UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

ON THE LIMITATIONS OF DISCRIMINATING OUTBREAKS OF SEVERE CONVECTION

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

CHAD SHAFER
Norman, Oklahoma
2010

ON THE LIMITATIONS OF DISCRIMINATING OUTBREAKS OF SEVERE CONVECTION

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF METEOROLOGY

BY

____
Dr. Lance Leslie, Co-Chair

____
Dr. Michael Richman, Co-Chair

____
Dr. Chuck Doswell

____
Dr. Alan Shapiro

____
Dr. David Stensrud

____
Dr. Theodore Trafalis

Dedication

This paper is dedicated to Maxine Laird, Jim Simpson, Phil Williamson, and Rachel Green. Their support was immeasurable, and I miss them dearly.

Acknowledgments

I wish to thank my advisors for giving me such a wonderful opportunity to do research on a topic that will fascinate me for the rest of my days. Lance Leslie, Mike Richman, and Chuck Doswell have provided exceptionally valuable guidance, support, and encouragement for this research and for being a research meteorologist in general. I hope to continue research in this area with the three of them in the future. Andrew Mercer has also provided substantial support and was instrumental in teaching me many of the statistical techniques used in Section 2.

I also would like to thank David Stensrud, Alan Shapiro, and Theodore Trafalis (members of the Ph.D. committee) for their support. Peer reviewers for much of the work presented herein include Valliappa Lakshmanan, Harold Brooks, Rich Thompson, Roger Edwards, Kim Elmore, David Stensrud, Michael Coniglio, Paul Roebber, and Corey Mead. All of their suggestions have proven instrumental in improvements in manuscripts and the work presented here. Several forecasters from the Storm Prediction Center provided excellent suggestions, particularly with the areal coverage objective technique. John Hart and Jared Guyer are acknowledged for SvrPlot, which is used for all of the storm reports figures in this paper. Additionally, John Hart is acknowledged for the development of the “middle-50% parameter”, a focus for much of the research presented in Shafer and Doswell (2010a) and Section 3 herein.

I thank Henry Neeman, Joshua Alexander, Brett Zimmerman, and the University of Oklahoma Supercomputing Center for Education & Research (OSCER) for all of their help when something went wrong computationally. All of the research shown herein was produced on OSCER resources, including the machines named “topdawg”, “boomer”, and “sooner”. Kevin Thomas also provided computing assistance on numerous occasions.

Several former and current graduate students at the University of Oklahoma provided suggestions, moral support, and much-needed distractions. These include but are not limited to Kevin Goebbert, Patrick Marsh, Bryan Burkholder, Alex Schenkman, Matt Kumjian, Charlotte Wainwright, Somer Erickson, Ryan Sobash, Amanda Kis, Kim Klockow, Angelyn Kolodziej, Stephanie Weiss, Pat Hyland, David Sherman, Darrel Kingfield, Sean Arms, Jose Galvez, Owen Shieh, Nick Engerer, Tim Wagner, Robin Tanamachi, Dan Dawson, Nate Snook, Virginia Silvis, Jeremy Gibbs, and Kiel Ortega. A big thank-you goes to Celia Jones, Marcia Pallutto, Judy Henry, Becky Steely, and Nancy Campbell for their unbelievable capability to keep me calm when disaster strikes. These secretaries for the School of Meteorology deserve recognition on *all* of the research conducted in the department. Additionally, Mark Laufersweiler provided computational help on several occasions, and an ear to bend on many more.

None of this research would have been possible without the eternal support of my family and friends. Friends deserving special recognition include Aaron Anckner, Chet McGowen, Mehdi Tajouri, Jenny Becic, Kyna Lang, Amber Bailey, Tony Westenhofer, Mike Grogan, Valery Dagostaro, Wilson Shaffer, Michael Schenk, Jerry Gorline, Arthur Taylor, Kristine Poptanich, Paul Dallavalle, Justin Jensen, Liz Sinclair, Andrya Pierce, and Toran Erickson. My brother and sister, Doug Rau and Carrie Dow, were kind enough to read my first journal articles without complaint. My grandfather, Roland Shafer, is the true source of my love for meteorology. My grandmother, Kay Shafer, supports me even though she hates severe weather. My father, Mike Shafer, still makes sure we talk about the weather first. And my mother, Donna Shafer, is my source of inspiration – thank you for everything.

Table of Contents

Page

Page

List of Tables

Page

List of Figures

Page

Page

Abstract

Although much has been learned in recent decades regarding the synoptic and subsynoptic environments associated with major tornado outbreaks, very little research has been conducted using a large sample of outbreak cases to distinguish major tornado outbreaks from less significant events. Preliminary investigations suggested that there were systematic differences between the synoptic-scale environments of tornado outbreaks and primarily nontornadic outbreaks, and that synoptic-scale processes occurring up to three days in advance of the outbreaks were a useful means of discriminating the events. Subsequent investigations showed that the time of year in which the outbreaks occurred played a non-negligible role in the discrimination of these events, but that precursor synoptic-scale environments could be used as initial data in mesoscale models to simulate outbreak environments that were systematically different between the two types of outbreaks.

Because of the promising results of these initial studies, research encompassing all types of outbreaks should be investigated, as most outbreaks of severe convection cannot be classified readily as major tornado outbreaks or primarily nontornadic outbreaks. This is the focus of the present study. For this work to be implemented, a new scheme to rank and classify severe weather outbreaks of any type must be developed. Initial work implemented a multivariate, linear-weighted index to rank the events based on the characteristics of the severe reports during individual 24-h periods. Although the technique was relatively effective in the identification of the most significant events, days with multiple regionally distinct clusters of severe reports were considered as single events. A modification to this ranking scheme will be introduced in this study to account for these days, using a kernel density estimation technique. Two objective techniques (the areal coverage and principal component analysis methods) will be introduced and their ability to discriminate major severe weather outbreaks from less significant events will be compared. Finally, comparison of these objective techniques to current operational forecasting will be examined to determine the potential utility of these methods in future forecasts.

The results of this work suggest that there is reasonable accuracy and skill in discriminating major tornado outbreaks from less significant events. However, a substantial false alarm problem exists with any of the objective techniques developed, which is also observed in current operational forecasts. Implications of these findings and potential topics for future research will be discussed.

1. Introduction

Finding links between synoptic-scale processes and tornado outbreaks has been a focus of severe weather research for several decades. This is true despite a relatively consistent upper-air network of radiosondes, surface observational networks, and continuously improving computer technology. Improvements of computing have allowed for the development of mesoscale models used to predict severe convection in numerous studies (e.g., Cortinas and Stensrud 1995; Stensrud et al. 1997; Stensrud et al. 2000; Stensrud 2001; Kain et al. 2006; Deng and Stauffer 2006). Although our understanding of synoptic-scale links to tornado outbreaks certainly has improved because of these advances, the specific synoptic-scale processes involved and the degree to which they influence the characteristics of a severe weather outbreak remain unclear (Johns and Doswell 1992; Doswell et al. 1993; Johns and Hart 1993; Doswell and Bosart 2001; Shafer et al. 2009 – hereafter, S09).

The need for increased understanding of the synoptic-scale processes regulating the occurrence of tornado outbreaks versus severe weather outbreaks with a relatively small number of significant tornadoes is undeniable. Tornado outbreaks pose a substantial risk to life and property, generally much greater than outbreaks of severe weather without many damaging tornadoes. Consistent discrimination of tornado outbreaks from other types of outbreaks also would allow the National Weather Service and emergency management personnel to prepare appropriately for warning operations, public awareness, and relief efforts. Because many outbreaks clearly are associated with strong synoptic-scale systems (termed "synoptically evident" outbreaks by Doswell et al.

1993), research identifying synoptic-scale processes associated with these outbreaks is essential to our understanding of these events.

Early research on the prediction of severe convection investigated typical severe-weather environments. Fawbush and Miller (1952) identified the “loaded gun” sounding. Fawbush and Miller (1954) described three characteristic types of soundings associated with severe weather. Beebe (1955) called attention to the “inverted V” sounding. Research soon expanded to identify prototypical synoptic-scale map types associated with severe weather events. Beebe (1956) developed local composites of rawinsonde data in proximity to tornado reports. Increased understanding of severe weather environments and the synoptic-scale patterns associated with them led to the identification of five map types commonly associated with regional tornado outbreaks (Miller 1972). See Schaefer (1986) and Doswell (2007a) for a review of early severe weather research.

Recent research has focused on an ingredients-based methodology (Doswell et al. 1996). Several severe weather parameters have been developed to assess (in a diagnostic and/or prognostic sense) the potential for severe weather, the mode of severe convection (e.g., discrete cellular versus linear), and the types of severe weather observed. These parameters include but are not limited to convective available potential energy (CAPE; e.g., Moncrieff and Green 1972), storm-relative environmental helicity (SREH; Davies-Jones et al. 1990), storm-relative flow (SRFL; Kerr and Darkow 1996), bulk Richardson number (Moncrieff and Green 1972), bulk Richardson number shear (BRNSHR; Droegemeier et al. 1993), bulk shear (Bunkers 2002), lifting condensation level (LCL; e.g., Rasmussen and Blanchard 1998), the energy helicity index (EHI; Hart and Korotky

1991), the supercell composite parameter (SCP; Thompson et al. 2003), and the significant tornado parameter (STP; Thompson et al. 2003).

However, most research has focused on distinguishing *storm* environments. A number of proximity sounding studies (e.g., Davies and Johns 1993; Johns et al. 1993; Brooks et al. 1994a,b; Rasmussen and Blanchard 1998; Craven et al. 2002a,b; Doswell and Evans 2003; Thompson et al. 2003, 2007; Potvin et al. 2010) have focused on the ability of various severe weather parameters to distinguish tornadic and nontornadic storms, supercellular and nonsupercellular convection, significant severe weather from less significant storms, etc. Similar studies using reanalysis data (Brooks et al. 2003b) and mesoscale model output (Stensrud et al. 1997) have been conducted. However, relatively few studies have focused on distinguishing *outbreak* types (Johns and Hart 1993; S09).

Doswell et al. (2006 – hereafter, D06) developed a technique for ranking tornado outbreaks (TOs) and primarily nontornadic outbreaks (PNOs). The goal of that study was not to define these outbreak types, but instead was to rank these outbreaks to identify *prototypical* cases of each type *associated with a particular synoptic-scale system*. This work provided a convenient way to obtain a large sample of outbreaks of each type to determine if antecedent synoptic-scale processes were associated with the type of outbreak that occurred (S09; Mercer et al. 2009 – hereafter, M09).

The subjective (objective) analysis conducted by S09 (M09) showed that the meteorological fields of the mesoscale model simulations of TOs and PNOs initialized with synoptic-scale input could be distinguished consistently up to three days in advance of the outbreaks. However, S09 discussed several limitations of these initial studies. (1)

The samples of 50 TOs and 49 PNOs used in these studies were unequally distributed throughout the year. Specifically, the TOs were most common in April, May, and early June, whereas the PNOs were most common in June, July, and August. No TO in the sample occurred between 17 June and 23 September (i.e., during the peak time of year of the sampled PNOs). (2) As a result of (1), the ability to discriminate TOs and PNOs, as described in S09 and M09, may not have been limited to synoptic-scale processes. Rather, the summer months typically feature high-instability, weak-shear severe-weather environments, whereas the spring months more commonly feature the collocation of moderate-to-high instability and strong shear. These seasonal differences likely played a role in the ability to discriminate the outbreaks. (3) TOs occurring in the late spring were more likely to be misclassified by the subjective and objective techniques. Similarly, PNOs occurring during the time of the year in which the sample of TOs peaked were misclassified more often. (4) The initial study was a "best-case scenario". Specifically, if a mesoscale model using synoptic-scale initial data could not distinguish prototypical TOs and PNOs, then there would be no need to incorporate outbreaks that fall somewhere in between TO and PNO classification. Because the initial work showed promise, these "intermediate" outbreaks should now be included in subsequent studies. In fact, such inclusion is essential, as most severe weather outbreaks cannot be classified readily as either type (Shafer and Doswell 2010a – hereafter, SD10a).

Regarding limitations (1)-(3), a follow-up study by Shafer et al. (2010b – hereafter S10b) modified the PNO case list to include cases only during the times of year in which the sample of TOs used in S09 and M09 occurred. Such modification resulted in a noticeable degradation in the accuracy and skill of the statistical algorithms using the simulated fields of severe weather parameters to distinguish TOs from PNOs. In general, hit rates for the best parameters were ~70-75%, false alarm ratios were ~25-35%, and skill scores were ~50% for 1-day forecasts, with statistically significant deterioration for 3-day forecasts. As found in S09, low-level shear and low-level helicity were the best predictors of outbreak type. Additionally, S10b found that synoptic parameters such as geopotential heights, mean sea-level pressure, and wind speeds were useful predictors of outbreak type. Finally, the utility of some parameters (in particular, the LCL) was found to deteriorate substantially when the PNO case list was altered, suggesting a sensitivity to the time of year in which the events occurred. Clearly, the time of year in which the outbreaks occurred played a substantial role in the discrimination of events in S09 and M09; however, synoptic-scale processes clearly were important, given the statistically significant accuracy and skill (95% confidence) with which these events could be discriminated during the same times of year.

Regarding limitation (4), SD10a modified the ranking indices developed by D06 to include outbreaks of *any* type. Their study included the top 30 days of each year from 1960-2006, based on the total number of severe reports in a given 24-h period, and used a linear-weighted multivariate scheme to rank the 1410 events. As in D06, secular trends in the data were pronounced, requiring temporal detrending. SD10a found that the scores used to rank the events (hereafter, ranking index scores) had characteristic curves (their Fig. 6), suggesting three basic types of events based on their relative severity, supported by subjective analysis of these events: major severe weather outbreaks (~200 cases, primarily major tornado outbreaks), intermediate events (~1000 cases), and marginal events (~200 cases). Marginal events were cases with substantial geographic scatter in the reports or multiple clusters of events on a given day, identified readily by incorporation of the "middle-50%" parameter (described in D06 and SD10a). Although the rankings of the major and marginal events were relatively robust to modifications of the ranking indices, the intermediate outbreak rankings were subject to large volatility. Thus, the prediction of future events using past events as training using the ranking index scores was doubtful. As a result, classification was deemed more appropriate. Besides the 3-class designation for relative severity, cluster analysis was used to classify outbreaks by the characteristics of the severe reports. SD10a found that five types of outbreaks were observed: major tornado, hail-dominant, wind-dominant, mixed-mode, and large-scatter events.

As S09 observed that areal coverage of parameters favorable for significant severe weather seemed to be a relatively useful means of discriminating TOs from PNOs, this technique was used in an objective manner to determine its capability of distinguishing the major severe weather outbreaks in SD10a from the less significant events (i.e., the intermediate and marginal events in SD10a). This was the focus of the study by Shafer et al. (2010a – hereafter S10a). S10a found that the areal coverage of severe weather parameters favorable for severe weather was a relatively useful way of diagnosing the severity of the outbreaks. However, a substantial false alarm problem was present, similar to results from previous studies discriminating storm environments (e.g., Rasmussen and Blanchard 1998; Thompson et al. 2003) and confirmed the findings of Hamill et al. (2005), who specifically investigated multi-day tornado outbreaks. Although several severe weather parameters were tested, it was found that SCP, STP, and EHI were relatively good predictors.

With the development of two techniques (the principal component analysis first introduced in M09, and areal coverage first introduced in S10a), a comparison of the efficacy of these methods for outbreak discrimination is appropriate. This is so because (1) the areal coverage technique is simpler to develop, automate, and interpret; (2) the principal component analysis technique is more computationally demanding; and (3) the principal component analysis technique requires knowledge of the location of the center of the outbreak, which is not known before the event occurs. If it is found that the areal coverage method is as good as or better than the principal component analysis technique, there would be no need for implementation of the principal component analysis technique in an environmental setting (currently). This comparison is the motivation for Section 2 of this study.

In their study, S10a found that the marginal events identified by SD10a were not treated in a way that agreed with subjective notions of these events. Specifically, days with multiple clusters of events were considered single events, whereas the synoptic-scale environments on these days suggested that they should be considered as distinct events. Moreover, days with large geographic scatter may not be outbreaks at all and should be discounted entirely. These limitations are the motivation for Sections 3 and 4 of this study.

Finally, Section 5 elaborates on the overall findings of the research conducted so far in determining our present ability to distinguish major severe weather outbreaks (primarily major tornado outbreaks) from less significant events. In this section, the limitations of outbreak discrimination are identified. Additionally, future work based on these limitations is proposed.

2. Diagnosing Major Severe Weather Outbreaks: Comparison of NARR and NCEP Reanalysis Datasets and Evaluation of Principal Component and Areal Coverage Techniques

a. Introduction

Two objective methods of diagnosing characteristic types of outbreaks have been proposed in recent studies. M09 introduced a technique in which model-simulated fields of meteorological covariates could be used as input into statistical models (such as support vector machines [SVMs; Haykin 1999]) via principal component analysis (PCA; Hotelling 1993, Richman 1986) for training, and using independent cases to assess the models. Henceforth, this technique will be referred to as the PCA method (Fig. 2.1). Specifically, M09 investigated the ability of the Weather Research and Forecasting model (WRF; Skamarock et al. 2008) to discriminate tornado outbreaks and primarily nontornadic outbreaks (D06), using synoptic-scale data for initialization (S09). M09 found that the proposed technique was effective in distinguishing the two types of convective outbreaks at least three days in advance. However, S10b found that this capability was diminished when the selection of cases was restricted to events occurring outside of the summer months.

S10a introduced a simplified technique to distinguish major severe weather outbreaks from intermediate and marginal outbreak days. Operational severe weather forecasting experience suggested that days in which spatially extensive regions of severe weather parameters favorable for the occurrence of significant severe weather occurred in conjunction with significant convective outbreaks. In contrast, days with relatively small regions of observed severe weather frequently feature comparably small regions with

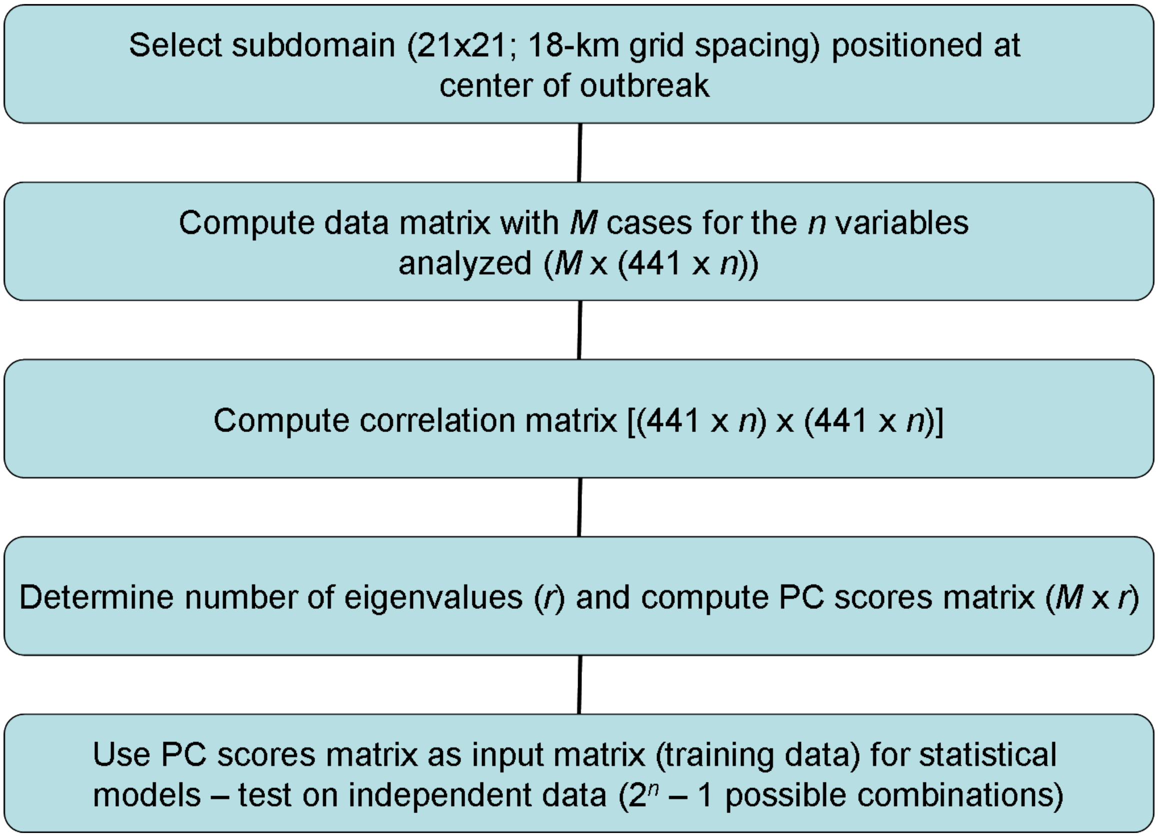

Select subdomain (21x21; 18-km grid spacing) positioned at center of outbreak
Compute data matrix with *M* cases for the *n* variables analyzed (*M* x (441 x *n*))
Compute correlation matrix [(441 x *n*) x (441 x *n*)]
Determine number of eigenvalues (*r*) and compute PC scores matrix (*M* x *r*)
Use PC scores matrix as input matrix (training data) for statistical models – test on independent data (2^n – 1 possible combinations)

Fig. 2.1. Flow chart outlining the PCA method.

favorable severe weather parameters. Thus, S10a used areal coverage of severe weather parameters exceeding a predetermined threshold as input for statistical models trained to identify the major severe weather outbreaks. Henceforth, this technique will be referred to as the areal coverage method. S10a used North American Regional Reanalysis (NARR; Mesinger et al. 2004, Mitchell et al. 2004) data valid at or near the peak times of the outbreaks to compute values of areal coverage (either as grid point sums or mean storm trajectories in the favorable region). The results found by S10a suggested that areal coverage was indeed associated with the relative severity of these convective outbreaks, despite a high (>70%) false alarm ratio.

