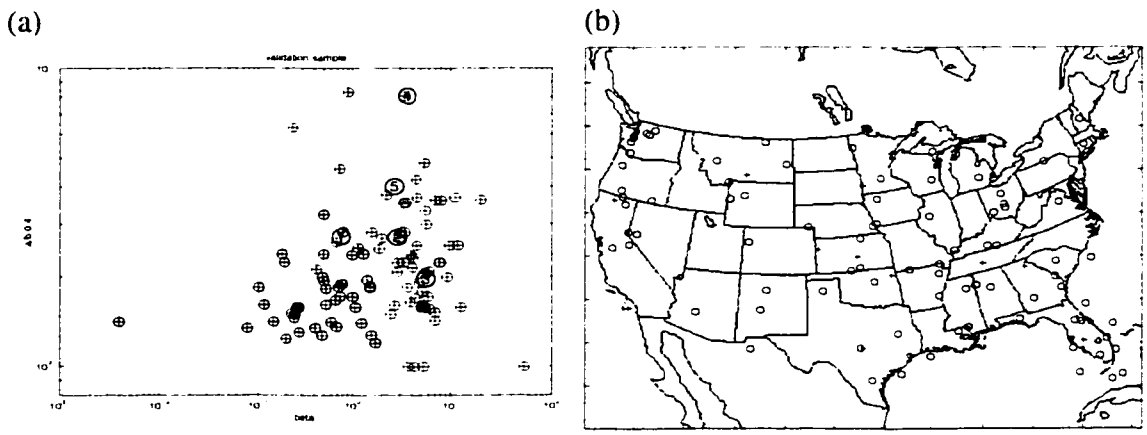


Figure 5-17: Left panel (a), scatter plot of validation sample in β, *a/b* 0.4 space. Right panel (b), geographic distribution of correctly (circles) and incorrectly (crosses) classified cases. In left panel (a), objects are color coded by their classification, blue for stratiform, red for cellular, and green for linear. Colored circles indicate the automated classification, colored crosses in the center of each circle indicate the subjective classification. Locations of the five cluster means used in the automated classification are indicated by their cluster numbers printed in black.

In figure 5-17a, most of the incorrectly classified cases were subjectively considered non-convective and classified as convective by the automated system. In fact, seven of the cases were classified as linear by the automated system and stratiform by the subjective classification. For these cases, high values of *a/b* tended to place them into a linear class even though the value of β 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 05 UTC 16 May 2002 (Obj_Five, figure 5-18). On figure 5-17, this

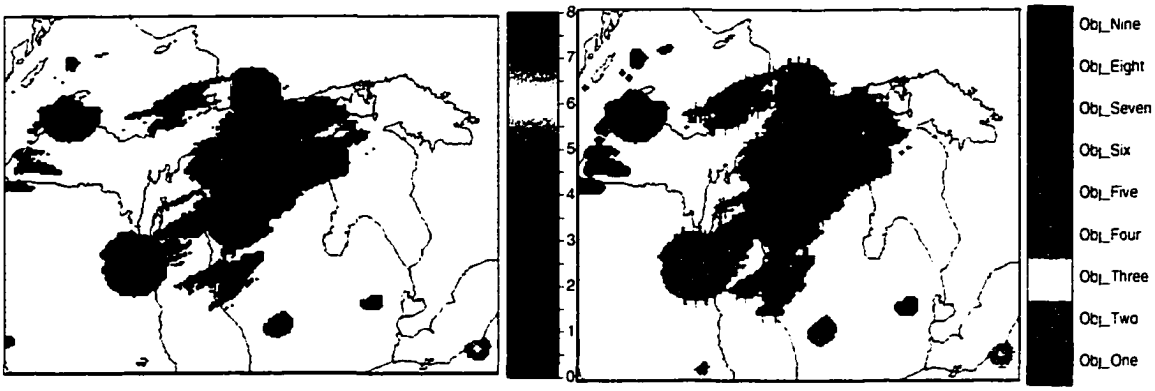


Figure 5-18: Object #3 from the validation sample. Left panel, 1h accumulated rainfall (mm) valid 05 UTC 16 May 2002. Right panel, result of object identification process. Object of interest is labelled as Obj_Five in right panel.

object is a green circle with blue cross located at $\beta = 0.7$ and $a/b = 4.6$. The rainfall associated with this object is generally light and widespread, likely leading to the subjective stratiform classification. At the same time, the rainfall is somewhat organized along a line, represented by the relatively high a/b values, such as the one for the 0.4 contour listed above. In terms of Euclidean distance to the five cluster means, this object was closest to the linear/hybrid cluster 5 due to the high values of a/b.

Further tuning of the automated classification to put greater weight on β or perhaps to first classify events into convective/non-convective classes based upon α, β and then further subdivide the convective cases into linear and cellular based on a/b may potentially improve the system. This sort of fine-tuning will be left for future work.

5.5 Summary

The overall goal of this work is to develop a completely automated rainfall pattern classification system. To accomplish this task, the discovery of a set of attributes that allow for automated characterization of rainfall patterns was required. This was accomplished via a relatively small target data set, comparing the results of various classification

experiments with a subjective classification. However, the target data set was a small sample of all possible rainfall objects that might occur in nature. The objects were located by hand, since at that point in the research an automated object identification system had not been developed. Since the automated classification was "trained" using the target data set, an independent validation of the system was required. In order to obtain a representative, random sample of the rainfall object population, analysis of the characteristics of a large data set had to be performed. To accomplish this, a completely automated rainfall object classification system was developed. The entire year of 2002 was analyzed using the automated rainfall object identification and analysis system. 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 data set was representative of the population. Once this had been confirmed, the sample was classified both subjectively and objectively via the automated classification system. Comparison of these results showed that the classification system accurately placed 85% of the objects into correct classes, and 89% of objects into their correct parent convective/non-convective class. Therefore, an independent confirmation of the accuracy of this classification system has been provided.

As a final step, each object in the 2002 data set regardless of size was run through the automated classification system, returning the distribution of objects into the five classes shown in figure 5-19. For objects of any size (figure 5-19a), the dominant class is stratiform, with over 90% of the cases from 2002 belonging to that class. Since the small objects (figure 5-19b) are almost unanimously classified as stratiform and represent over