A comparison of the efficacy of the two techniques for outbreak discrimination using the same dataset and the same type of outbreak classification is appropriate, for a number of reasons. (1) The areal coverage technique is simpler to execute and is easier to interpret. The computational demands required for the PCA method (described in S10b; their Section 2) are reduced considerably using the areal coverage technique.[1] The a priori determination of the outbreak centroid, necessary for operational implementation of the PCA method (M09), also is eliminated. Questions regarding the spatial distribution of the outbreak (in terms of the observed severe weather and the observed severe weather environment) are unaccounted for in the current PCA technique (S10b). These concerns are not relevant for the computation of areal coverage. Automation of the PCA method is challenging, because of these limitations. (2) S09 observed that the areal coverage of severe weather parameters was useful for discriminating between tornado and primarily nontornadic outbreaks. Thus, it is possible that the PCA and areal coverage methods may be providing solutions that are largely (statistically significantly) indistinguishable. If so, the areal coverage method is preferable, based on the aforementioned increased complexity of the PCA method. (3) Because of the large number of false alarms when distinguishing major from intermediate and marginal outbreak days using the areal coverage techniques, developing methods to improve upon these initial results is desirable. The PCA method is one possible alternative.

The PCA method used by M09 and S10b (refer to Fig. 2.1) uses a 21x21 domain with 18-km horizontal grid spacing located at the center of the relevant outbreak. This was designed to represent the mesoscale environment in the vicinity of the outbreak in order to determine if mesoscale models can produce consistently distinct environments with synoptic-scale data. However, as outbreaks are associated with synoptic-scale systems (Doswell et al. 1993; Glickman 2000; D06), the synoptic-scale environment at

[1] The computational demand also is discussed briefly in Section 2b.

the valid time of the outbreak may also be a means of discriminating outbreak type. This has been suggested numerous times in the literature (e.g., Miller 1972; Johns and Doswell 1992; Doswell et al. 1993; Doswell and Bosart 2001; S09). Therefore, use of a second dataset, in which the grid spacing is relatively coarse, can be implemented to determine how grid spacing influences outbreak type discrimination.

The purpose of this study is to attempt to determine which of two methods tested (the PCA vs. the areal coverage method) is the best choice in the discrimination of major severe weather outbreaks from intermediate and marginal events, as identified by SD10a, using reanalysis data valid at or near the peak time of the outbreak, as done for the areal coverage method in S10a. The utility of the PCA method is investigated further by using both the NARR dataset and the National Centers for Environmental Prediction (NCEP)/National Center for Atmospheric Research (NCAR) reanalysis (hereafter, NNRP; Kalnay et al. 1996) dataset. As the NNRP dataset is consistent with the synoptic scale (see S09), whereas the NARR dataset is of higher resolution (see Mesinger et al. 2004; Mitchell et al. 2004; Section 5b of this study), comparison of each dataset's ability to diagnose major severe weather outbreaks will provide insight into the additional value (if any) provided by enhanced spatial resolution data for this purpose. Section 2b discusses the data and methods incorporated in this study. Section 2c presents the results comparing the NARR and NNRP data using the PCA method. Section 2d compares the PCA and areal coverage techniques. Section 2e provides additional discussion on possible reasons for the findings and implications for future research. This research also is discussed in Shafer et al. (2010c).

b. Data and methods

1) The reanalysis datasets

The NARR dataset was described by Mesinger et al. (2004) and Mitchell et al. (2004), as well as by S10a. Data are available from 1 January 1979 to the present, allowing for many cases to be analyzed. Horizontal grid spacing is 32 km, with 45 vertical layers. Such a dataset allows for analysis of the mesoscale fields of various meteorological variables associated with severe weather (known as meteorological covariates – see Brown and Murphy 1996; Brooks et al. 2003b). The reanalysis data are converted to a 31-layer, 300x200 18-km Lambert conformal grid centered on the conterminous United States (CONUS), which is consistent with the studies by M09, S10a, and S10b. This conversion incorporated the WRF Preprocessing System (WPS) Version 3.1 (Skamarock et al. 2008), using bilinear interpolation.

The NNRP dataset is documented in detail in Kalnay et al. (1996). Data are available from 1 January 1948 to the present, which allows many more cases to be considered than with the NARR data. Horizontal grid spacing is 2.5° latitude-by-longitude, with 17 vertical levels. The process of optimal interpolation-based quality control (Woollen 1991; Woollen et al. 1994) removes observations that are spatially and temporally inconsistent with the grid spacing present in the NNRP dataset. This procedure minimizes subsynoptic-scale features from the dataset, as the horizontal and vertical grid spacing is of the synoptic scale (e.g., Orlanski 1975). The NNRP data were converted to a 31-layer, 150x100 50-km Lambert conformal grid centered on the

CONUS[2], using the same method as for the NARR data. The reasons for the differences in the converted grids are discussed in the following subsection.

2) The analysis techniques

The PCA method (Fig. 2.1) uses vector fields of meteorological variables, which are mapped from a loading matrix that is column orthogonal, individually or in combination. The PC loadings represent a compact representation of the variance structure of the similarity matrix of a sample of data. The assumptions are (1) that the information in the similarity matrix is statistically reliable representing the coherent signal in the data which, in turn, assumes that sufficient sampling of recurring patterns in the data exists to generate a stable variance structure[3] and (2) that the relationships among the variables are largely linear. For the present study, a subdomain of the reanalysis fields positioned at the center of the outbreak is obtained. The center of the outbreak was assumed to be the median latitude and median longitude of all reports within six hours of the outbreak valid time. The subdomain contains 441 grid points (21x21), shown to be of practical size in previous work (M09). One of the advantages of using relatively coarse grids is that a larger area centered on the outbreak can be analyzed with the PCA method without additional computational demand. That is, when a subdomain of the same *x*- and *y*-dimensions is selected for the PCA method using the NARR and NNRP datasets, the subdomain for the NNRP dataset will comprise a larger area than that of the NARR subdomain.

[2] Although the data were converted to lower horizontal grid spacing, this does not mean additional information was extracted, by definition. The conversions of the NARR and NNRP data to Lambert conformal grids allowed for relatively simple comparison using the same map projection.

[3] It will be shown later in this section that this assumption likely is not met in the present work.

S10b (their Section 2d) found that increasing the size of the subdomain, so that a larger area surrounding the observed outbreak center was analyzed, did not change the results in a statistically significant manner. This was true despite the relatively small size of the domain (25 600 km^2) and without accounting for phase/timing errors of a model's forecast location of the outbreak. The latter is not a factor in this study, as I am looking at reanalysis fields valid at the time of the outbreak (rather than model forecast fields). Thus, one should expect the small size of the NARR subdomain compared to the NNRP subdomain to be of limited concern. As in S10b, preliminary tests indicate that changing the NARR subdomain to 41x41 in this study does not alter the results in a statistically significant manner (not shown). Therefore, the objective is to compare the NARR and NNRP datasets using the same dimensions (number of grid points) of the data matrix and not of the same region within the subdomain.[4]

Next, the 441 grid points for each of the n variables to be analyzed are converted to an M x (441 x n) data matrix, where M is the total number of cases. The correlation matrix of the S-mode decomposition (Richman 1986) is computed and is of size (441 x n) x (441 x n), which results in increased computational demand when using larger subdomains or multiple variables in the analysis. Once the number of eigenvalues to retain is computed, the principal component scores matrix is used as the input matrix for statistical algorithms used to discriminate major severe weather outbreaks from less significant events (see the following subsection). These scores correspond to linear combinations of the PC loading fields, which contain information on the gradients of the variables.

[4] However, for completeness, I also will be comparing NARR and NNRP data with the same grid spacing. See Section 2c.

The areal coverage technique is a much simpler approach. For each outbreak day, the number of grid points exceeding a specified threshold for a particular variable or combination of variables is determined to be the areal coverage.[5] As S10a observed, using these values led to some undesirable results. For example, grid points over water initially were included, leading to large areas of parameters favorable for severe weather over the Atlantic Ocean and the Gulf of Mexico on some days where no severe weather could actually be observed. Water points were excluded hereafter. Some severe weather parameters, notably indices such as EHI, SCP, and STP (see Doswell and Schultz 2006 for a discussion of severe weather indices), were very large on high-CAPE, low-shear days (which generally are not observed on major tornado outbreak events). As a result of these tendencies, two types of areal coverage calculations were examined. The *unconstrained* values used the original number of grid points. The *constrained* values included the sum of grid points exceeding specified thresholds, after excluding all points with surface-based convective available potential energy (SBCAPE) < 1000 J kg^{-1} and 0-1 km storm relative environmental helicity (SREH) < 100 m^2 s^{-2}.

Areal coverage was also examined by computing hypothetical storm trajectories within the region considered to be favorable for significant severe weather. These pseudo-trajectories were calculated for each grid point within the region, using the 500-hPa winds to develop forward and backward trajectories. Assuming the winds and the favorable region were unchanged, the distances and times within the favorable region were calculated for which a hypothetical storm at each grid point. The sum and mean

[5] No subdomain is computed for the areal coverage method. The entire domain is used for the calculation of areal coverage. Furthermore, the areal coverage technique only is computed using NARR data. Although the values of areal coverage change using NNRP data, the relative areal coverage values are similar to those of NARR data.

distances were computed for each outbreak day. Although these storm motions obviously are crude estimates, these computations are acceptable in comparing each outbreak to the others (as in S10a).

3) The outbreaks

As NARR data are available from 1979 to the present, the top 30 outbreak days each year, determined by SD10a, from the period 1970-2006 were considered in this study (840 cases). An outbreak day was considered to be 1200 UTC on the nominal date to 1159 UTC the following day. The outbreak valid time was determined subjectively using the times of each of the reports for the 24-h period.

Outbreak classification was identical to that of S10a. That is, an outbreak day was considered to be *major* if the score of the outbreak ranking index selected was greater than unity. The selection of this value was based on the finding that the mean of the scores was approximately zero for each of the indices and that the standard deviation of the scores was approximately 1 for each of the indices. Any score matching or exceeding the mean plus one standard deviation was classified as a *major* severe weather outbreak, and any outbreak below this threshold was considered a null case. This study, for reasons discussed in S10a, will focus on the N15 index (Fig. 2.2; see SD10a for a description of the various ranking indices), though results are similar when selecting alternative indices. SD10a showed that almost every major severe weather outbreak was a tornado outbreak, whereas the null cases clearly were not tornado outbreaks.[6]

[6] As S10a discussed, the selection of the outbreak classification threshold was subjective, based on the tendency for outbreaks above this threshold to be major tornado outbreaks and events below the threshold not to be. S10a suggested modifying this threshold to test, and this is discussed at length in Section 4.

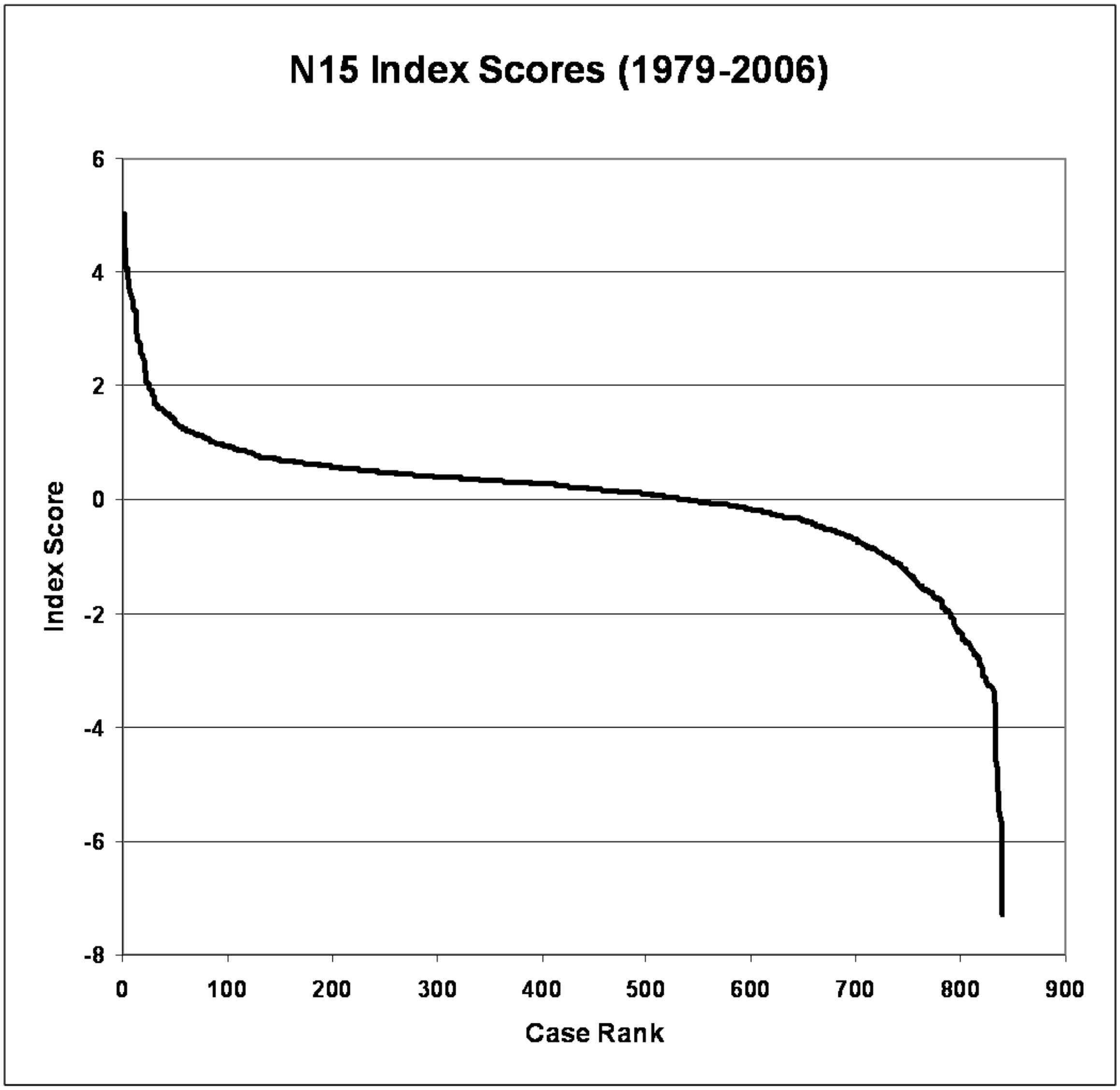


Fig. 2.2. Scores of the N15 index for the 840 cases analyzed in this study.

4) Statistical algorithms

Once the PCA and areal coverage methods were completed for each case, an input matrix was created. For the PCA method, the input matrix was of the principal component scores (*M* x *r*), where *r* is the number of retained principal components. For the areal coverage method, the input matrix was the number of grid points exceeding predetermined thresholds, or the sum or mean times and distances of hypothetical storm pseudo-trajectories within the favorable region. This matrix was of size *M* x *p*, where *p* is the number of parameters for which the areal coverage is computed. Two sets of matrices were computed for each method: a training matrix and a testing matrix. The training input matrix was used to develop the statistical model, and the testing input matrix was used to provide independent events for evaluating the accuracy and skill of the statistical models created.

Of the 840 available cases, a subset of 630 randomly selected cases was used for training, and the remaining cases were used for testing. The selection of 75% of the cases for training was based on the increasingly poor representativeness of the training set as the number of cases chosen for training decreased (see Section 3b of S10a for a discussion). Of course, leaving a small number of cases to test leads to higher uncertainty with the testing data, and this is discussed in the following sections.

Results were analyzed in two ways. (1) To test the uncertainty of the training models, a set of 63 cases (10%) was removed randomly from the training cases, and the training model was developed with the remaining 567 cases.[7] This procedure was conducted 25 times. The accuracy and skill of the 25 statistical models were then evaluated *on the same testing data*.[8] These contingency statistics for each of the statistical models were compared directly to determine training model variability. (2) The results of the training model using all of the 630 training cases were sampled with replacement. Bootstrap confidence intervals (Efron and Tibshirani 1993) were analyzed to assess the uncertainty of the model diagnoses of the testing data.

[7] As discussed in S10a, the removal of 10% of the cases was a compromise. Removing a larger proportion of cases led to reduced accuracy and skill of the training model, whereas removing a smaller proportion of cases led to substantially larger uncertainty in the statistics.

[8] Increasing the number of statistical models from the selection of 25 yielded similar findings. Thus, selection of a relatively small number of statistical models was preferred as computational demand is lower. As the purpose of this analysis was to gain a general sense of statistical model variability, a large number of statistical models was deemed unnecessary.

Table 2.1. Statistical algorithms and identification numbers in the relevant figures for this section.

Statistical Algorithm	**ID Number**
Linear discriminant analysis (multivariate normal density; pooled covariance estimate)	1
Quadratic discriminant analysis (multivariate normal density; covariance estimates stratified by groups)	2
Linear discriminant analysis (multivariate normal density; diagonal covariance matrix estimates – naïve Bayes classifiers)	3
Quadratic discriminant analysis (multivariate normal density; diagonal covariance matrix estimates – naïve Bayes classifiers)	4
Decision trees	5
Support vector machines (SVMs) – radial basis kernel function (RBF); quadratic programming (QP)	6
SVMs – linear; QP	7
SVMs – quadratic polynomial; QP	8
SVMs – third-order polynomial; QP	9
SVMs – RBF; minimal sequential optimization	10

Based on the small sample size of the testing data, relatively large uncertainty was expected and observed, as the following sections will show. The same ten statistical techniques as used in S10a are used for this study as well (Table 2.1). Using several statistical techniques permitted investigation of deficiencies with particular methods of statistical modeling or with particular combinations of meteorological covariates.

Several binary contingency statistics were analyzed. These included the hit rate (HR), probability of detection (POD), false alarm ratio (FAR), probability of false detection (POFD), critical success index (CSI), Heidke's skill score (HSS), Peirce's skill score (PSS), Clayton's skill score (CSS), and Gilbert's skill score (GSS). These statistics are discussed in Doswell et al. (1990), Wilks (1995), Murphy (1996), Richardson (2000), Wandishin and Brooks (2002), and many others. The selection of these statistics and their equations are discussed in S10a. Results in this study will focus on HR, POD, FAR, CSI, PSS, and HSS.

The meteorological variables analyzed in this study were identical to those of S10b (their Table 4). Although combinations of covariates were analyzed in this work, only results for individual variables will be presented here, owing to space constraints and only minimal observed improvement in contingency scores with inclusion of multiple predictors. Finally, as this study compares the ability to distinguish outbreak types using the PCA and areal coverage methods, and using the PCA method with NNRP and NARR data, *this is a diagnostic study*. The meteorological covariates used in this study cannot be assumed to be forecast parameters and are not forecast diagnostic variables (Doswell and Schultz 2006).

c. Comparing NNRP and NARR using the PCA method

1) The training models

Using the significant tornado parameter (STP; Thompson et al. 2003) as an example meteorological field to be analyzed for discriminating major severe weather outbreaks from less significant events (with this categorization based on the N15 index), the PCA method reveals similar results using NARR data (Fig. 2.3) and NNRP data (Fig. 2.4). Specifically, the HR and FAR using either reanalysis were generally in the 0.7-0.8 range, with CSIs generally between 0.2 and 0.3. The POD was subject to relatively large variability, which is expected, given that the number of major severe weather outbreaks is much lower than that of the null cases. The PSS and HSS appeared to be generally higher for the NNRP dataset, though the best-performing statistical models featured quantitatively similar scores between the reanalysis datasets.