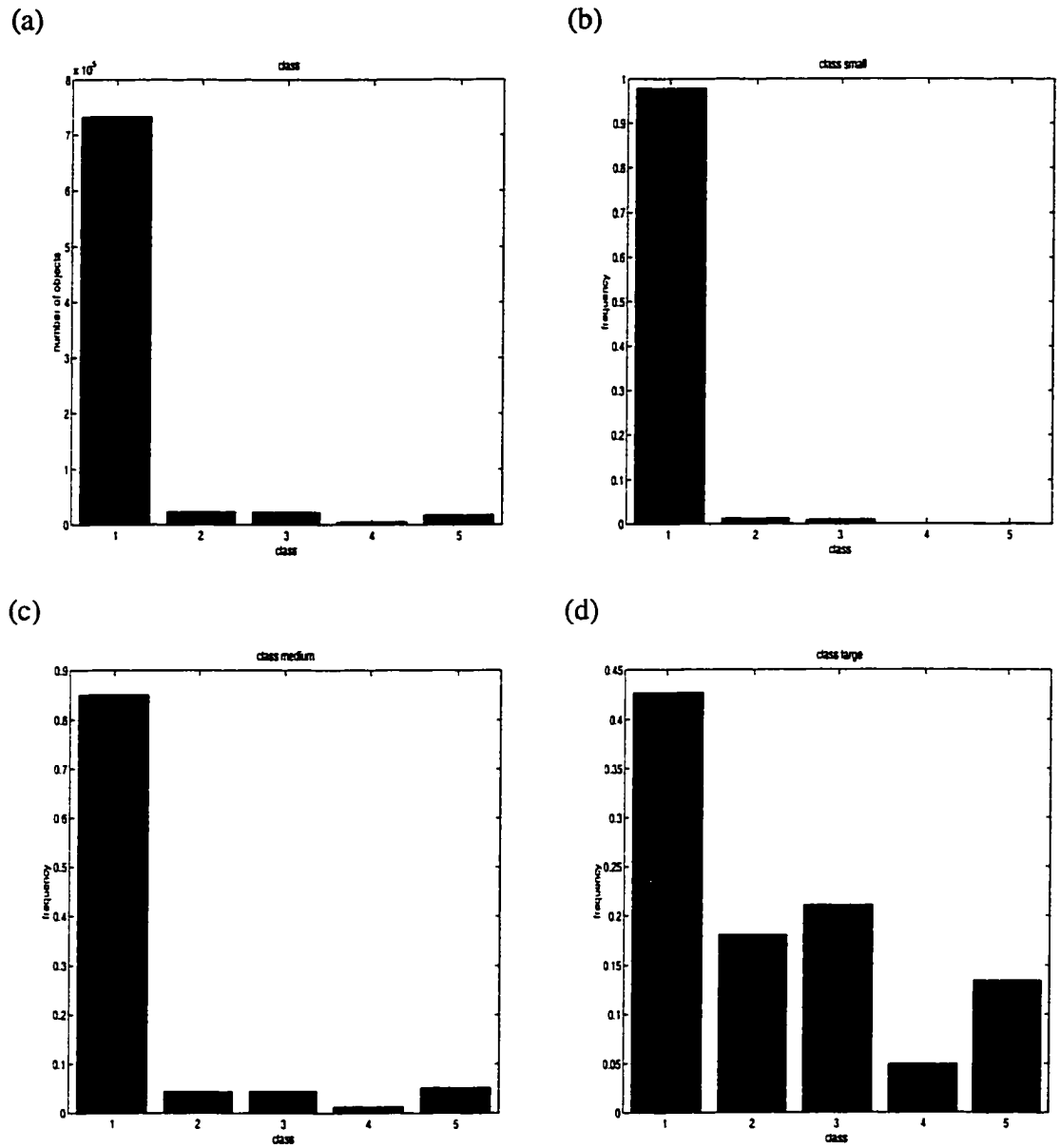


Figure 5-19: Results of automated classification of 2002 rainfall objects (1=stratiform, 2=cellular/hybrid, 3=cellular, 4=linear, 5=linear/hybrid). Top left panel (a) shows number of objects for all sizes, top right panel (b) shows relative frequency of small objects, lower left panel (c) shows relative frequency of medium-sized objects, and lower right panel (d) shows relative frequency of large objects.

65% of the entire data set, this result is not very surprising. As previously discussed, the small objects uniformly contain small values of β and a/b=1, making them closest to the stratiform cluster in Euclidean distance. Similarly, the dominant class for medium-sized objects is stratiform, again since a large fraction of these objects possess small values of β

and a/b=1. As discussed in the previous section, the classification of the large objects proves to be the most varied. The most popular class for these objects is also stratiform, followed closely by the cellular classification.

A summary of this research, concluding comments, and description of future work will be provided in the next chapter.

Chapter 6

Conclusions

6.1 Summary

The overall goal of this work was to develop a completely automated rainfall pattern classification system. To accomplish this task, the discovery of a set of attributes that allow for accurate characterization of rainfall patterns was required. This task was accomplished via a relatively small target data set, comparing the results of various classification experiments with a subjective classification. The first step in this process involved a "baseline" automated classification using the raw values of rainfall at each point in space as attributes. The next 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 distribution was selected as a compact model of the observed histograms. The parameters of the gamma distribution were fit to each histogram using the generalized method of moments technique. The automated classification algorithm using attributes produced by analysis of the observed histograms successfully separated the target data set into convective and non-convective classes. However, when the next level of classification hierarchy was considered, the classification experiments based upon these histogram-related attributes were not able to separate the linear and cellular events within the parent convective class. This did not come as a surprise since the attributes only contain information about the overall distribution of rainfall within the object. In order to provide information on aspects of the spatial continuity and variability within rainfall objects, it was determined that additional attributes

related to the shape and structure of the spatial patterns were needed.

Information regarding the degree of spatial organization of the rainfall systems was obtained via geostatistical measures. The correlogram or auto-correlation function was selected for analysis because it did not depend on the magnitude of the rainfall values. By fitting ellipses to various correlation levels in the correlogram, useful information on the degree of organization within each rainfall system was obtained. The automated classification algorithm using attributes produced by summary measures of the geostatistical properties of rainfall patterns successfully separated the target data set into linear, cellular, and stratiform classes. Therefore, it was concluded that a useful set of attributes for classification had been obtained.