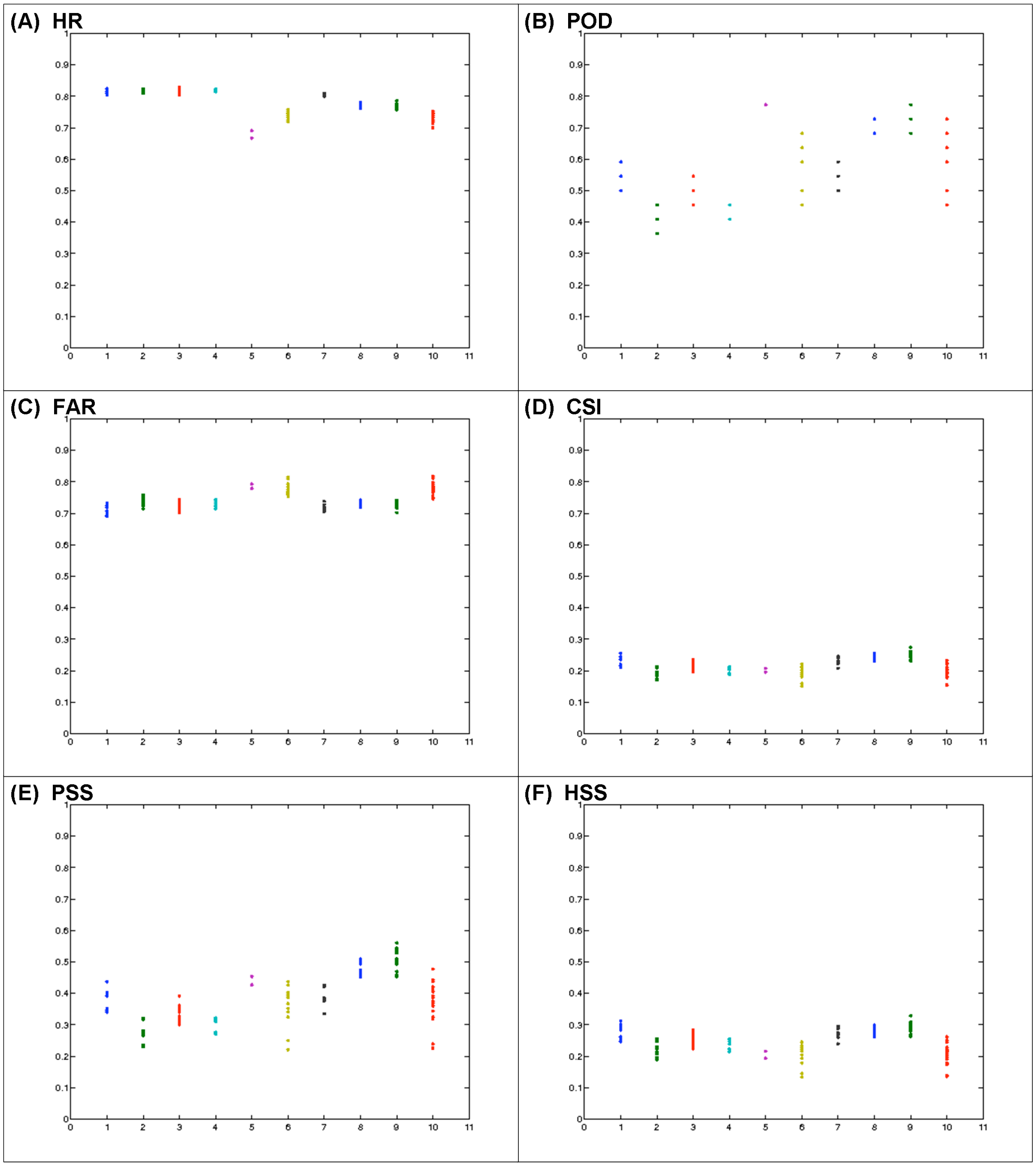


Fig. 2.3. Contingency statistics (labeled) for each of the 25 training models developed, when tested on the same set of 210 independent cases. Redundant statistics are overlain on the same values. Results are for the NARR STP fields at the outbreak valid time. Classification of outbreaks is based on the N15 index scores. Each statistical model is labeled on the *x*-axis and is identified in Table 2.1.

Variability among the 25 statistical models was small but not negligible. For HR, FAR, and CSI, variations of the scores generally were around 0.05, whereas skill scores

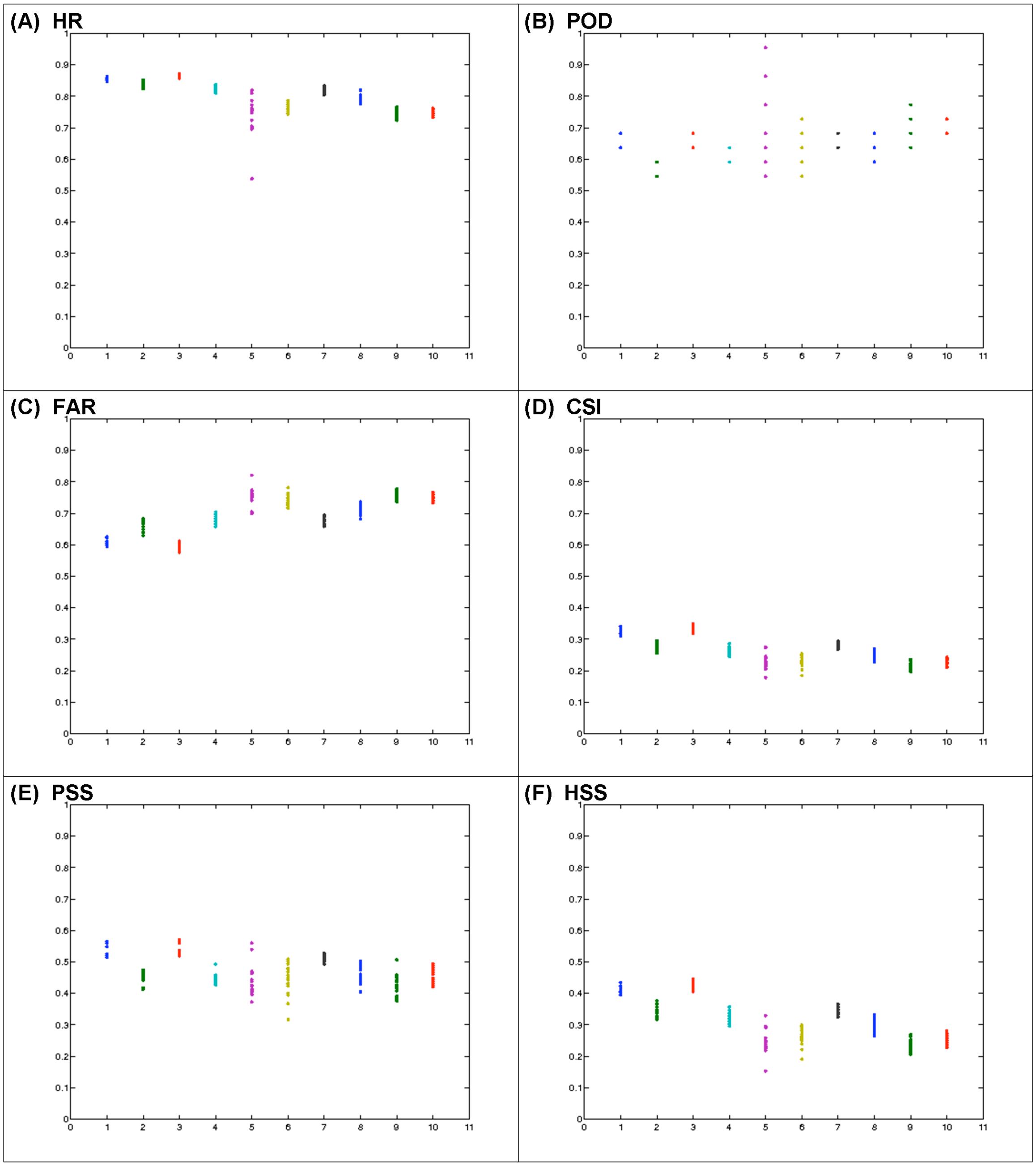


Fig. 2.4. As in Fig. 2.3, using the NNRP STP fields.

ranged from 0.05 to 0.1. Differences in the ranges of the statistics between NNRP and NARR data were minor (generally on the order of 0.05), with some exceptions. For example, the training models using decision trees featured limited variability when using

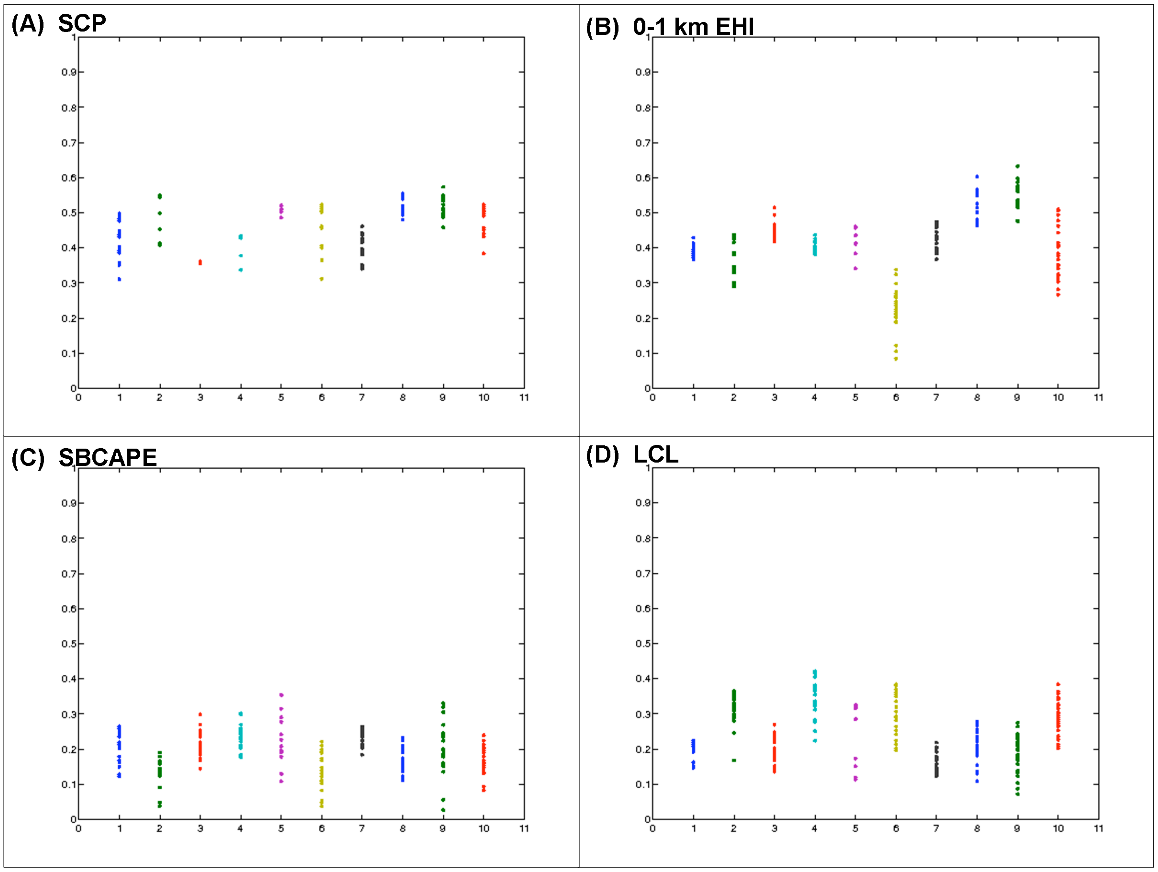


Fig. 2.5. Peirce skill scores of the PCA method for (a) SCP, (b) 0-1 km EHI, (c) SBCAPE, and (d) LCL using NARR data at the valid times of the outbreaks. Statistical models labeled on the *x*-axis as identified in Table 2.1.

the NARR data but pronounced variability using NNRP data. Additionally, differences were not universal among all meteorological covariates (e.g., Figs. 2.5-2.7).

Qualitatively, some statistical algorithms appeared to distinguish the major severe weather outbreaks somewhat better than others. For example, analyzing STP with the NARR data, LDA, the decision tree, and the SVM with polynomial kernel functions performed well compared to QDA and the SVMs using the radial basis kernel function (refer to Table 2.1 for statistical algorithm abbreviations). When using the NNRP data, LDA and the SVM incorporating the linear kernel function performed somewhat better than the other statistical algorithms. These results may be indicative of noise that

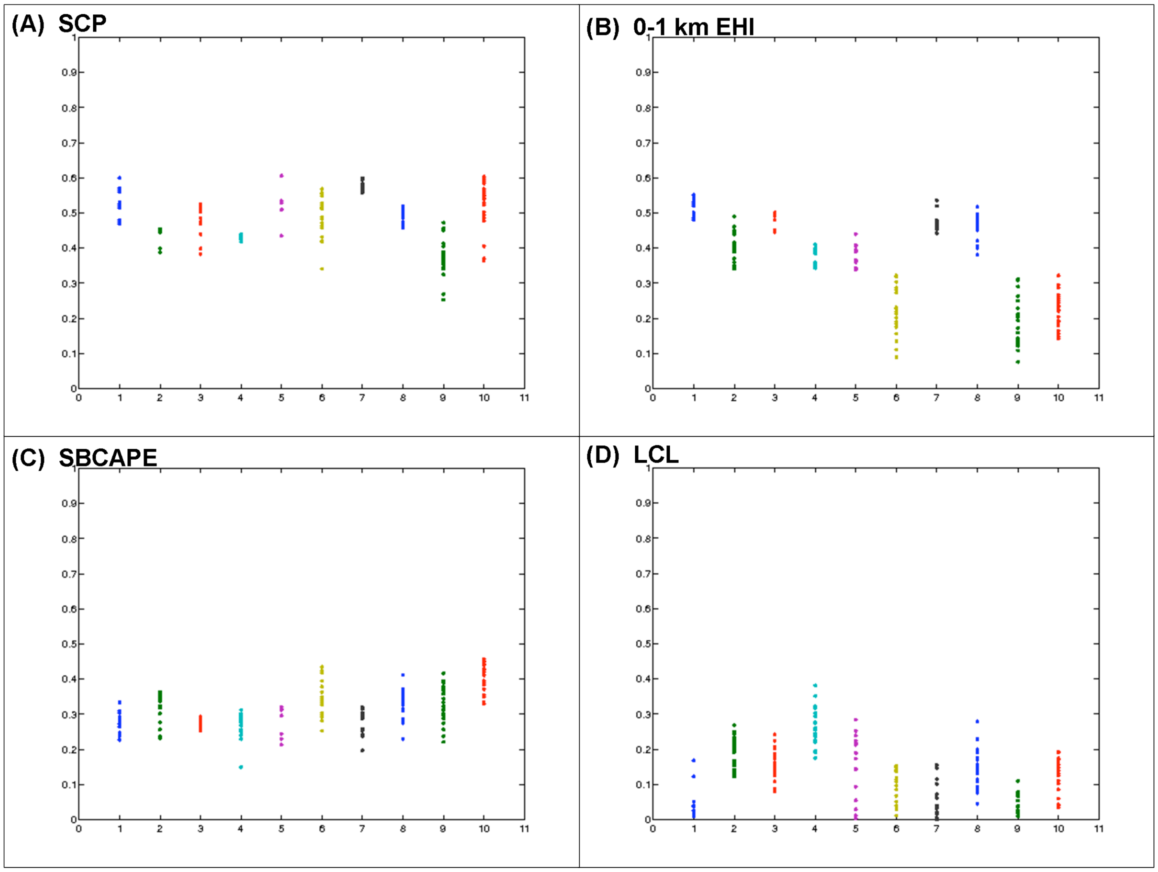


Fig. 2.6. As in Fig. 2.5, using NNRP data.

prevents nonlinear techniques from generalizing adequately to diagnose independent data accurately. However, these differences among the statistical models are not universal when analyzing other meteorological covariates. For example, the statistics for the SVMs with radial basis kernel functions perform somewhat better than for the SVMs with polynomial kernel functions for LCL but not for SBCAPE and 0-1 km EHI (Fig. 2.4). Additionally, using the SVM with the linear kernel function results in statistics similar to SVMs with higher-order polynomial kernel functions for SBCAPE and LCL but not for SCP and EHI.

These findings suggest that no statistical algorithm used in this study is universally superior to the others. Moreover, *multiple algorithms should be tested for*

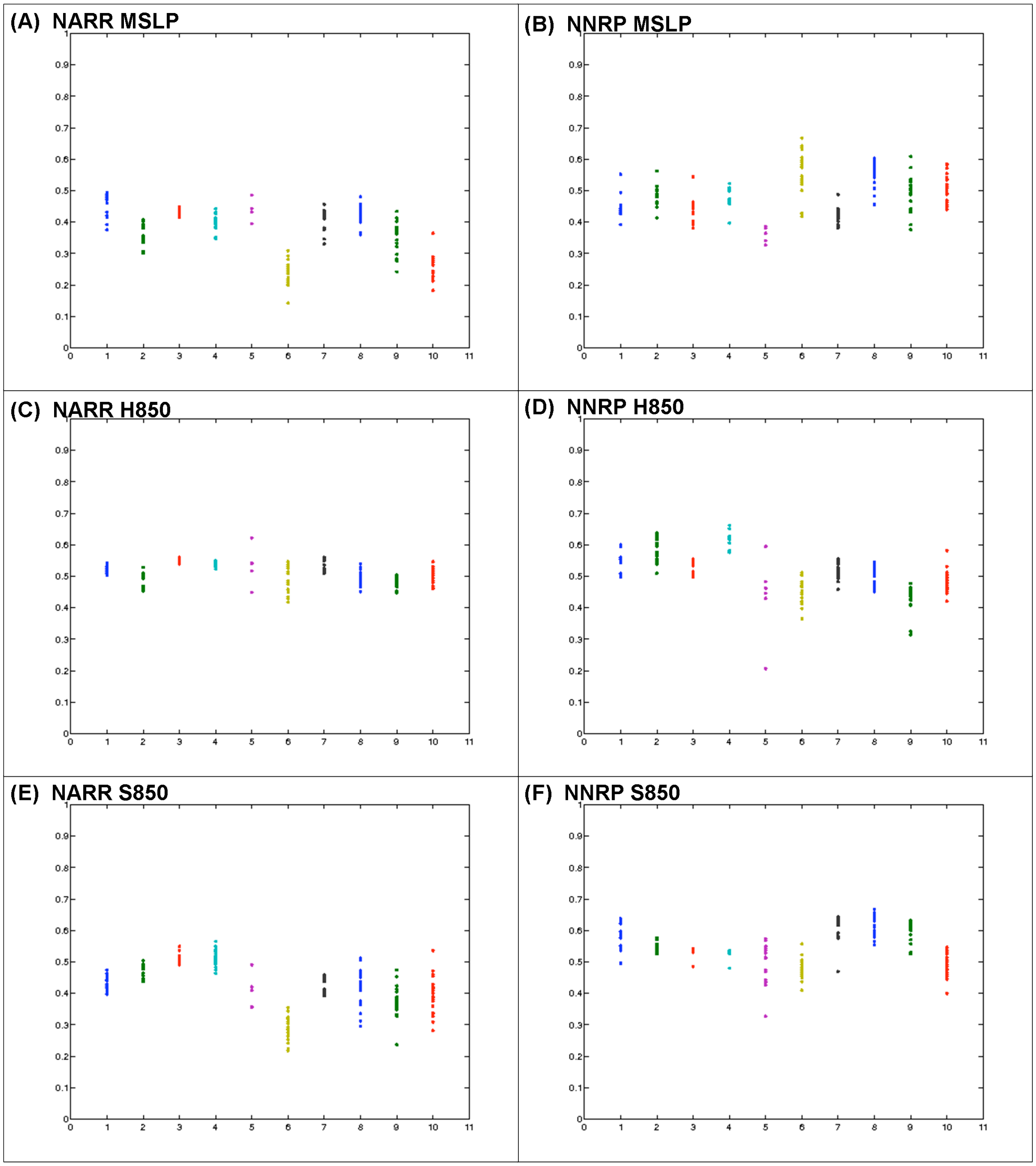


Fig. 2.7. (a) Peirce skill scores of the 25 statistical models for the PCA method, using NARR data valid at the time of the outbreaks, for the mean sea level pressure fields. (b) As in (a), using NNRP data. (c) As in (a), for the 850-hPa geopotential heights. (d) As in (c), using NNRP data. (e) As in (a), for the 850-hPa wind speeds. (f) As in (e), using NNRP data. Statistical algorithms labeled on the *x*-axis and identified in Table 2.1.

every parameter (meteorological covariate) analyzed because of parameter-to-parameter variability with the relative performance of the discriminating algorithms. Because the analysis of the variability of the training models is limited to that of 25 statistical models per algorithm and because uncertainty in the testing data is large (see the following subsection), preference of a particular statistical technique for a particular meteorological covariate is a dubious task.[9] Moreover, computation of a sufficient number of models for confidence interval comparison is possible but computationally intractable (Section 2b).