The target data set was a small sample of all possible rainfall objects that might occur in nature. The objects were located by hand, since an automated object identification system had not yet been developed. Since the automated classification was "trained" using the target data set, an independent validation of the system was required. In order to obtain a representative, random sample of the rainfall object population, analysis of the characteristics of a large data set were performed. A completely automated rainfall object classification system was developed to accomplished this task. Rainfall objects (or systems) were simply defined as contiguous regions of measurable precipitation. The classification system was based upon a nearest-neighbor approach, using the best results from the previous classification experiments using the target data set. A large data set covering the entire year of 2002 was then analyzed using the automated rainfall object identification and analysis system. Summary statistics of attributes from this year were examined, and a random sample of interesting objects was pulled from the 2002 data. The distribu-

tion of the random sample was compared with the summary statistics in order to confirm that this validation data set was representative of the population. Once this had been confirmed, the sample was classified both subjectively and objectively via the automated classification system. Comparison of these results showed that the classification system accurately placed 85% of the objects into correct classes, and 89% of objects into their correct parent convective/non-convective class. Therefore, an independent confirmation of the accuracy of this classification system was provided. Finally, the complete set of rainfall systems for the year of 2002 was classified.

6.2 Conclusions

The goal of developing a general, completely automated procedure for classifying rainfall systems has been met. A desirable property of the technique is that any rainfall system can be classified regardless of size, location, time of day or year, degree of organization, etc. The process of knowledge discovery in databases was followed to develop a relatively straightforward and unique classification system using statistically-based attributes. To ensure that the method performed well, results of this technique were validated against a subjective classification based upon objective criteria. From an independent random sample of interesting cases, the automated classification system accurately placed events into stratiform, linear, and cellular classes 85% of the time. The classification will be applied to forecast fields from research NWP models in the near future as part of an object-oriented verification system. Other applications, such as climatological studies, ensemble forecast diagnosis, and weather-related decision support systems, may also benefit from the use of this system.

6.3 Future work

While the automated rainfall system classification procedure developed in this work produces satisfactory results, further refinement of the methods used 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. These methods utilize texture-related attributes; the inclusion of these sorts of attributes in the classification system while keeping the current object identification system might also lead to more accurate classification. The inclusion of other sources of rainfall-related data, such as lightning, reflectivity, VILS, satellite radiances, etc., may also help to improve the classification.

Further refinements in the classification hierarchy are also desirable. For example, the degree to which the attributes used in this work will dissect 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 sub-classes, then additional attributes that do have this ability should be discovered.

There are many potential applications for the automated rainfall system classification procedure developed in this work. Forecast verification and predictability studies

may also benefit from such a classification system. As mentioned in chapter 1, this was the primary motivation for this work. For example, Anthes (1983) argued for expanding verification information to include the validation of the "realism" of a forecast. One specific method that Anthes (1983) suggested was to verify the characteristics of significant meteorological phenomena. Along these lines, several "object-oriented" or "feature-specific" approaches to verification have been attempted or proposed (Somerville 1977; Williamson 1981; Neilley 1993; Smith and Mullen 1993; Weygandt and Seaman 1994; Baldwin et al. 2001). In order to accomplish the task of verifying significant meteorological phenomena, an automated system for identifying, characterizing, and classifying such phenomena is required. Rainfall systems are certainly an excellent candidate for this type of verification technique. For example, information on errors of displacement, amplitude, orientation, mode, from numerical guidance related to specific classes of MCSs, for example, would be quite useful for operational forecasters, such as those at the Storm Prediction Center (Greg Dial 2003, personal communication).

Climatological studies, similar to those undertaken by Bluestein and Jain (1985), Houze et al. (1990), Geerts (1998), and Parker and Johnson (2000) would benefit greatly from the use of an automated rainfall system classification procedure. A much larger and more comprehensive database of events could be obtained. Since a multi-year 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 (Perica and Foufoula-Georgiou 1996). Severe

weather reports could also be associated with the various classes of co-located rainfall systems, possibly leading to improved forecasts of hazardous weather.