The STP was among the meteorological covariates that performed relatively well (i.e., statistical models with average PSS > 0.4). This was true for the NARR (cf. Figs. 2.3 and 2.5) and NNRP (cf. Figs. 2.4 and 2.6) datasets. Other variables performing reasonably well included the supercell composite parameter (SCP; Thompson et al. 2003), the energy-helicity index (EHI; Hart and Korotky 1991), storm-relative environmental helicity (SREH; Davies-Jones et al. 1990), and low-level and deep-layer bulk shear (Bunkers 2002). Thermodynamic instability parameters, such as CAPE and the lifting condensation level (LCL), performed relatively poorly, which agrees with numerous previous studies focused on discriminating storm types (e.g., Johns et al. 1993; Rasmussen and Blanchard 1998; Monteverdi et al. 2003) or outbreak types (S09). As discussed in S10b, synoptic parameters such as mean sea-level pressure (MSLP), 850-hPa geopotential heights (H850), and 850-hPa wind speeds (S850) also performed reasonably well (cf. Figs. 2.5 and 2.7). Finally, the variability of scores with the 25 statistical models was larger with some parameters than with others. The ranges of PSSs of the 25 statistical models for some variables exceeded 0.2, depending on the statistical

[9] However, some of the techniques could be excluded from the analysis. For example, the relatively high variability of the decision tree models with some covariates suggests susceptibility to overfitting, even after pruning. This tendency, and a lack of evidence showing statistically significant superiority over other algorithms in most analyses, may provide sufficient reasoning for its exclusion in general.

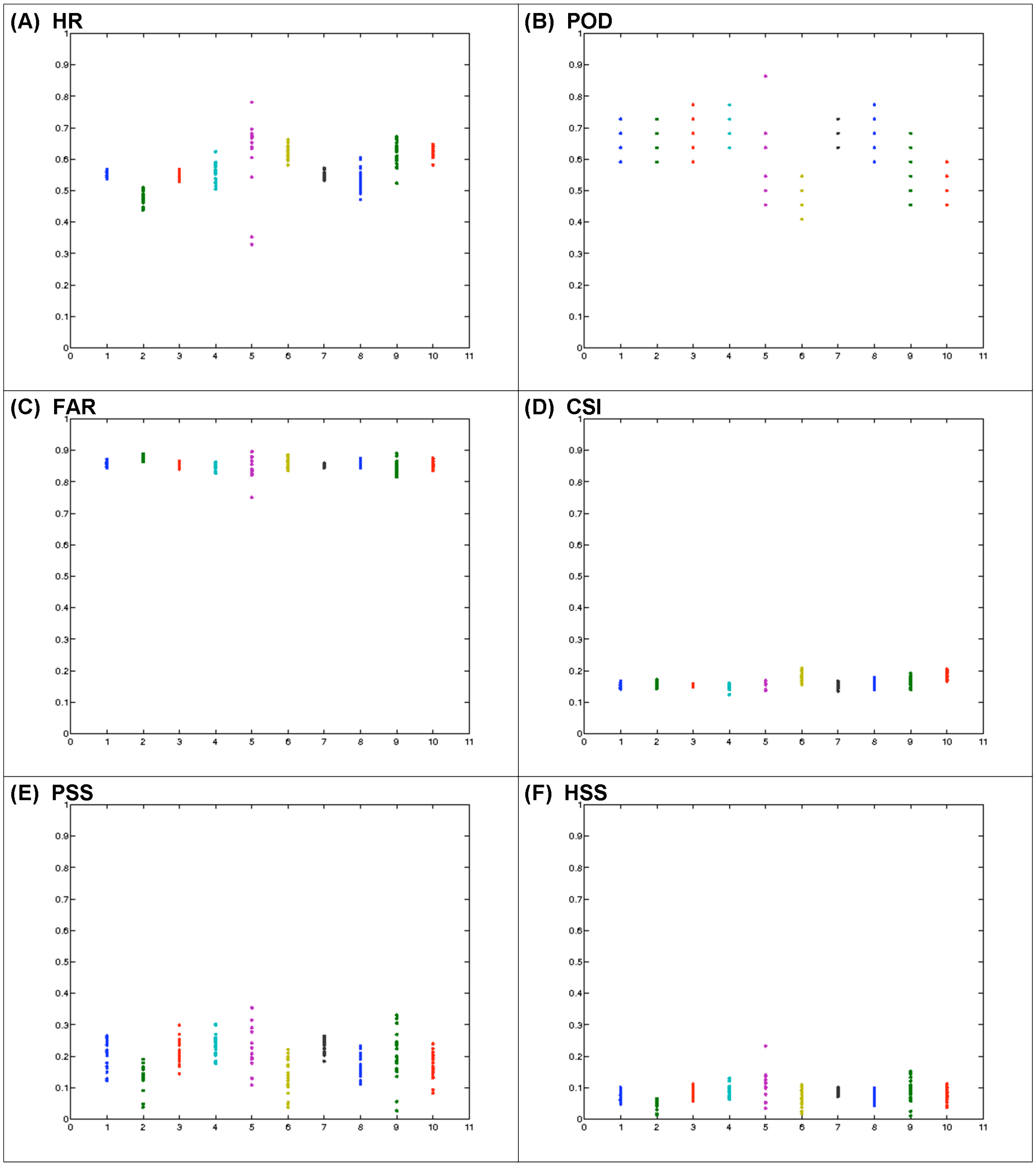


Fig. 2.8. As in Fig. 2.3, analyzing SBCAPE rather than STP.

algorithm incorporated (Figs. 2.5-2.6). The ranges appeared to be somewhat larger for the thermodynamic instability parameters, in general.

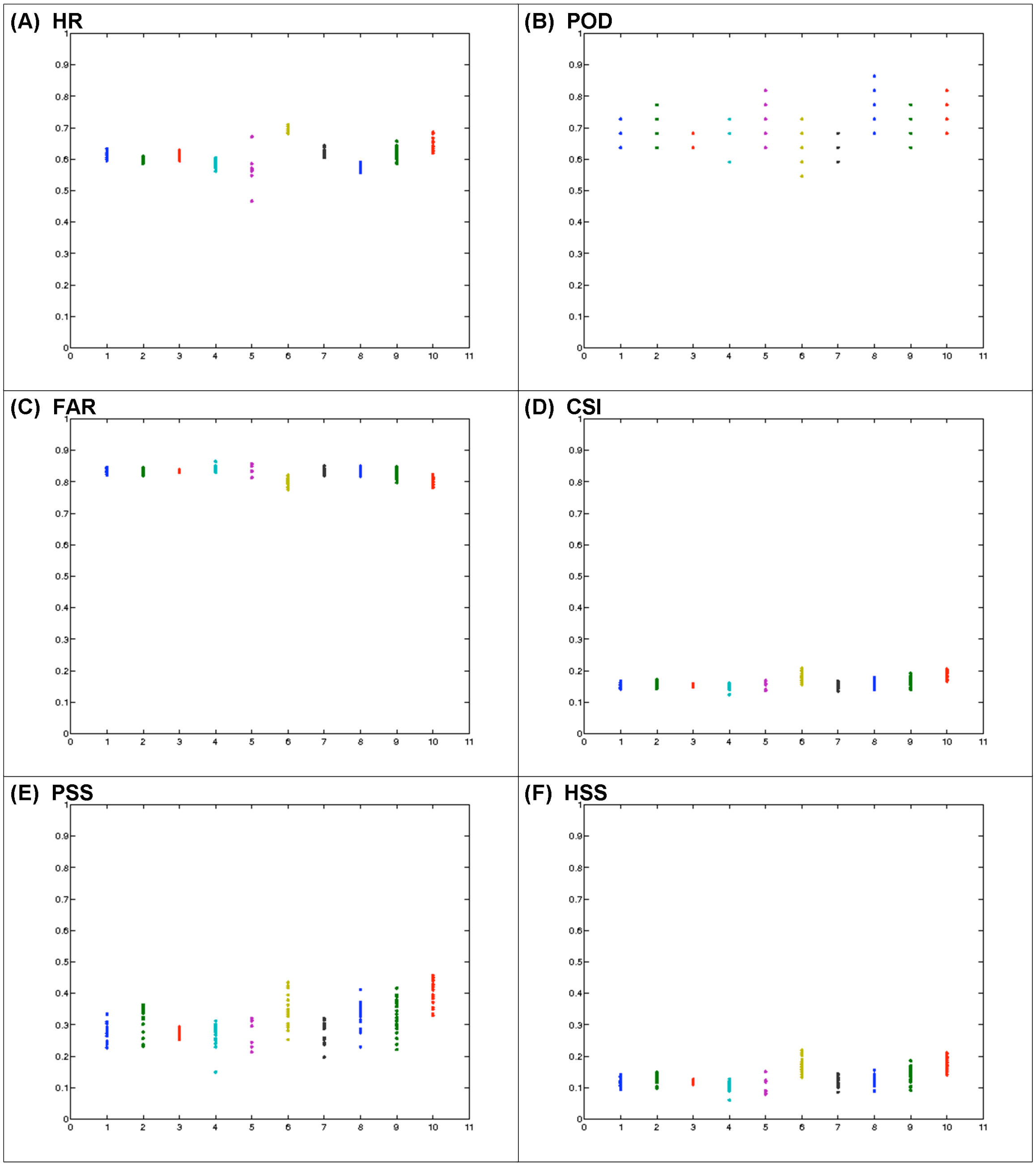


Fig. 2.9. As in Fig. 2.8, using NNRP data.

Comparisons of NNRP and NARR results with other variables were similar to the findings for STP. For example, the statistics for SBCAPE (Figs. 2.8 and 2.9) were similar between datasets. Clearly, the results were worse overall for SBCAPE than for STP (cf. Figs. 2.8 and 2.3; e.g., PSS decreased by 20% on average) and indicated little

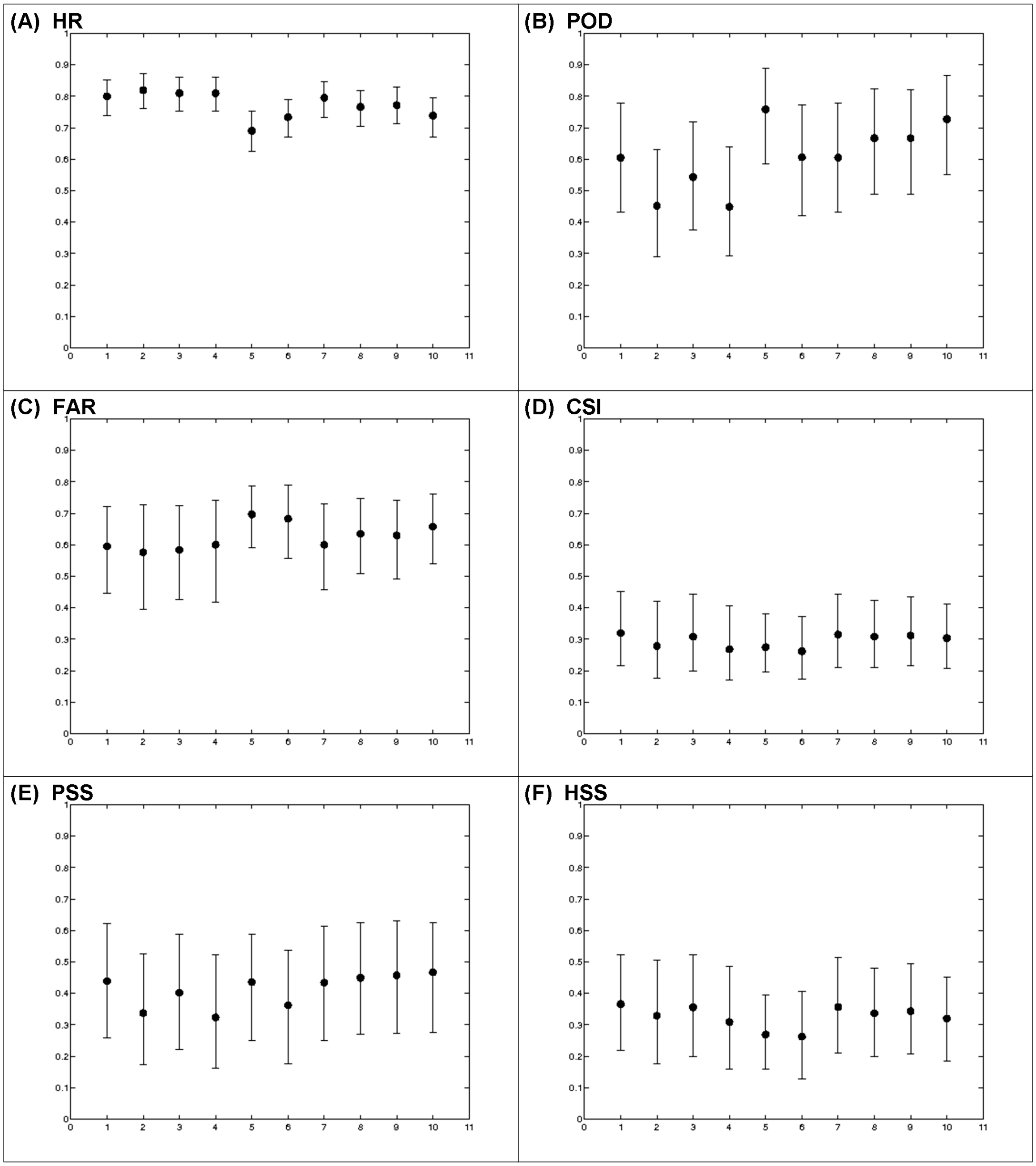


Fig. 2.10. Contingency statistics (labeled) of the testing data, incorporating all of the training data, for the PCA method. Confidence intervals (95%) are shown using error bars, with the median of the confidence interval indicated by a black dot. Results are for the STP fields using the NARR dataset. Statistical algorithms as labeled in Fig. 2.3.

difference in the performance of the PCA method using either reanalysis dataset for the same variable.

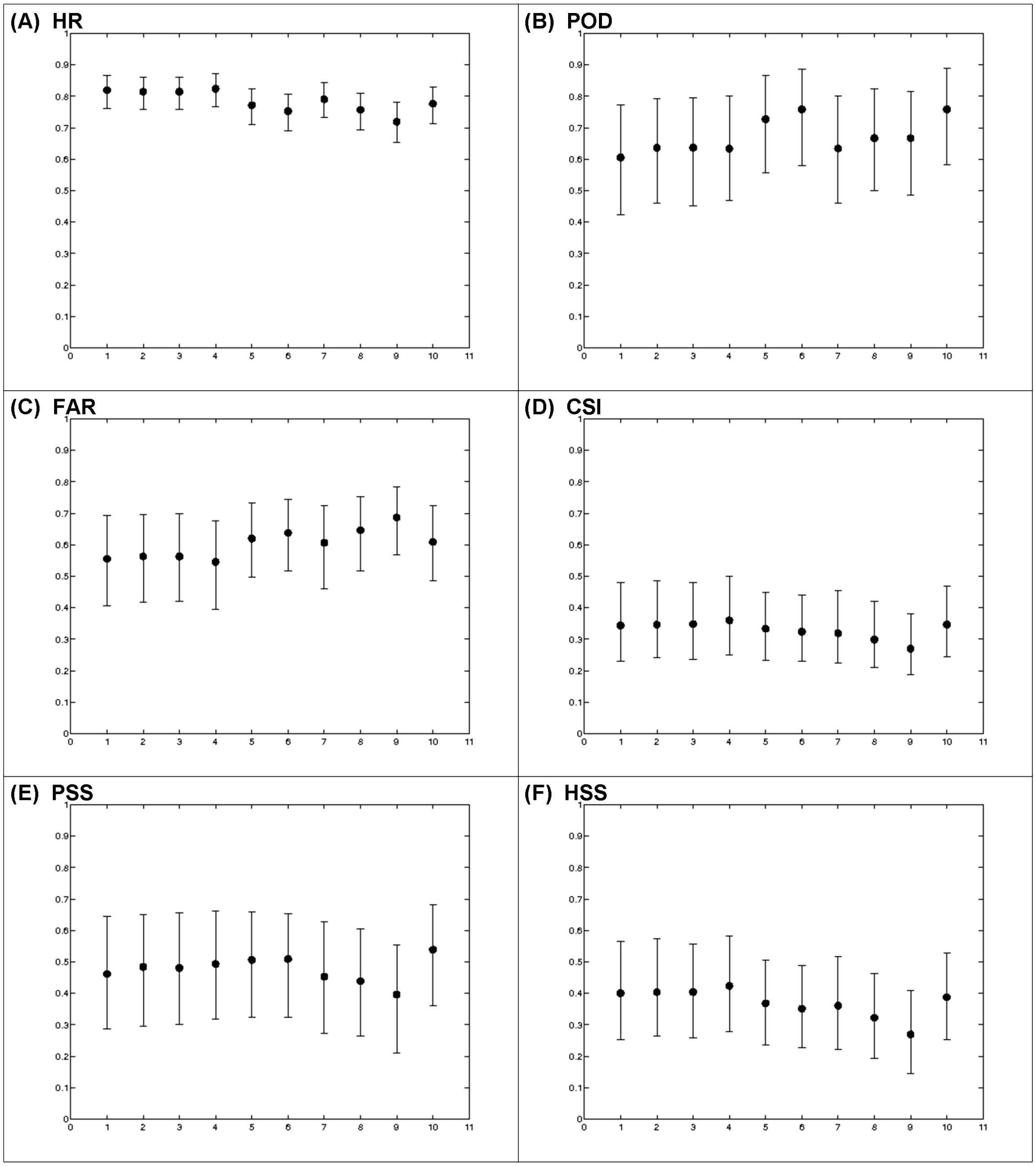


Fig. 2.11. As in Fig. 2.10, using the NNRP dataset.

2) The testing data

Based on the relatively small number of cases in the testing dataset (210) versus the number of cases in the training dataset (630), the uncertainty of the contingency statistics of the testing data was expected to be large. The results of bootstrapping (bias corrected and accelerated; Efron and Tibshirani 1993) the contingency statistics of the testing data using STP fields for the NARR (Fig. 2.10) and NNRP (Fig. 2.11) datasets confirmed this expectation. The 95% confidence intervals (CIs) for HR commonly ranged near 10%, with FARs and CSIs ranging from 20-30%, the skill scores ranging 25-35%, and PODs ranging up to 40%. The limitation of a small sample size for the testing data was exacerbated by the rare-events nature of the dataset, with only approximately 15% of the test cases qualifying as major severe weather outbreaks. The observed variability of the statistical models for POD also was evident for the testing data owing to the small number of major outbreak days in the sample. Unfortunately, S10a showed that decreasing the ratio of training cases to testing cases decreased representativeness, which generally leads to decreased accuracy and skill of the statistical models.

The CIs between the NARR and NNRP datasets for the same analyzed variables clearly overlap in virtually all cases. The medians of the CIs among the statistical algorithms and between the reanalysis datasets are quite close. Uncertainty would have to be very low for statistical significance to be observed. However, as uncertainty is quite large, no statistical algorithm appears to have a statistically significant advantage using the STP field as the meteorological covariate. Additionally, a more fundamental complication may be present here. Based on the sample size concerns discussed in S10a, the uncertainty of the actual means or medians (of the bootstrap samples) of the contingency statistics for the testing data may be large, rendering any conclusions regarding preference of statistical algorithm dubious.

The results observed with STP are consistent with all other severe weather and synoptic meteorological variables tested. Additionally, the variables observed to perform

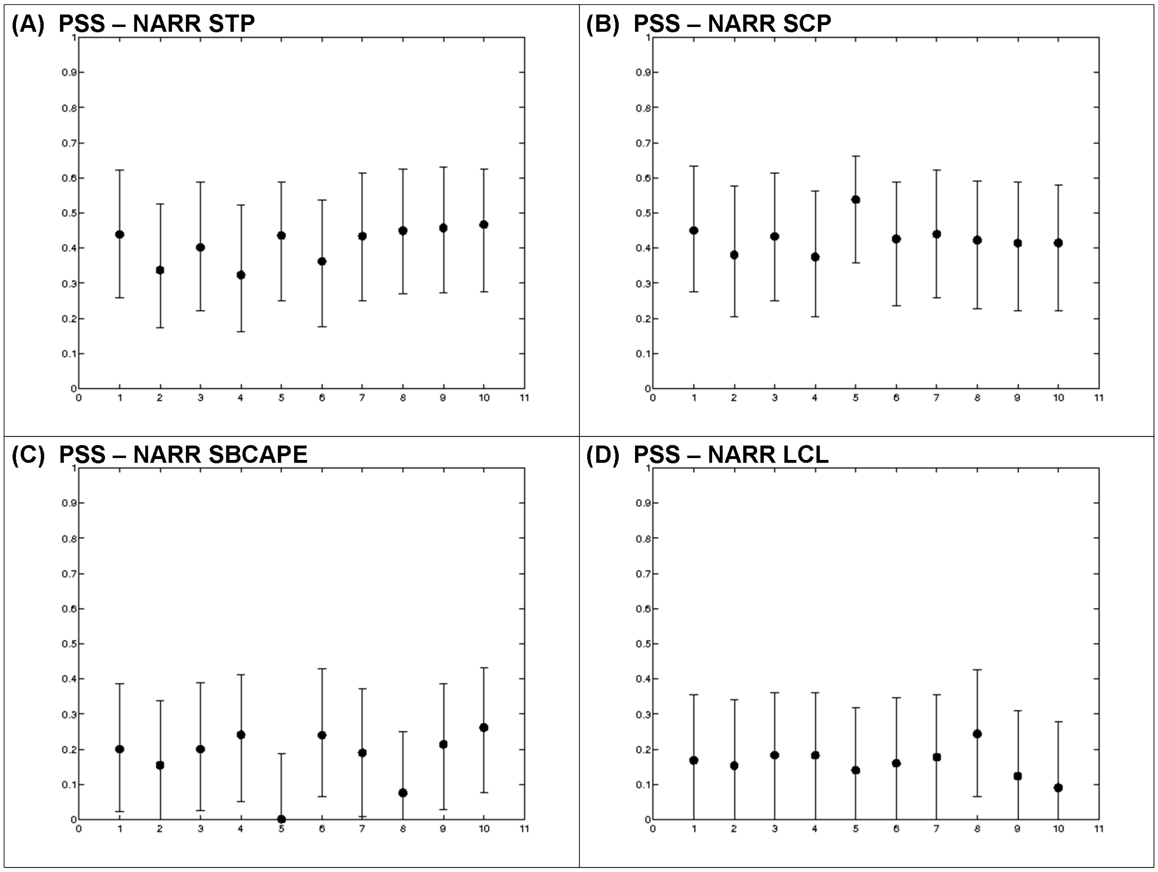


Fig. 2.12. Peirce skill scores using the PCA method on the NARR dataset for (a) STP, (b) SCP, (c) SBCAPE, and (d) LCL fields at the valid times of the outbreaks. Confidence intervals and medians as shown in Fig. 2.10. Statistical algorithms as labeled in Fig. 2.3.

somewhat better when testing the variability of the statistical models were found to perform better when analyzing testing data uncertainty. However, given the large uncertainty in the results, statistical significance was not always observed (e.g., compare PSS CIs for STP, SCP, SBCAPE, and LCL in Fig. 2.12).

3) Comparing NNRP and NARR results using the same grid spacing

An additional method for comparing the NNRP and NARR results includes analyzing subdomains of the same size and grid spacing. The advantage of using a coarse grid means that a large area surrounding the outbreak can be analyzed without additional computational demand; however, comparing grids of the same size and grid

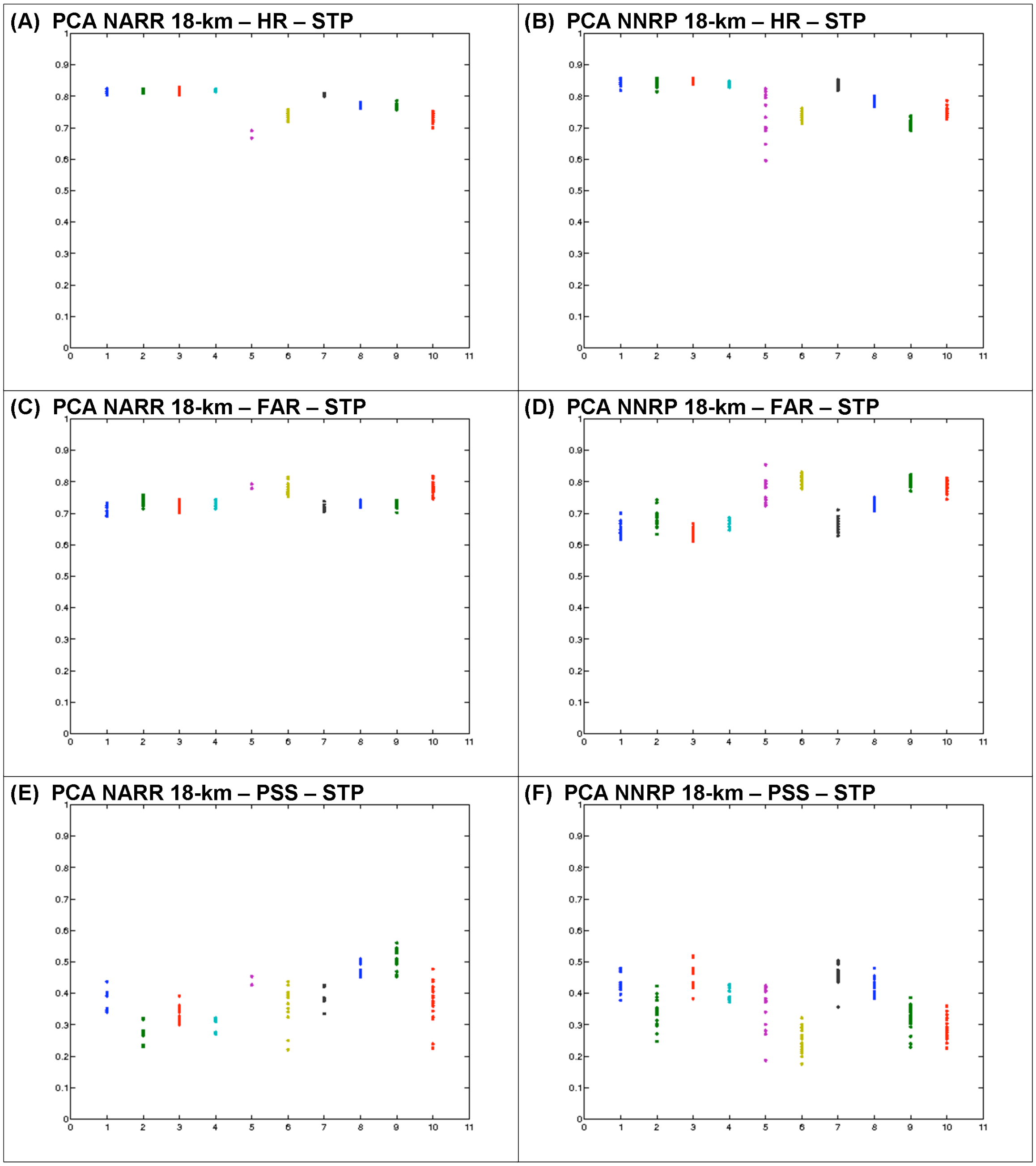


Fig. 2.13. Contingency statistics (labeled) of 25 statistical models developed as subsets of the training data, using the PCA method and STP as the meteorological covariate. For (a), (c), and (e), the NARR 18-km subdomains were analyzed. For (b), (d), and (f), the NNRP 18-km subdomains were analyzed. Statistical values that are repeated are overlain on the previous results.

spacing, using data of lower resolution (NNRP) versus data of higher resolution (NARR) interpolated to the same grid, can provide insight into how much additional beneficial information higher-resolution data provide discriminating outbreaks. For this analysis, the NNRP data were bilinearly interpolated to a 300x200 18-km grid positioned exactly as that of the NARR data, and the subdomain selected for analysis was identical to that of the NARR analysis.

The variability among the training models between the NARR and NNRP datasets was relatively similar (e.g., using STP as the meteorological covariate; Fig. 2.13). Although overall there were slightly larger ranges of the statistics using the NNRP dataset, this was not always true for every statistic and for every statistical algorithm. This comparison also demonstrates the variable relative performance of the statistical algorithms. For example, comparing the SVMs using linear and polynomial kernel functions (models 7-9), skill scores for the NARR data are relatively high for the second- and third-order polynomial kernel functions (models 8 and 9; Fig. 2.13e), whereas their relative utility using NNRP data is reduced (Fig. 2.13f). Finally, the values of the statistics, overall, are quite similar between the datasets – implying little additional discrimination capability using the higher-resolution NARR dataset. Although not shown, this was true for nearly all of the variables analyzed.

Comparison of the 95% CIs of the contingency statistics for the testing data between the NNRP and NARR results using STP show few differences in the medians and the ranges of the CIs (Fig. 2.14). Aside from some differences among the statistical algorithms (e.g., the noticeably worse performance of the decision trees and the SVMs with higher-dimensional kernel functions for the NNRP data, possibly because of noise hampering the performance of nonlinear discrimination techniques), the statistics are overall in good agreement. Although the uncertainty is large for testing data, the results

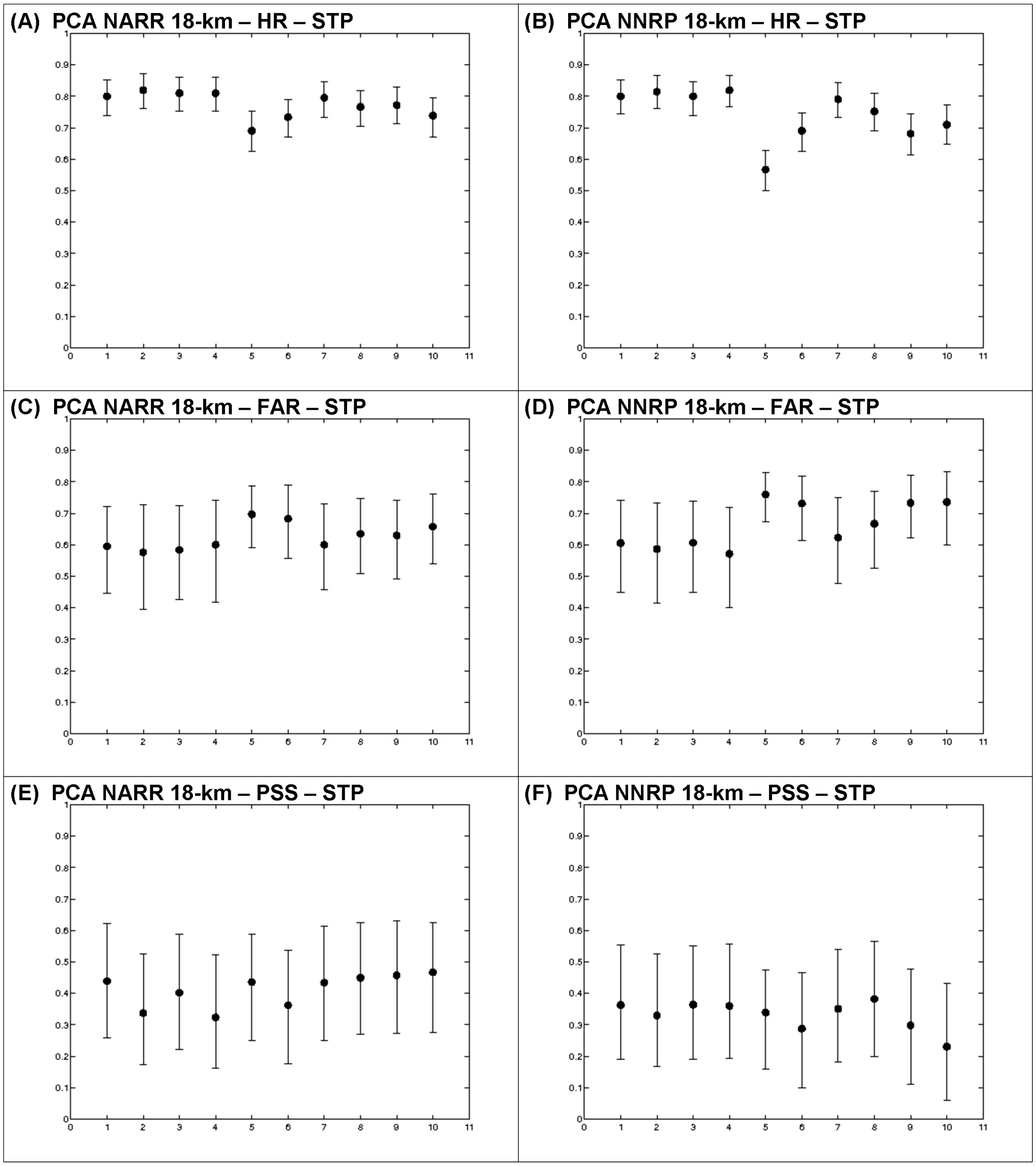


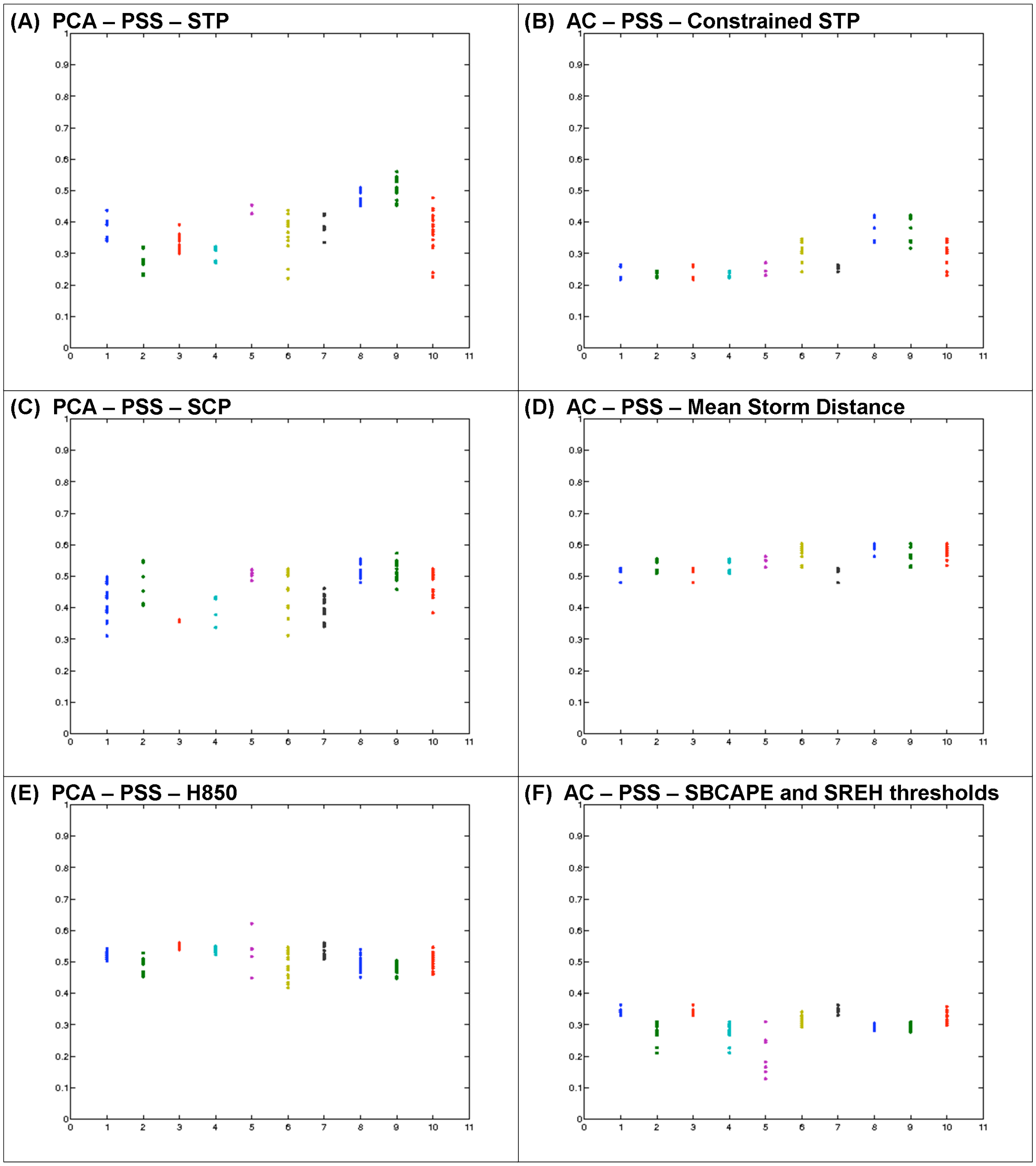
Fig. 2.14. Bootstrap 95% confidence intervals of various contingency statistics of the testing data using the STP field and the principal component analysis technique, when training the 10 statistical algorithms (labeled as in Fig. 2.3, identified in Table 2.1) on the entire set of training cases. For (a), (c), and (e), the NARR dataset was used with 18-km grid spacing. For (b), (d), and (f), the NNRP dataset was used with 18-km grid spacing.

between the NARR and NNRP data are remarkably similar, suggesting there is little benefit in using a high-resolution dataset to identify major severe weather outbreaks using the PCA method.

4) Discussion

These results show no clear advantage of using an enhanced spatial resolution reanalysis dataset *when discriminating major severe weather outbreaks from less significant events using the PCA method*. Obviously, this result may not be true and should not be assumed for other types of outbreak discrimination or other techniques employed to discriminate these events. Additionally, our study only compares NNRP and NARR data. As the types of data included in the two reanalysis datasets are similar (e.g., atmospheric soundings and surface observations) and of relatively large scales, the relative lack of improvement in outbreak discrimination may not be surprising. A higher-resolution reanalysis dataset does not necessarily provide additional fine-scale information if the data incorporated into it cannot resolve finer-scale meteorological processes.

Additionally, the relatively small sample size of the testing data may be a factor in the similar results of the two reanalysis datasets. However, given the relatively similar medians of the 95% CIs for the testing data and the comparable results using 25 statistical models for both of the reanalysis datasets, the findings suggest that differences in discrimination capability are small between the two datasets. As a result, outbreak discrimination may not require high-resolution model initializations. This remains untested in the current work and is beyond the scope of this study. However, past work in outbreak discrimination (S09; M09) certainly supports this possibility.



(A) PCA – PSS – STP

(B) AC – PSS – Constrained STP

(C) PCA – PSS – SCP

(D) AC – PSS – Mean Storm Distance

(E) PCA – PSS – H850

(F) AC – PSS – SBCAPE and SREH thresholds

Fig. 2.15. Peirce skill scores of 25 statistical models on the 210 testing cases, using NARR data at the valid outbreak times, for (a) the PCA method using the STP field, (b) the areal coverage method using the constrained STP (see text for description), (c) the PCA method using the SCP field, (d) the areal coverage method using the mean hypothetical storm distance within regions where STP ≥ 1, (e) the PCA method using 850-hPa geopotential height fields, and (f) the areal coverage method using grid points in which SBCAPE ≥ 1000 J kg^{-1} and 0-1 km SREH ≥ 100 m^2 s^{-2}. Statistical algorithms labeled as in Fig. 2.3.

d. Comparing the PCA and areal coverage methods

Next, the results using the PCA and areal coverage methods are compared, using the same two approaches used in the previous section.

1) The training models

A set of 25 statistical models for each of ten statistical algorithms was compared between the PCA and areal coverage methods. The purpose here is the same as that of Section 2c. Comparisons of the variability of the statistical models and of the overall accuracy and skill when tested on the same 210 cases were conducted. When comparing the PCA and areal coverage methods, the NARR dataset was used exclusively. The objective of this portion of the project was to compare methods of discrimination.

The STP, SCP, and 850-hPa geopotential height fields using the PCA method were compared to various areal coverage computations, in which the N15 index was used to categorize the outbreaks (Fig. 2.15). For the constrained STP (Fig. 2.15b), the sum of the grid points in which STP ≥ 1, SBCAPE ≥ 1000 J kg^{-1}, and 0-1 km SREH ≥ 100 $m^2 s^{-2}$ was computed for each outbreak day at the valid time of the outbreak. For the mean storm distance (Fig. 2.15d), the average distance traversed by a hypothetical storm over each grid point within the region in which STP ≥ 1 was computed. For the SBCAPE and SREH constraints (Fig. 2.15f), the sum of the grid points in which SBCAPE ≥ 1000 J kg^{-1} and SREH ≥ 100 $m^2 s^{-2}$ was computed.[10]

Comparison of the PSSs for 25 statistical models using the PCA and areal coverage methods indicated the following: (1) There were some meteorological variables in which the areal coverage method performed similarly to the PCA method.

[10] Other areal coverage thresholds were tested besides those discussed in this section. Results using different thresholds (not shown) generally were similar to or worse than those discussed in the paper.

For example, using mean hypothetical storm distance within the region in which STP ≥ 1 resulted in similar PSSs to the SCP and 850-hPa geopotential height fields (cf. Figs. 2.15c-e). (2) The variability of the statistics for the 25 training models is generally similar or lower for the areal coverage method than for the PCA method. (3) As with the PCA method, statistical model results are not always similar, and relative performance can vary with meteorological parameter selection. For example, the higher-order polynomial kernels using the SVM statistical algorithm have noticeably higher PSSs for the constrained STP compared to the other algorithms (Fig. 2.15b). This was not the case, however, when using SBCAPE and SREH thresholds in combination (Fig. 2.15f). These three findings were true for various contingency statistics and for a number of meteorological covariates.