An automated rainfall system classification may also address interesting issues related to the predictability of smaller-scale rainfall features. Past research involving the use of band-pass, Fourier, or wavelet analysis techniques have applied the filtering properties of these methods to select certain spatial scales within the fields for subsequent verification (e.g. Stamus et al. 1992; Briggs and Levine 1997). As a result, smaller-scale features were discarded as "noise" and only the larger-scale "signal" was verified. Consequently, various measures of forecast skill showed that the smoothed fields verified better than those that contained smaller-scale features, which are considered "unpredictable". Unfortunately, what may be categorized as "noise" might actually be interesting, realistic, and potentially valuable smaller-scale detail. Removing this "noise" might be akin to throwing out the baby with the bath water, so to speak. This idea of filtering "unpredictable" scales has been touted as a primary benefit of ensemble forecasting. For example, Hamill and Colucci (1997) claimed that the mean of an ensemble of reduced-resolution model forecasts provided better forecasts than a single higher-resolution model. Similarly, Germann and Zawadzki (2002) show that if rainfall forecasts are filtered, the predictability limit for such forecasts is extended.

Perhaps the traditional definition of "predictability" should be modified. Typically, phenomena are considered "predictable" as long as errors associated with their prediction are smaller than the length or time scales associated with the lifetime of the phenomena. Predictability is often measured by the correlation between predicted and observed variables (e.g., Zawadzki et al. 1994; Germann and Zawadzki 2002). The point during the

forecast when the correlation drops below some threshold (typically $1/e$) is considered the predictability time scale. At this point, phase and displacement errors (in time and/or space) are thought to be as large as the scales of the phenomena. When considering larger-scale phenomena, such as planetary waves, this definition is sensible. However, when considering smaller-scale events, such as tornadoes for example, this definition may not be appropriate. By this definition, a tornadic feature will only be considered predictable if it can be forecasted accurately with timing errors less than the lifetime of a tornado (typically on the order of 10min) and less than the length scale of a tornado (typically on the order of 100m). This is certainly not a very useful definition of predictability for tornado forecasting, as most emergency managers, weather forecasters, etc. would consider forecasts of these systems to be extremely valuable even with timing and distance errors much larger than the lifetime and length scales of a typical tornado. A more appropriate definition for predictability may be obtained through the use of a classification methodology. For rare events in particular, if the occurrence/non-occurrence of certain classes of events are predicted within a certain degree of accuracy, even with considerable errors in timing and displacement, those events should be considered predictable. An automated system for locating and classifying such events could be used to confirm that the occurrence of such events is accurately predicted, with information provided on typical errors in displacement, amplitude, orientation, etc.

In addition, recent research into ensemble forecasting techniques using cloud-resolving models (Elmore et al. 2002) has demonstrated the benefits of such a technique. The automated rainfall system classification system will help in analyzing a large set of high-resolution forecasts, providing meaningful information on the range of possible rainfall

systems that are predicted by the ensemble members.

Finally, an automated rainfall classification system might useful as part of a weather-related decision support system (e.g., Peak and Tag 1994, Brody et al. 1997). These expert systems assimilate large volumes of data and return some form of interpretation of the data, thereby speeding up the data analysis process so the human decision maker can concentrate on the important task is at hand.

References

Adriaans, P. and D. Zantinge, 1996: *Data Mining*, Addison-Wesley, 158 pp.

Alhamed, A., 2000: *Clustering methodologies applied to short-term ensemble forecasting: an exercise in data mining.* Univ. of Oklahoma, PhD dissertation, 298pp.

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

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

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.

Austin, P. M., 1987: Relation between measured radar reflectivity and surface rainfall. *Mon. Wea. Rev.*, **115**, 1053–1070.

Bak, P., C. Tang, and K. Wiesenfeld, 1987: Self-organized criticality: An explanation of 1/ f noise. *Phys. Rev. Lett.*, **59**, 381-384.

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, 2nd Symposium on Integrated Observing Systems, 78th AMS Annual Meeting, January 11-16, 1998, Phoenix, Arizona, 10-11.