2) The testing data

As observed comparing NARR and NNRP datasets using the PCA method, bootstrapping the contingency statistics of the testing data for the PCA and areal coverage methods resulted in large 95% CIs as a result of the relatively small sample of cases (e.g., Fig. 2.16). However, results among the best performing meteorological parameters were similar between the two techniques, for a variety of contingency statistics. Comparisons of the STP fields using the PCA method versus the mean hypothetical storm distances within the region in which STP ≥ 1 found accuracy (HR; Figs. 2.16a,b) and skill scores (PSS; Figs. 2.16e,f) that were reasonably close, well within the CIs of the alternative technique. Discrepancies in POD (not shown) and FAR (Figs. 2.16c,d) were somewhat larger, although the median statistics generally remained within the CIs of the other method. Once again, the POD and, to a lesser extent, the FAR were

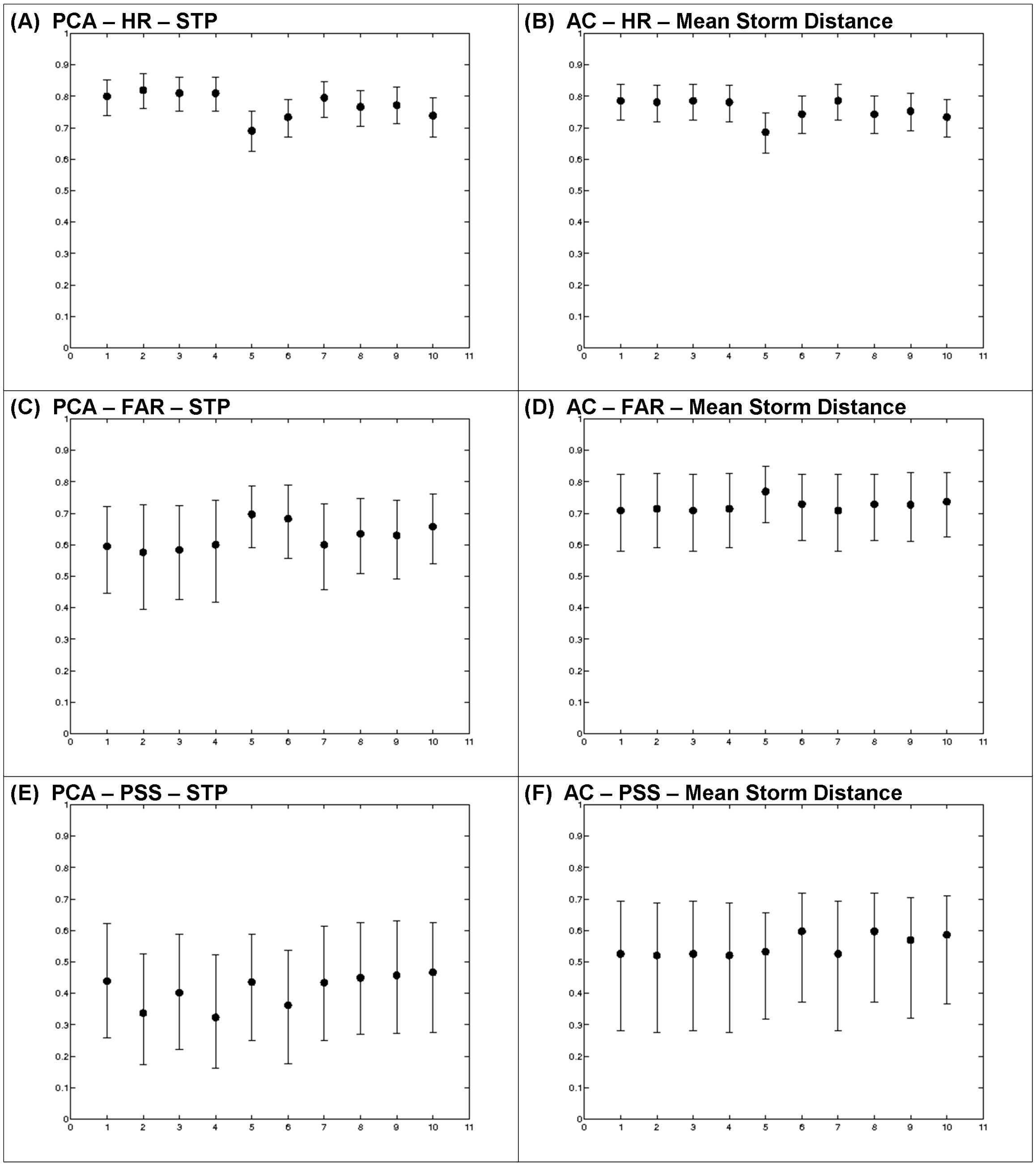


Fig. 2.16. Bootstrap 95% confidence intervals of various contingency statistics of the testing data, when training the 10 statistical models (labeled as in Fig. 2.3 identified in Table 2.1) on the entire set of training cases. The analysis technique, contingency statistics, and variables selected are labeled as in Fig 2.15.

subject to relatively large variation because of the rare-events nature of the dataset, in addition to the relatively small sample of testing cases.

3) Discussion

Comparisons of the PCA and areal coverage technique revealed that there was no obvious preference for choice of technique. The areal coverage method appeared to have somewhat less variation in the statistical models compared to the PCA method in general, which is a desirable characteristic. However, some statistical algorithms with some of the meteorological fields analyzed using the PCA technique featured comparably low variability. Additionally, the best contingency statistics using the two techniques were quite similar, despite the relatively large CIs associated with the limited sample available for the testing data.

Given the relatively comparable performance between the two techniques, there is little evidence at this point to encourage the use of the relatively complicated PCA method over the simpler, easily interpretable, and computationally less demanding areal coverage method. These results agree with previous research investigating discrimination of tornadic and nontornadic outbreaks (S09; M09). Specifically, the subjective technique developed by S09 used areal coverage as a means of distinguishing the events, and the PCA method introduced by M09 produced similar results. The reason for this is obvious from inspection of Fig. 2.2: the undersampling of a sufficient number of severe events in the two tails of the distribution to yield meaningful coherent patterns of variation that are unique to a specific type of severe weather, a basic requirement of PCA. As Figure 2 in S10a shows, the areal coverage of parameters favorable for severe weather for the highest-ranked severe weather outbreaks tends to be large and, as a group of events, tends to have smaller variance than that for the combination of major severe weather outbreaks and less significant events. However, the small sample of major severe weather outbreaks available for training and testing, using both areal coverage and the PCA method, leads to the possibility of underrepresentation of the environments associated with major severe weather outbreaks as well as a clear false alarm problem with either technique (see Figs. 2.16c,d). This important finding of insufficient sampling is consistent with previous work (S10a,b) that discrimination analyses to date have been undersampled for each category.

Note that the index scores greater than 1 or less than -1 occupy only a small fraction of events in this dataset. Ideally, for discrimination purposes, the N15 curve would have two relative plateaus of index values in the tails with a sharp descent in the center (i.e., comparable to the prototypical "logit" curve – http://en.wikipedia.org/wiki/Logistic_function). However, the N15 curve has a steep descent in each of the tails and a broad center, indicating no two events in the tails were similar. This shows that the variance of severe weather report variables, used to rank the severe weather outbreaks in SD10a, is highest with the highest- and lowest-ranked cases and relatively minimal for the intermediate events, which agrees with subjective analyses described in D06 and SD10a. However, this is not observed for severe weather *environments*. The nature of the N15 index curve suggests that classification of major severe weather outbreaks and null events is subject to uncertainty. Section 4 will approach the discrimination problem differently by varying the threshold used to classify events as major severe weather outbreaks.

As discussed in SD10a, the lower tail of the rankings for the N15 index (and for all of the ranking indices developed in their study) is associated with outbreak days with spatially distinct clusters of severe reports and/or large geographic scatter in the reports.

A new technique to eliminate these cases from consideration has been introduced (Shafer and Doswell 2010b; see also Section 3 herein). The results of this study modified the scores of the ranking indices by providing one of the two plateaus necessary for outbreak discrimination (removing the lower tail). However, major severe weather outbreaks are rare events; thus, even for a period as long as 1960-2008 (as is used in Shafer and Doswell 2010b), developing a ranking scheme with two plateaus is a very challenging task with the data currently available. The relative severity of, e.g., the 3-4 April 1974 tornado outbreak (Corfidi et al. 2010) compared to the 3 May 1999 tornado outbreak (Thompson and Edwards 2000), and the exceedingly rare nature of these events in the past 50+ years suggests such a ranking scheme is unlikely to agree with subjective notions regarding major severe weather outbreaks. The fact than an event comparable to that of 3-4 April 1974 has not occurred since then is a clear indication of the undersampling problem. Therefore, the task of discriminating severe weather outbreaks likely will continue to be difficult for the foreseeable future and will depend on improved observations and increased physical understanding of these events.

e. Reasons for and implications of the findings

This study compared two reanalysis datasets and two techniques for discriminating major severe weather outbreaks from less significant events. The NNRP dataset, with horizontal and vertical grid spacing consistent with the synoptic scale, was compared to the NARR dataset, with grid spacing consistent with subsynoptic-scale phenomena. The PCA method, a data mining technique that examines meteorological fields in the vicinity of the outbreak, was compared to the areal coverage technique, which computes the size of the region considered favorable for significant severe weather. Our findings show that there is no statistically significant advantage to using subsynoptic data to distinguish major severe weather outbreaks from less significant events. Furthermore, the computationally less demanding, easily interpretable, and less complex areal coverage technique performs as well as the PCA method. The potential trade-off of using the simpler method is the loss of information on the gradients of the fields being analyzed. However, in cases where there are not sufficient samples to identify reliable structures (e.g., in the present work), the simpler areal coverage technique is preferred.

Major tornado outbreaks are virtually always associated with strong synoptic-scale systems, and the environments in which these events unfold commonly fit into a small set of conceptual map types (e.g., Miller 1972; S09). These "synoptically evident" events (Doswell et al. 1993) typically are associated with large regions of thermodynamic instability and vertical wind shear supportive of significant severe weather (Brooks et al. 2003b; Hamill et al. 2005). Large regions of favorable severe weather parameters are easy to diagnose using coarse datasets (e.g., see Hamill et al. 2005), rendering identification of these days possible using the areal coverage method or with coarse grids using the PCA method.

Physical understanding regarding tornadogenesis remains elusive. Though we have associated tornado outbreaks with widespread regions of sufficient thermodynamic instability and vertical wind shear, other factors are associated with the occurrence or nonoccurrence of tornado outbreaks. In particular, the effects of mesoscale boundaries, convective inhibition, and boundary layer moisture content (among other factors) have been examined and continue to be the focus of current research. Their specific roles in tornadogenesis remain unclear, however, and the analysis and prediction of these fields (especially subsynoptic fields) is imperfect. Furthermore, tornado outbreaks are rare events, and most studies have focused necessarily on conditions enhancing rather than mitigating their occurrence (Doswell et al. 2002). Such focus on “events” versus “nonevents” may lead to an incomplete understanding of what diagnostic variables to analyze in order to distinguish these events more accurately and skillfully.

One obvious result of the work by S10a and of this study is that the detection of major severe weather outbreaks typically is associated with a substantial number of false alarms, agreeing with past work investigating storm environments by Rasmussen and Blanchard (1998), Doswell and Evans (2003), Thompson et al. (2003), Hamill et al. (2005), and others. S10a identified particular synoptic patterns that resulted in large regions of sufficient CAPE and wind shear but failed to produce a large number of significant tornadoes (their Section 5). Such synoptic-scale map types sometimes feature midlevel flow oriented parallel to a surface boundary. In the case of zonally-oriented midlevel flow and surface boundaries (e.g., Fig. 2.17), one or two supercells typically develop along the (warm or stationary) front – possibly producing tornadoes – before dissipating outside of the favorable region or developing upscale into a mesoscale convective system. In the case of meridionally-oriented midlevel flow and surface boundaries, rapid upscale growth to a squall line is common (e.g., Fig. 2.18; see also Dial et al. 2010). Nevertheless, a substantial portion of false alarms feature synoptic patterns and mesoscale fields similar to major tornado outbreaks (S10a, their Fig. 17). The reasons for the failure to produce a major tornado outbreak in such cases remain elusive.

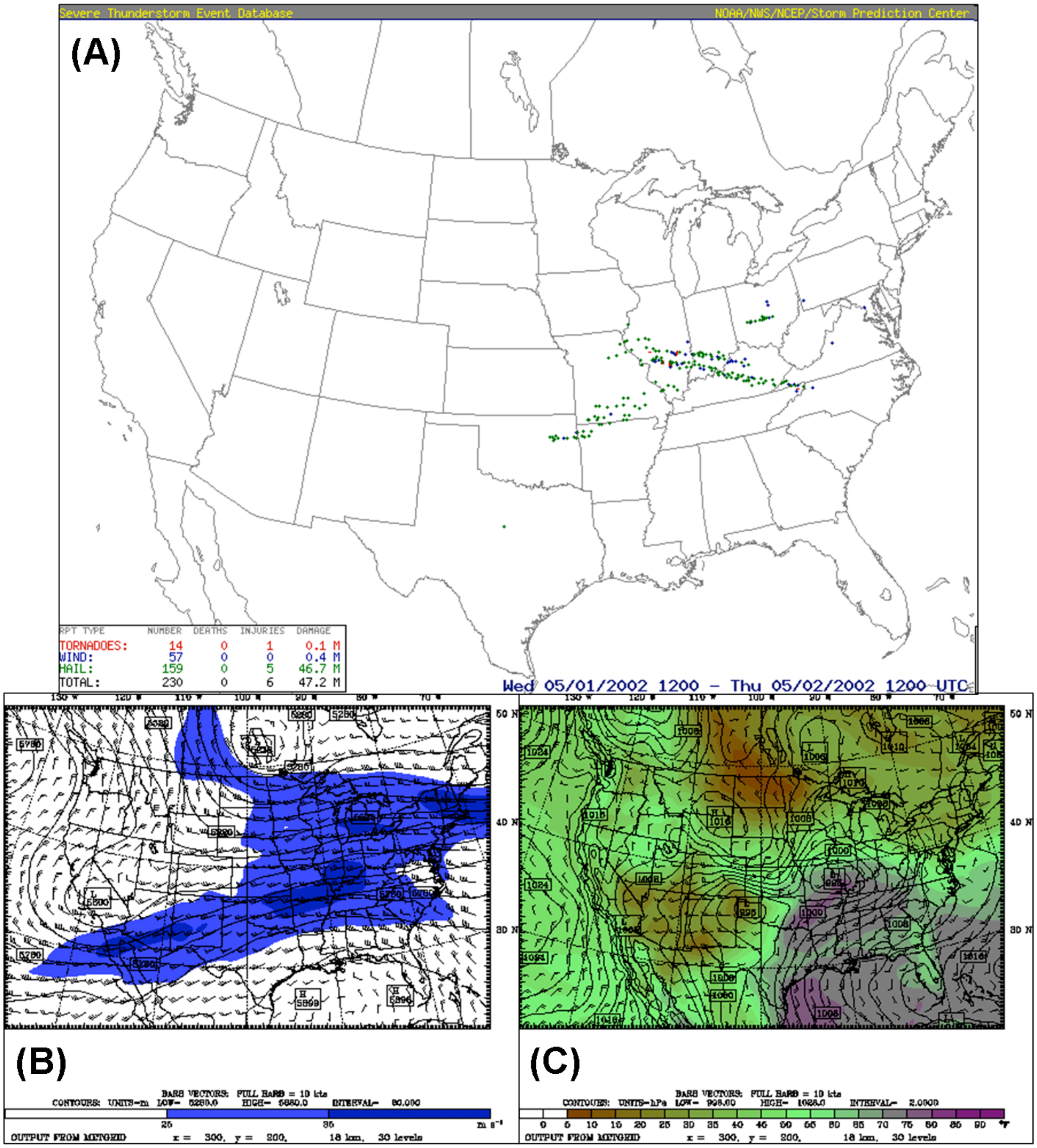

Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(A)
RPT TYPE NUMBER DEATHS INJURIES DAMAGE
TORNADOES: 14 0 1 0.1 M
WIND: 57 0 0 0.4 M
HAIL: 159 0 5 46.7 M
TOTAL: 230 0 6 47.2 M
Wed 05/01/2002 1200 - Thu 05/02/2002 1200 UTC
(B)
BARB VECTORS: FULL BARB = 10 kts
CONTOURS: UNITS=m LOW= 5280.0 HIGH= 5880.0 INTERVAL= 60.000
25
35
m s⁻¹
OUTPUT FROM METGRID x = 300, y = 200, 18 km, 30 levels
(C)
BARB VECTORS: FULL BARB = 10 kts
CONTOURS: UNITS=hPa LOW= 998.00 HIGH= 1028.0 INTERVAL= 2.0000
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 °F
OUTPUT FROM METGRID x = 300, y = 200, 18 km, 30 levels

Fig. 2.17. (a) Severe reports for the 1 May 2002 outbreak day. Tornado reports as red dashes or lines, and hail (wind) reports in green (blue) dots. (b) Plot of 500-hPa winds (speeds above 25 m s^{-1} in filled contours by increments of 10 m s^{-1}; barbs in kts) and geopotential heights (contours in m) valid at 0000 UTC 2 May 2002 using NARR data. (c) Surface dew point temperature (filled contours in ºF), winds (barbs in kts), and mean sea level pressure (contours in hPa) valid at 0000 UTC 2 May 2002 using NARR data.

Therefore, future work must investigate the differences between the major severe weather outbreaks and the false alarm cases that are similar in terms of synoptic map type

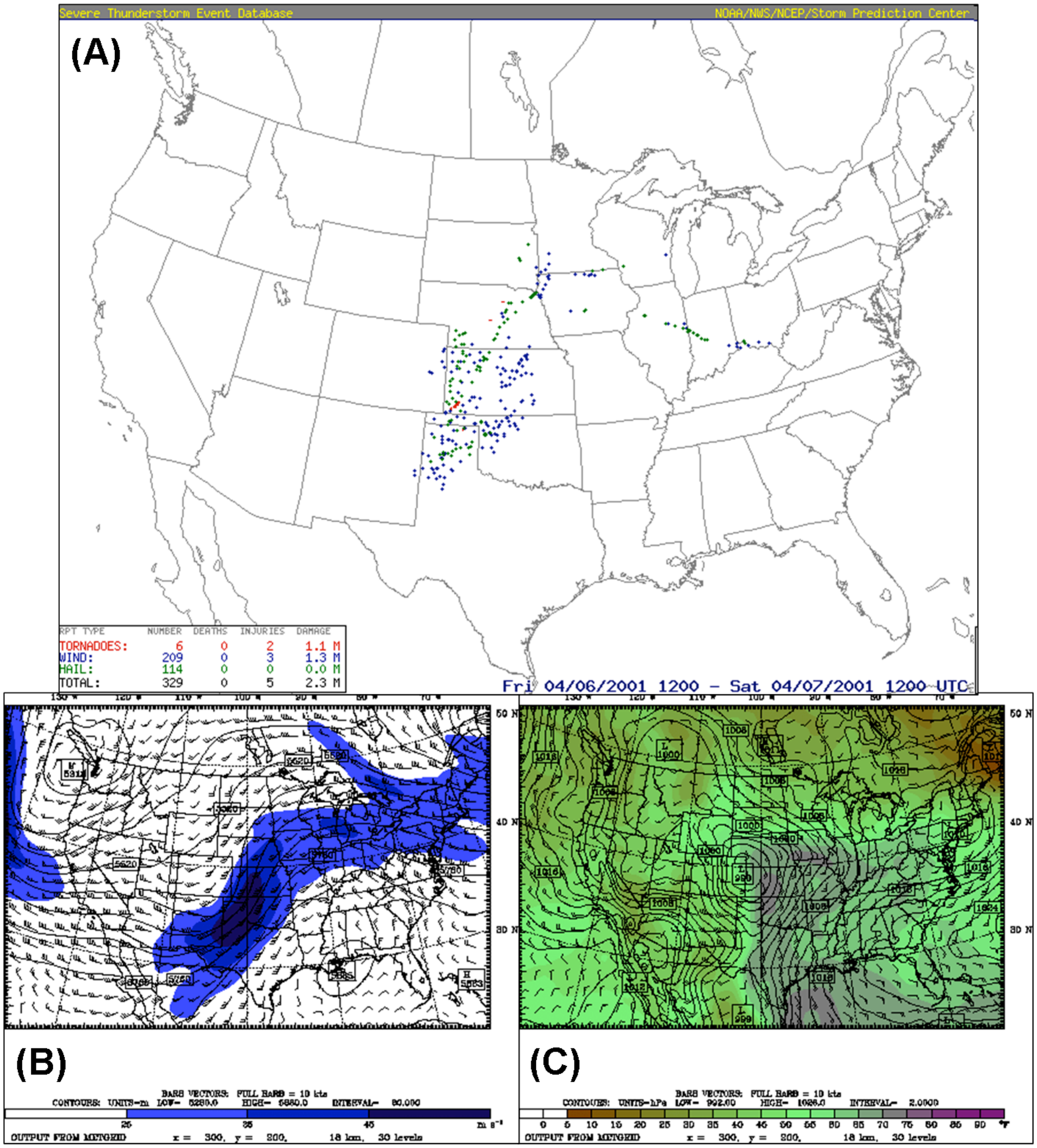

Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(A)
RPT TYPE NUMBER DEATHS INJURIES DAMAGE
TORNADOES: 6 0 2 1.1 M
WIND: 209 0 3 1.3 M
HAIL: 114 0 0 0.0 M
TOTAL: 329 0 5 2.3 M
Fri 04/06/2001 1200 - Sat 04/07/2001 1200 UTC
(B)
BARB VECTORS: FULL BARB = 10 kts
CONTOURS: UNITS=m LOW= 5280.0 HIGH= 5880.0 INTERVAL= 60.000
25 35 45 m s-1
OUTPUT FROM MM5/GRID x = 300, y = 200, 18 km, 30 levels
(C)
BARB VECTORS: FULL BARB = 10 kts
CONTOURS: UNITS=hPa LOW= 992.00 HIGH= 1026.0 INTERVAL= 2.0000
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 °F
OUTPUT FROM MM5/GRID x = 300, y = 200, 18 km, 30 levels
50 N
40 N
30 N

Fig. 2.18. As in Fig. 2.17, for the 6 April 2001 outbreak day.

and fields of diagnostic meteorological parameters. Such research should investigate the meteorological processes that *inhibit* tornado formation, despite the presence of seemingly favorable environmental conditions. The capability to "analyze" these processes using current or newly-proposed diagnostic variables should be addressed in such studies. Additionally, the identification of systematic differences in the synoptic environments *preceding* events and null cases and the modification of precursor *composite* synoptic fields (as in Mercer et al. 2010) in idealized model simulations to determine what changes bring about the occurrence or absence of widespread tornado formation is essential for our increased understanding of these events. Finally, the improvement of techniques to identify outbreak days in the hopes of increasing sample size, as outlined in SD10a, is necessary. That is the focus of the following section.