Baldwin, M. E., S. Lakshmivarahan, and J. S. Kain, 2001: Verification of mesoscale features in NWP models. Preprints, *9th Conf. on Mesoscale Processes*, Ft. Lauderdale, FL, Amer. Meteor. Soc., 255-258.

Baldwin, M. E., and S. Lakshmivarahan, 2002: Rainfall classification using histogram analysis: An example of data mining in meteorology. *Intelligent Engineering Systems Through Artificial Neural Networks*, Volume 12, C.H. Dagli, A.L. Buczak, J. Ghosh, M.J. Embrechts, O. Ersoy, S.W. Kercel, Eds., ASME Press, 429-434.

Baldwin, M. E., and S. Lakshmivarahan, 2003: Development of an events-oriented verification system using data mining and image processing algorithms. Preprints, *3rd Conf. on Artificial Intelligence*, Long Beach, CA, Amer. Meteor. Soc., paper 4.6.

Benjamin, S. G, K. J. Brundage, P. A. Miller, T. L. Smith, G A. Grell, 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.

Bluestein, H. B., 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.

Boas, M. L, 1983: *Mathematical methods in the physical sciences.* John Wiley and Sons, 793pp.

Briggs, W. M., and R. A. Levine, 1997: Wavelets and field forecast verification. *Mon. Wea. Rev.*, **125**, 1329–1341.

Brody, F. C., R. A. Lafosse, D. G Bellue, and T. D. Oram, 1997: Operations of the National Weather Service Spaceflight Meteorology Group. *Wea. Forecasting*, **12**, 526–544.

Burpee, R. W., 1989: A summer day without significant rainfall in South Florida. *Mon. Wea. Rev.*, **117**, 680-687.

Byers, H.R., and R. R. Braham, Jr., 1949. *The Thunderstorm.* U.S. Government Printing Office, Washington, D.C., 187pp.

Davies, E. R., 1997: *Machine vision: Theory, algorithms, practicalities.* Academic Press, 750pp.

Deutsch, C. V. and A. G Journel, 1998: *GSLIB: Geostatistical software library and user's guide. Second edition.* Oxford University Press, 369pp.

Dixon, M., and G Wiener, 1993: TITAN: Thunderstorm Identification, Tracking, Analysis, and Nowcasting—a radar-based methodology. *J. Atmos. Oceanic Technol.*, **10**,

785–797.

Ebert, E., 1987: A pattern recognition technique for distinguishing surface and cloud types in polar regions. *J. Climate Appl. Meteor.*, **26**, 1412–1427.

Eckert, P., D. Cattani, and J. Ambuhl, 1996: Classification of ensemble forecasts by means of an artificial neural network. *Meteor. Appl.*, **3**, 169–178.

Edwards, R., S. F. Corfidi, R. L. Thompson, J. S. Evans, J. P. Craven, J. P. Racy, D. W. McCarthy, and M. D. Vescio, 2002: Storm Prediction Center forecasting issues related to the 3 May 1999 tornado outbreak. *Wea. Forecasting.*, **17**, 544–558.

Elmore, K. L., D. J. Stensrud, and K. C. Crawford, 2002: Ensemble cloud model applications to forecasting thunderstorms. *J. Appl. Meteor.*, **41**, 363–381.

Fayyad, U. M., G. Piatetsky-Shapiro, and P. Smyth, 1996: From data mining to knowledge discovery: An overview. *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. MIT Press. 1-34.

Fine, S. S. and A. B. Fraser, 1990: A simple technique for multiple-parameter pattern recognition with an example of locating fronts in model output. *J. Atm. and Ocean. Tech.*, **7**, 896–908.

Freund, J. E. and R. E. Walpole, 1987: *Mathematical Statistics*, Fourth edition, Prentice Hall, Englewood Cliffs, NJ, 608pp.

Fritsch, J. M., R. J. Kane, and C. R. Chelius, 1986: The contribution of mesoscale convective weather systems to the warm-season precipitation of the United States. *J. Climate Appl. Meteor.*, **25**, 1333–1345.

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.