3. Using kernel density estimation to identify, rank, and classify severe weather outbreak events

a. Introduction

Recent studies have attempted to rank severe weather events using archived reports of tornadoes, severe winds, wind damage, and hail for the purposes of identifying prototypical tornado and primarily nontornadic outbreaks (D06) and for determining the relative severity of outbreaks of any type (SD10a). Both studies used linear-weighted multivariate indices to rank cases that met initial criteria for their inclusion (e.g., a day in which seven or more tornadoes occurred was considered for the ranking of tornado outbreaks in D06). These studies resulted in rankings of severe weather events that agreed with subjective notions, were reproducible, and turned out to be relatively robust to modifications of the weights for the multiple variables used to rank the cases.

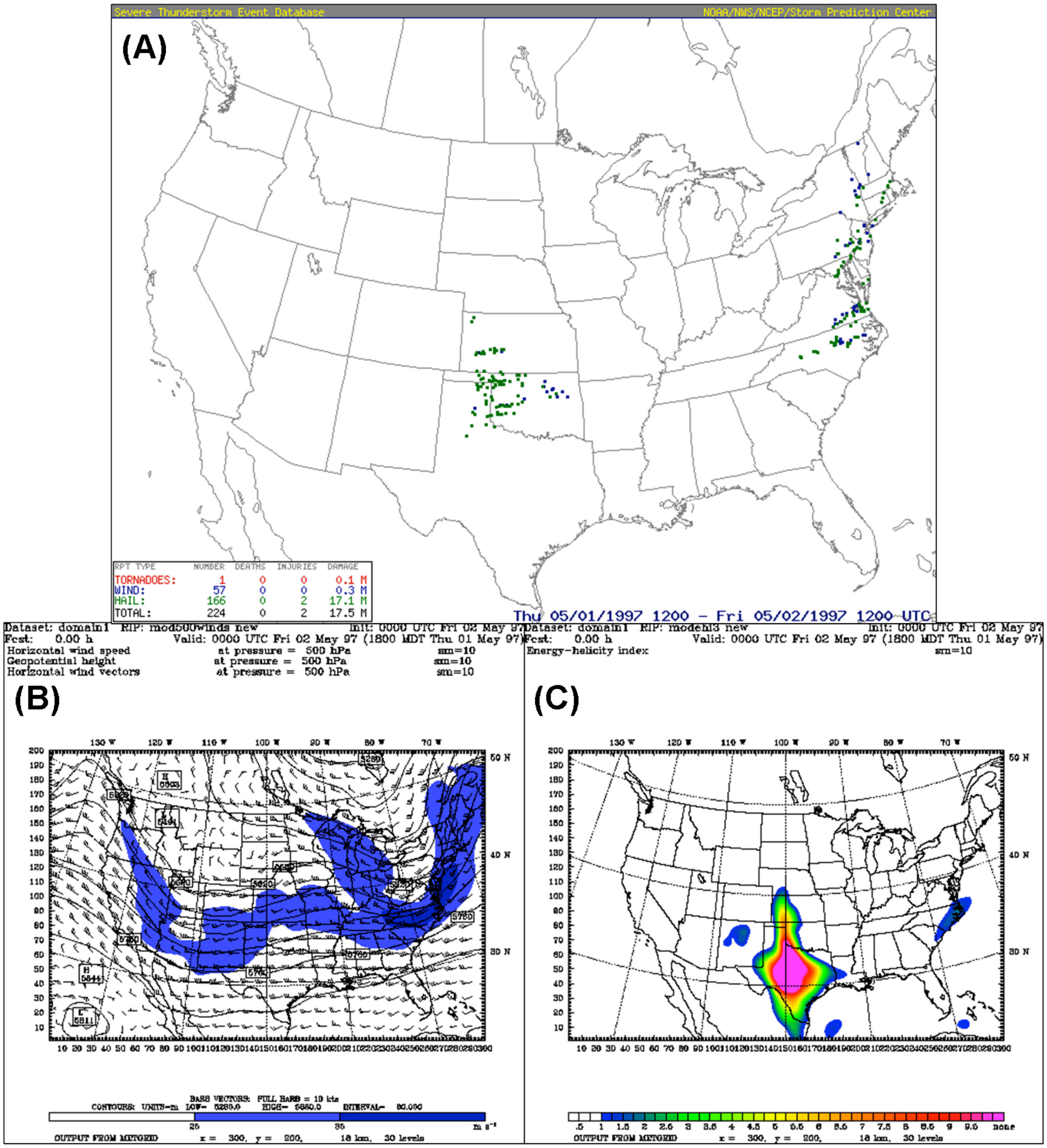
A complicating factor in the ranking of severe weather outbreaks is the presence of large geographic scatter of the reports on a subset of the days considered. Such scatter can manifest itself in various ways (see Fig. 2 in SD10a). For example, some days feature widely dispersed reports of severe weather throughout the United States. Other days consist of multiple clusters of severe reports separated by large areas of little or no observed severe weather. Some days exhibit a combination of the two effects, with a cluster of severe reports and a large number of widely dispersed reports separated from the cluster of reports.

As severe weather outbreaks generally are perceived to consist of a large number of severe reports over a geographically compact region, accounting for days featuring large geographic scatter is critical for the identification of prototypical outbreak days or for the ranking of these events in a way that agrees with these subjective perceptions. Elimination of these days based solely on the number of reports (as in SD10a) is ineffective, as many days exhibiting such large geographic scatter also comprise a large number of severe reports.

D06 introduced a method to account for large geographic scatter, using the distributions of the latitudes and longitudes of the reports. For both latitude and longitude, the middle 50% of the distribution (i.e., between the 25th and 75th percentiles) was used as a range of latitude and longitude (SD10a, their Fig. 3). This results in a latitude/longitude "box", the area of which can be parameterized by the product of the latitude-longitude ranges. A large value suggests substantial geographic scatter, whereas a small value suggests limited geographic scatter (as suggested by SD10a, their Fig. 3). D06 defined this as the middle-50% parameter, which was found to be effective in eliminating appropriate cases from the top rankings of primarily nontornadic outbreaks (D06) and from the major and intermediate outbreak days (SD10a).

However, both studies raised questions about the middle-50% parameter's utility. Specifically, on days with multiple geographically-separated clusters of reports, the technique treated all of these clusters as one outbreak. Typically, such days feature multiple synoptic-scale systems, indicating these events should be considered separately (e.g., Fig. 3.1). The AMS glossary, for example, states tornado outbreaks are associated with a single synoptic-scale system (Glickman 2000), rendering the combination of separate clusters as single events undesirable.

The purpose of this study is to consider severe weather events based on clusters of severe reports on a given day, rather than based on the 24-h period alone, to rank these



Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(A)
RPT TYPE NUMBER DEATHS INJURIES DAMAGE
TORNADOES: 1 0 0 0.1 M
WIND: 57 0 0 0.3 M
HAIL: 166 0 2 17.1 M
TOTAL: 224 0 2 17.5 M
Thu 05/01/1997 1200 - Fri 05/02/1997 1200 UTC
Dataset: domain1 RIP: mod500winds new
Init: 0000 UTC Fri 02 May 97
Fcst: 0.00 h
Valid: 0000 UTC Fri 02 May 97 (1800 MDT Thu 01 May 97)
Horizontal wind speed at pressure = 500 hPa sm=10
Geopotential height at pressure = 500 hPa sm=10
Horizontal wind vectors at pressure = 500 hPa sm=10
(B)
130 W
120 W
110 W
100 W
90 W
80 W
70 W
50 N
40 N
30 N
BARB VECTORS: FULL BARB = 10 kts
CONTOURS: UNITS=m LOW= 5280.0 HIGH= 5880.0 INTERVAL= 60.000
25
35
m s⁻¹
OUTPUT FROM METGRID x = 300, y = 200, 18 km, 30 levels
Dataset: domain1 RIP: modehi3 new
Init: 0000 UTC Fri 02 May 97
Fcst: 0.00 h
Valid: 0000 UTC Fri 02 May 97 (1800 MDT Thu 01 May 97)
Energy-helicity index sm=10
(C)
.5 1 1.5 2 2.5 3 3.5 4 4.5 5 5.5 6 6.5 7 7.5 8 8.5 9 9.5 none
OUTPUT FROM METGRID x = 300, y = 200, 18 km, 30 levels

Fig. 3.1. (a) Severe reports on 1 May 1997, with severe wind gusts or wind damage in blue, severe hail in green, and tornadoes in red. (b) North American Regional Reanalysis (NARR) 500-hPa wind speeds (filled contours in m s^{-1}), winds (barbs in kts), and geopotential heights (contours in m) valid at 0000 UTC 2 May 1997. (c) NARR 0-3 km energy helicity index (EHI) valid at 0000 UTC 2 May 1997.

events based on relative severity, and to classify these events based on the characteristics of the severe weather reports associated with the particular cluster. Section 3.2 describes the data and methods used to identify, rank, and classify these severe weather events.

Section 3.3 demonstrates the characteristics of the techniques on various types of severe weather report clusters. Section 3.4 details the results of the rankings of the severe weather events. Section 3.5 presents the findings when classifying the severe weather events. Section 3.6 summarizes the study and discusses remaining issues associated with the current work. This work is also discussed in Shafer and Doswell (2010b).

b. Data and methods

As in D06 and SD10a, the Storm Prediction Center severe weather database (Schaefer and Edwards 1999) was used to obtain the severe reports on each day from 1960-2008. The database includes information on the type of report (tornado, hail, or straight-line wind), the intensity (e.g., hail size or wind speed) or Fujita-scale rating, and various geographic and societal aspects of the reports (e.g., location or track, number of casualties, etc.). The variables considered when ranking the outbreaks were the same as those in SD10a (their Table 3.1), except for the middle-50% parameter.

Each 24-h period from 1 January 1960 to 31 December 2008 was considered separately. The period of consideration was 1200 UTC on the nominal date to 1159 UTC the following day. Any severe weather event that continued past 1200 UTC on the following day, perhaps for multiple days, was considered separately, though these events were rare in our dataset.

As the goal of this work was to consider clusters of severe reports as a severe weather outbreak, rather than just the outbreak day, a technique for overcoming the limitations of the middle-50% parameter was necessary. The use of kernel density estimation (KDE; Bowman and Azzalini 1997) was employed for this particular purpose.

KDE approximates the probability density function at a particular point. Specifically, a one-dimensional KDE can be represented as the following:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} K_h\left(x - x_i\right) \qquad (3.1),$$

where n is the number of severe reports on a given day, K_h is a kernel function, and h is a tunable smoothing parameter (bandwidth). Typically, the kernel function implemented is Gaussian (e.g., Brooks et al. 1998), and that is the case for this study:

$$K_h\left(x - x_i\right) = \frac{1}{h\sqrt{2\pi}}\exp\left[\frac{-(x - x_i)^2}{2h^2}\right] \qquad (3.2).$$

It can be shown that for multivariate KDE, (6.1) can be represented as:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n}\left[\prod_{k=1}^{d} K_{h_k}\left(x^{(k)} - x_i^{(k)}\right)\right] \qquad (3.3),$$

where d is the number of dimensions. For this study, d was 2, as the severe reports are reported as latitudes and longitudes. The bandwidth (which can be different for each dimension, but was not in this study) and the threshold value of the approximated probability density function (PDF) can be used to determine the reports associated with a particular geographic cluster. Because d is 2, Eqn. (3.2) is modified for two dimensions by taking the square of itself [as suggested by Eqn. (3.3)], such that the quantity $(x - x_i)$ becomes a two-dimensional distance.

The observed reports for a given day either were associated with a grid point for various map projections using objective analysis techniques (as in Brooks et al. 1998), or were computed as distances from all of the grid points for a particular map projection directly. Thus, the distance quantity in Eqns. (3.2)-(3.3) either was defined in terms of grid point separation or in terms of actual distance. The bandwidth is a measure of the uncertainty associated with these distances (see Brooks et al. 1998) and requires modification based on the technique used to identify the clusters of severe reports. As Section 3.3 will show, differences among the techniques and map projections were minor, as expected, so long as the bandwidth and PDF thresholds were modified accordingly.[11]

If the reports were converted to a grid initially (i.e., positioned at the relevant grid point that encompasses that location), the KDE was computed for each of the points on the same map projection the report locations were converted to, and contours of the KDE were drawn on these projections. This will be referred to as the *grid point method* henceforth. On the other hand, if the observations were not converted to a grid initially, the distances from each grid point of a map projection (of which various types were considered) to each severe report were calculated, and contours of the KDE were computed on these projections. This will be referred to as the *distance method* hereafter. After selection of bandwidth and PDF threshold value, any point falling on or within the contour was considered to be associated with the cluster of severe reports.

Modification of a map projection's grid spacing could be accounted for by selecting different values of bandwidth and probability thresholds. This was unnecessary, however, if the quantity $(x - x_i)$ was measured in terms of latitude and longitude, or in terms of direct distances, since the grid spacing would not affect these values for grid points of various size at the same location. Therefore, the choice of grid spacing for a map projection is essentially arbitrary[12], though relatively coarse grid

[11] Note that the PDF threshold must be modified if the bandwidth is modified, as $f(x)$ is a function of the bandwidth [see Eqs. (3.1)-(3.3)].

[12] This is true so long as the grid spacing is low enough to resolve the severe reports for adequate representation.

spacing is preferred because of reduced computational demand. For the latitude-longitude map projection, 1^{o} grid spacing was used.

After all of the clusters for the 49-year period were identified, a subset of these cases was removed to eliminate those events with a relatively small number of reports or relatively sparse coverage within the region determined to be associated with the event. This was done by calculating two variables for each cluster considered: the total number of reports within a cluster, and the ratio of reports to grid points associated with the cluster (hereafter, the *density ratio*). A cluster was removed from consideration if the total number of reports within the region associated with the cluster was below the detrended mean value for all of the clusters for that particular year, or if the density ratio for the particular cluster was below the detrended mean value for all of the clusters for that particular year.

As the total number of severe reports in a given year substantially increases from 1960 to 2008 (see Brooks et al. 2003a; Doswell et al. 2005; D06; Verbout et al. 2006; SD10a), the annual means of the two variables were detrended to account for these nonmeteorological artifacts. The process of detrending was the same as that incorporated by D06 and SD10a: a linear regression to the logarithm of the annual means was computed for each of the variables. The detrended annual mean is the value of the regression curve for the relevant year. Based on the small values and exponential increase of the detrended means over the 49-yr period (e.g., Fig. 3.2), a large number of the clusters on a given year featured a very small number of reports and/or sparse coverage of the reports within the event region. Unsurprisingly, the number of clusters on a given year increases from 1960-2008 (not shown), which means the number of cases

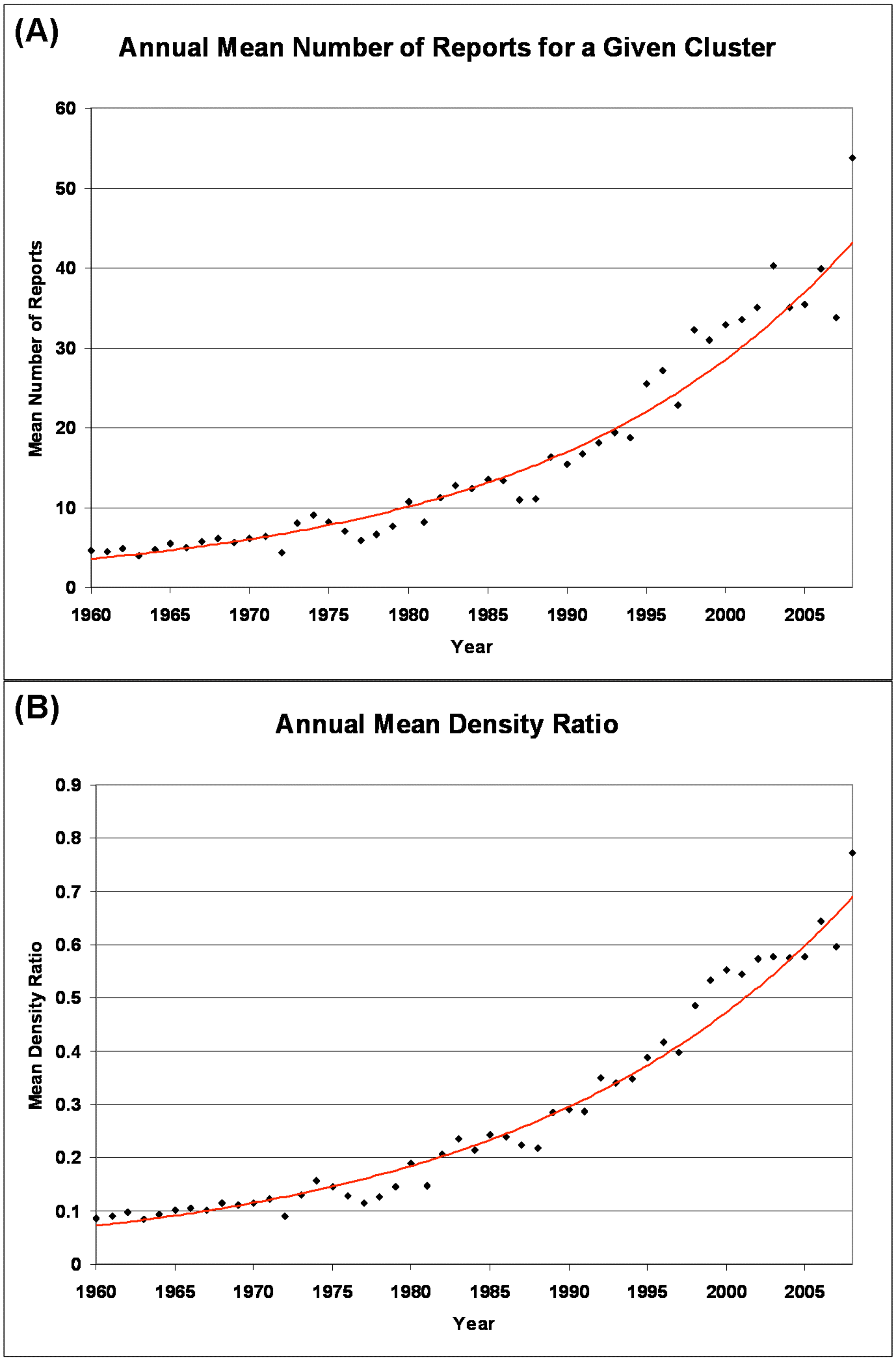


Fig. 3.2. Examples of detrending for the annual mean number of reports for a given cluster (a) and for the annual mean density ratio (b), when considering all of the clusters for a particular year. The results are for a latitude-longitude map projection with 1° grid spacing spanning the conterminous United States, using a bandwidth of unity for each dimension, and a threshold probability of 0.001 for a grid point to be associated with the severe weather event.

considered increases for more recent years. This is a result of the relative lack of reporting, particularly of nontornadic severe weather or of relatively minor severe weather events, in the early years of the period included in this study. Thus, although attempts to account in a relatively simple way for secular changes in the dataset were implemented, *some impact from the changes in reporting is inevitable*.