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

Germann, U., and J. Joss, 2001: Variograms of radar reflectivity to describe the spatial continuity of Alpine precipitation. *J. Appl. Meteor.*, **40**, 1042-1059.

Germann, U., and I. Zawadzki, 2002: Scale-dependence of the predictability of precipitation from continental radar images. Part I: Description of the methodology. *Mon. Wea. Rev.*, **130**, 2859–2873.

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. Clim.*, **8**, 897–931.

Hamill, T. M., and S. J. Colucci, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312–1327.

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

Harris, D., E. Foufoula-Georgiou, K. K. Droegemeier and J. J. Levit, 2001: Mutiscale statistical properties of a high-resolution precipitation forecast. *J. Hydrometeor.*, **2**, 406-418.

Hobbs, P. V., 1978: Organization and structure of clouds and precipitation on the mesoscale and microscale in cyclonic storms. *Rev. Geophys. Space Phys.*, **16**, 741–755.

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., Jr., 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, 561pp.

Jain, A. J., and R. C. Dubes, 1988: *Algorithms for Clustering Data.* Prentice-Hall, 320 pp.

Johns, R. H., and W. D. Hirt, 1987: Derechos: Widespread convectively induced windstorms. *Wea. Forecasting*, **2**, 32–49.

Johns, R. H., and C. A. Doswell III, 1992: Severe local storms forecasting. *Wea. Forecasting*, **7**, 588–612.

Johnson, J. T., P. L. MacKeen, A. E. Witt, E. D. Mitchell, G. J. Stumpf, M. D. Eilts, and K. W. Thomas, 1998: The storm cell identification and tracking algorithm: An enhanced WSR-88D algorithm. *Wea. Forecasting.*, **13**, 263–276.

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

Kendall, M. G., and A. Stuart, 1977: *The Advanced Theory of Statistics.* Vol. 2, *Inference and Relationship.* Griffin, 748pp.

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

Kessler, E., III, and J. A. Russo Jr., 1963: Statistical properties of weather radar echoes. Preprints, Proc. of 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, 397pp.

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

Lakshmanan, V., 2003: Motion estimator based on hierarchical clusters. *Preprints, 19th IIPS Conference*, Amer. Meteor. Soc., Long Beach, CA., paper 15.14.

Lovejoy, S., 1982: Area-perimeter relation for rain and cloud areas. *Science*, **216**, 185-187.

Marr D. and E. Hildreth, 1980: Theory of edge detection. *Proc. R. Soc. Lond.*, **B207**, 187-217.

Mohr, K. I., and E. J. Zipser, 1996: Defining mesoscale convective systems by their 85-Ghz ice-scattering signatures. *Bull. Amer. Meteor. Soc.*, **77**, 1179–1189.

Mohr, K. I., Famiglietti, J. S., and E. J. Zipser, 1999: The contribution to tropical rainfall with respect to convective system type, size, and intensity estimated from the 85-GHz ice-scattering signature. *J. Appl. Meteor.*, **38**, 596–606.

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

Neilley, P. P., 1993: Evaluating forecasts using an object-oriented approach. *Preprints, 13th Conference on Weather Analysis and Forecasting*, Vienna, VA, 298-300.

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

NOAA, 2002: Climate Atlas of the United States, National Climatic Data Center, Federal Building, 151 Patton Ave., Asheville, NC, 28801-5001. [available online at http:// www.nndc.noaa.gov/cgi-bin/climaps/climaps.pl]

NOAA, 2003: Climate of 2002 - Annual review, National Climatic Data Center, Federal Building, 151 Patton Ave., Asheville, NC, 28801-5001. [available online at http:// www.ncdc.noaa.gov/oa/climate/research/2002/ann/us-summary.html]

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.

Pearson, Frederick, II, 1990: *Map projections: Theory and applications.* CRC Press, 372pp.

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.

Rao, T. N., D. N. Rao, K. Mohan, and S. Raghavan, 2001: Classification of tropical precipitating systems and associated Z–R relationships. *J. Geophys. Res.*, **106D**, 17699–17711.

Richman, M. B., 1986: Rotation of principal components. *J. Climatol.*, **6**, 293-335.

Roebber, P. J., and L. F. Bosart, 1996: The complex relationship between forecast skill and forecast value: A real-world analysis. *Wea. Forecasting*, **11**, 544–559.

Rogers, E., M. Ek, Y. Lin. K. Mitchell, D. Parrish, and G. DiMego, 2001: Changes to the NCEP Meso Eta Analysis and Forecast System: Assimilation of observed precipitation, upgrades to land-surface physics, modified 3DVAR analysis. NWS Technical Procedures Bulletin No.479. [Available at http://www.emc.ncep.noaa.gov/mmb/ mmbpll/spring2001/tpb/ or Office of Meteorology, 1325 East-West Highway, Silver Spring, MD 20910].

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.

Schaefer, J. T., and R. L. Livingston, 1988: The typical structure of tornado proximity soundings. *J. Geophys. Res.*, **93-D5**, 5351-5364.

Smith, B. B. and S. L. Mullen, 1993: An Evaluation of Sea Level Cyclone Forecasts Produced by NMC's Nested-Grid Model and Global Spectral Model. *Wea. Forecasting*, **8**, 37–56.

Smith, J.A., and W.F. Krajewski, 1991: Estimation of mean field bias of radar rainfall estimates. J. *Appl. Meteor.*, **30**, 397-412.

Somerville, R. C. J., 1977: Pattern recognition techniques for forecast verification. *Contributions to Atmospheric Physics*, **50**, 403-410.

Song, W., W. Fan, and B. Wang, 2002: Influences of finite-size effects on the self-organized criticality of forest-fire model. *Chinese Science Bulletin*, **47**, 177-180.

Stackpole, J. D., 1994: *The WMO format for the storage of weather product information and the exchange of weather product messages in gridded binary form.* NMC/NCEP Office Note 388, 71 pp.

Stamus, P. A., F. H. Carr, and D. P. Baumhefner, 1992: Application of a scale-separation verification technique to regional forecast models, *Mon. Wea. Rev.*, **120**, 149-163.

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.

Stumpf, G. J., A. Witt, E. D. Mitchell, P. L. Spencer, J. T. Johnson, M. D. Eilts, K. W. Thomas, and D. W. Burgess, 1998: The National Severe Storms Laboratory Mesocyclone Detection Algorithm for the WSR-88D. *Wea. Forecasting*, **13**, 304–326.

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

Weckwerth, T. M., J. W. Wilson, R. M. Wakimoto, and N. A. Cook, 1997: Horizontal convective rolls: Determining the environmental conditions supporting their existence

and characteristics. *Mon. Wea. Rev.*, **125**, 505–526.

Weygandt, S. S. and N. L. Seaman, 1994: Quantification of predictive skill for mesoscale and synoptic-scale meteorological features as a function of horizontal grid resolution. *Mon. Wea. Rev.*, **122**, 57-71.

Wilks, D. S., 1989: Rainfall intensity, the Weibull distribution, and estimation of daily surface runoff. *J. Appl. Meteor.*, **28**, 52-58.

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

Wilks, D. S., 1995. Statistical Methods in the Atmospheric Sciences. Academic Press, 467 pp.

Williamson, D. L., 1981: Storm track representation and verification. *Tellus*, **33**, 513-530.

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.

Wilson, J. W., and E. A. Brandes, 1979: Radar measurement of rainfall—A summary. *Bull. Amer. Meteor. Soc.*, **60**, 1048–1058.

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.

Zawadzki, I., J. Morneau, and R. Laprise, 1994: Predictability of precipitation patterns: An operational approach. *J. Appl. Meteor.*, **33**, 1562–1571.

Zepeda-Arce, J., E. Foufoula-Georgiou, and K. K. Droegemeier, 2000: Space-time rainfall organization and its role in validating quantitative precipitation forecasts. *J. Geophys. Res.*, **105**, 10129-10146.

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.

Zipser, E. J., 1982: Use of a conceptual model of the life cycle of mesoscale convective

systems to improve very-short-range forecasts. *Nowcasting*, K. Browning, Ed., Academic Press, 191–204.

Zupanski, D., and F. Mesinger, 1995: Four-dimensional variational assimilation of precipitation data. *Mon. Wea. Rev.*, **123**, 1112–1127.