For the remaining cases, annual sums of the severe weather reports included in the linear-weighted, multivariate indices used to rank and classify the outbreaks (SD10a, their Table 1) subsequently were tabulated. These sums were divided by the number of clusters for the relevant year. These "cluster means" then were detrended, if necessary, in the same manner as in D06 and SD10a (examples in Fig. 3.3). As the values for each of the variables included in the indices can have markedly different magnitudes, all variables were standardized (transformed to have zero mean and a standard deviation of unity):

$$\tilde{x}_i^{(j)} = \frac{x_i^{(j)} - \mu_i}{\sigma_i}, \tag{3.4}$$

where the i^{th} member of the n variables is denoted as x_i, and the value of the variable for the j^{th} member of the m cases is denoted as $x_i^{(j)}$. The mean and standard deviation are defined in their usual manner:

$$\mu_i = \frac{1}{m}\sum_{j=1}^{m} x_i^{(j)}, \tag{3.5}$$

and

$$\sigma_i = \frac{1}{m-1}\sqrt{\sum_{j=1}^{m}[x_i^{(j)} - \mu_i]^2}\ . \tag{3.6}$$

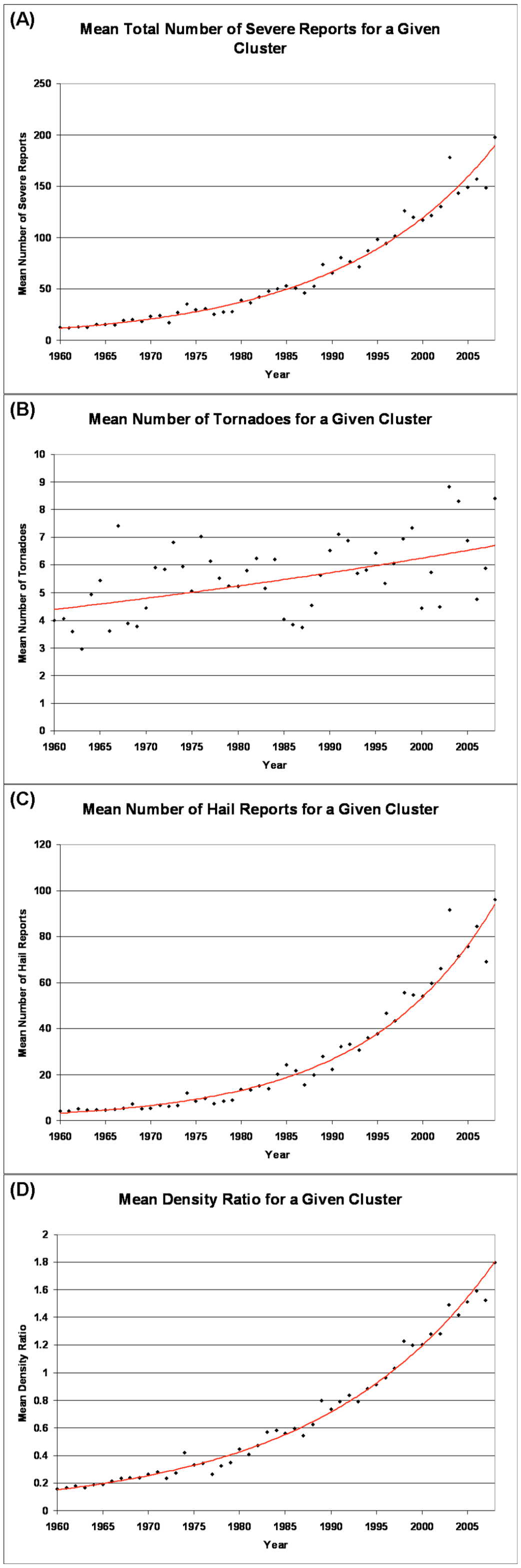


Fig. 3.3. Examples of detrending for the annual means of the clusters with total number of reports and density ratios above the detrended annual means. Variables labeled in each chart.

The standardized variable $\tilde{x}_i^{(j)}$ as computed in Eq. (3.4) is then given a weight w_i, and the final score of the index is given by:

$$I^{(j)} = \frac{\sum_{i=1}^{n} w_i \tilde{x}_i^{(j)}}{\sum_{i=1}^{n} w_i} \qquad (3.7).$$

Thus, the score of the index is the sum of the products of the weights and standardized values divided by the sum of the weights. In this manner, it is the *relative* weights of the variables that are pertinent. Variables were weighted with values ranging from 0 to 10, as all of the parameters were associated positively with the significance of severe weather events.

This method permits modifying the weights, for the computation of several different indices, to determine if the ranking of the severe weather events is susceptible to substantial variability. In general, the same weights that were used in SD10a (their Section 3a) were used in this study as well. However, the density ratio replaced the middle-50% parameter in this study, and it was given a weight of 3 for each of the indices (similar to the equivalent treatment of the middle-50% parameter for all of the indices used in D06 and SD10a).

The techniques used in this study and that of SD10a are intentionally similar, as the former developed a technique that was relatively simple to implement and easy to reproduce. Optimality of these methods cannot be shown but is not necessarily required, as no known "truth" of severe weather outbreak rankings exists. Various other methods could have been used to detrend the variables, and other types of severe weather reports could be used in the multivariate indices. See D06 and SD10a for the reasons involved

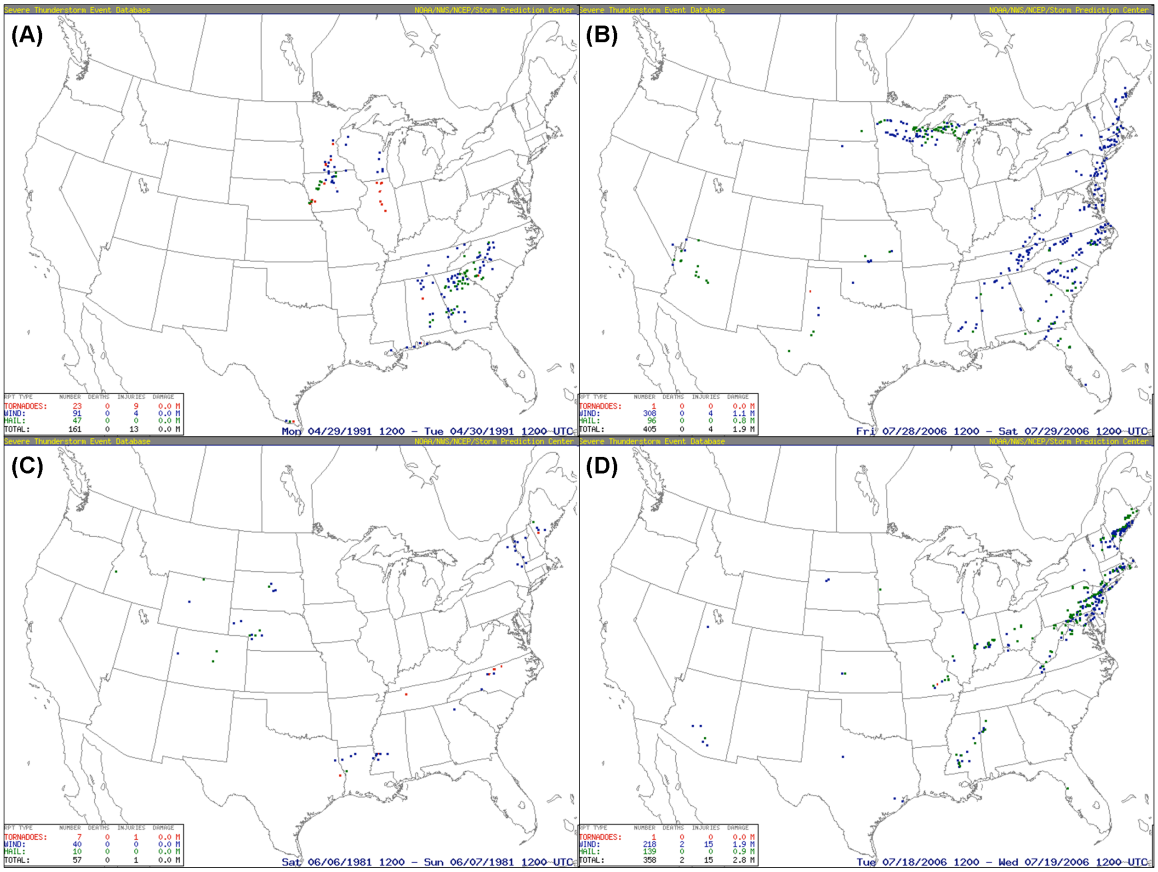

Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(A)
RPT TYPE NUMBER DEATHS INJURIES DAMAGE
TORNADOES: 23 0 9 0.0 M
WIND: 91 0 4 0.0 M
HAIL: 47 0 0 0.0 M
TOTAL: 161 0 13 0.0 M
Mon 04/29/1991 1200 – Tue 04/30/1991 1200 UTC
Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(B)
RPT TYPE NUMBER DEATHS INJURIES DAMAGE
TORNADOES: 1 0 0 0.0 M
WIND: 308 0 4 1.1 M
HAIL: 96 0 0 0.8 M
TOTAL: 405 0 4 1.9 M
Fri 07/28/2006 1200 – Sat 07/29/2006 1200 UTC
Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(C)
RPT TYPE NUMBER DEATHS INJURIES DAMAGE
TORNADOES: 7 0 1 0.0 M
WIND: 40 0 0 0.0 M
HAIL: 10 0 0 0.0 M
TOTAL: 57 0 1 0.0 M
Sat 06/06/1981 1200 – Sun 06/07/1981 1200 UTC
Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(D)
RPT TYPE NUMBER DEATHS INJURIES DAMAGE
TORNADOES: 1 0 0 0.0 M
WIND: 218 2 15 1.9 M
HAIL: 139 0 0 0.9 M
TOTAL: 358 2 15 2.8 M
Tue 07/18/2006 1200 – Wed 07/19/2006 1200 UTC

Fig. 3.4. As in Fig. 3.1a, for (a) 29 April 1991, (b) 28 July 2006, (c) 6 June 1981, and (d) 18 July 2006.

in the selection of the variables used herein and the various methods used in the ranking of these events.

c. KDE analysis

To identify severe weather events by clusters of severe reports for a given day, modifying Eqn. (3.3) by changing the bandwidth (*h*) for a given map projection and analyzing various threshold values of *f(x)* were required. Using the grid point method and a latitude-longitude map projection with 1° grid spacing, analysis of the severe reports (Fig. 3.4a) and the resultant two-dimensional PDF contour charts for various modifications of the bandwidth and contour thresholds (Fig. 3.5) illustrate the process. This day featured three distinct regions of severe weather: the Southeast, the Upper

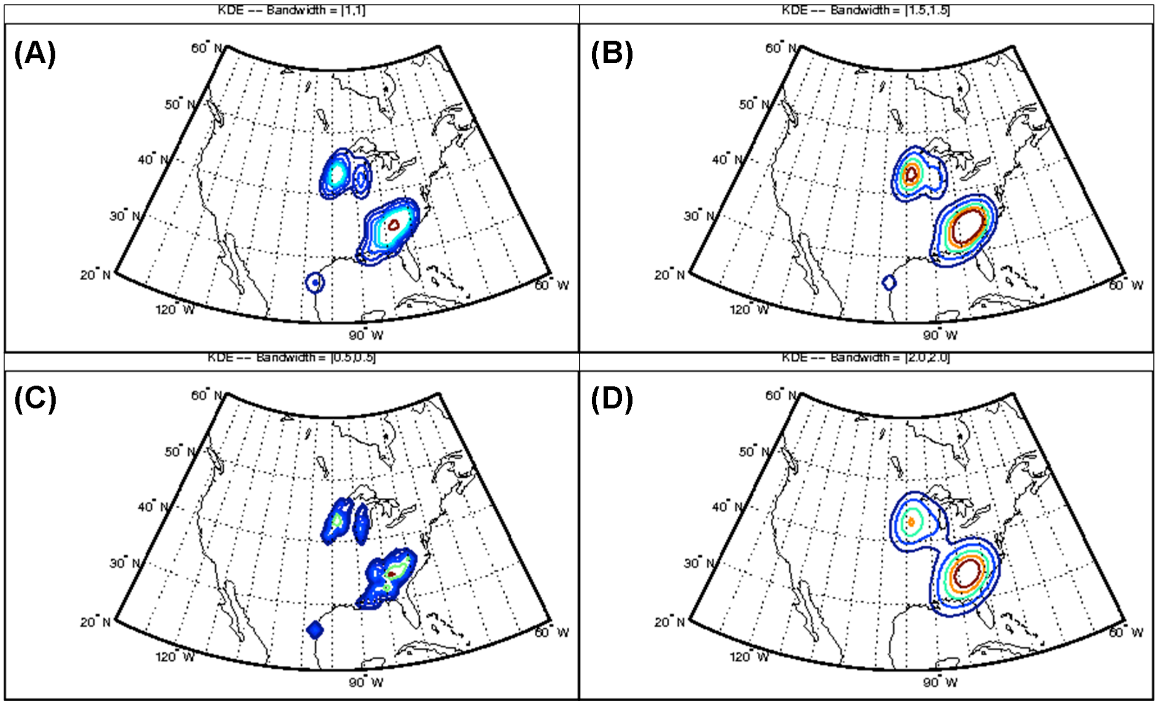


Fig. 3.5. Two-dimensional kernel density estimations of the probability density functions for severe reports from 1200 UTC 29 April 1991 to 1159 UTC 30 April 1991, using bandwidths of (a) 1, (b) 1.5, (c), 0.5, and (d) 2 for the latitudinal and longitudinal directions. These plots use severe reports converted to a latitude-longitude map projection with 1° grid spacing. The outermost contour is 0.001, and the contour within the outermost contour (second contour) is 0.005 for each plot.

Midwest, and Deep South Texas. Thus, the selected bandwidth and PDF threshold should capture these three locations as distinct events. This clearly excludes options with relatively high bandwidths (Fig. 3.5d), as the low-valued contours indicate the Midwest and Southeast events as one severe weather region. At the same time, the outermost contour (the 0.001 threshold) does not enclose the three reports in Deep South Texas.

Conversely, low bandwidths typically result in separate clusters for reports that are relatively close together. The two regions indicated in the Upper Midwest using bandwidths of 0.5 for both the latitude and longitude dimensions are undesirable (Fig. 3.5c), given their relatively close proximity. Though there is some separation of the reports (Fig. 3.4a), the relatively small distance between these areas suggests distinct synoptic-scale systems are not associated with the two regions. Furthermore, the contours are not smooth, which is also undesirable.

Varying the bandwidth in KDE is analogous to varying the smoothing parameter used in distance-dependent, weighted-average methods for objective analysis, described in Barnes (1964; see his Fig. 4 which shows the density of upper air stations using a Gaussian kernel). The objective, therefore, is to find a range of bandwidths falling in between the two extremes discussed above (as in Figs. 3.5a,b). In general, lower bandwidths in this range were preferred, as these had a tendency to include more minor, isolated events into separate clusters at thresholds that also did not combine regionally separate severe weather events. Based on Fig. 3.5, bandwidths of 1 for the latitude and longitude dimensions were preferred over 1.5.

Selection of PDF thresholds primarily was determined by the lowest threshold that included the most number of reports while also not merging regionally separate events. For example, in Fig. 3.5d, the second contour (0.005 threshold) would be selected over the outermost contour (0.001 threshold), as the 0.001 threshold combined the two regionally separate clusters of reports. However, the 0.005 threshold also does not enclose all of the reports associated with particular clusters (e.g., the Midwest region; cf. Figs. 3.4a and 3.5d). Selecting a threshold too high results in separating relatively close regions, such as the 0.005 threshold (second contour) in Fig. 3.5a.

Of course, bandwidth and threshold selection occurred only after analyzing a large number of cases. An analysis of additional cases shows that the bandwidth of 1 and the threshold of 0.001 (outermost contour; Figs. 3.6a,c,e) are reasonable selections for various types of events. For example, as shown in Fig. 3.4b on 28 July 2006, distinct

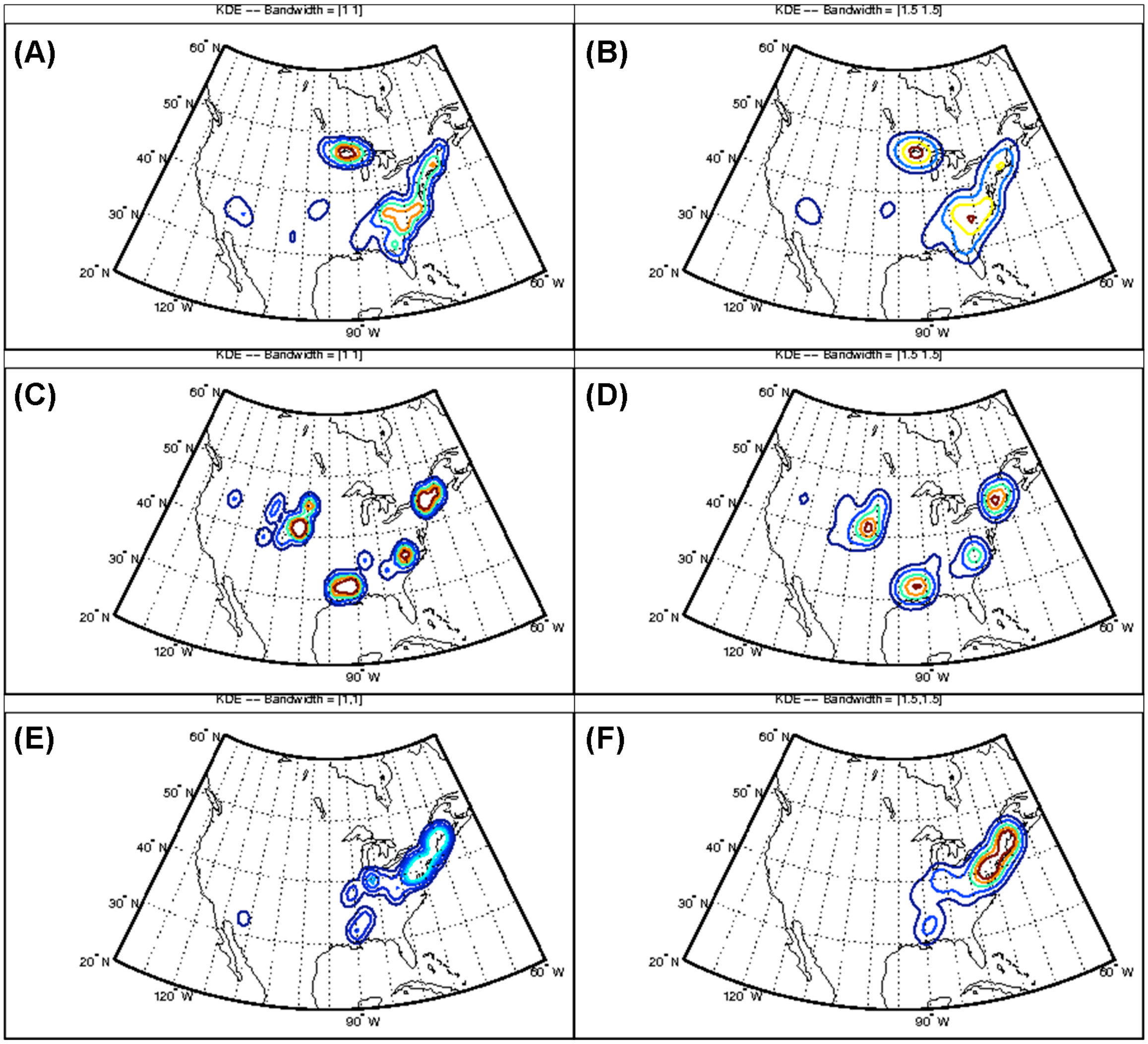


Fig. 3.6. (a) As in Fig. 3.5a, for 28 July 2006. (b) As in Fig. 3.5b, for 28 July 2006. (c) As in (a), for 6 June 1981. (d) As in (b), for 6 June 1981. (e) As in (a), for 18 July 2006. (f) As in (b), for 18 July 2006.

regions of reports are observed near Lake Superior and the East Coast, with dispersed reports in the central and southern plains and a small cluster in the Southwest. Bandwidths of 1 to 1.5 with thresholds near 0.001 identify the two clusters with large numbers of reports well with limited coverage of the reports in the central and southern plains (Figs. 3.6a,b). On 6 June 1981, relatively few severe reports were scattered through the northern High Plains, Southeast, and Northeast (Fig. 3.4c). The two-dimensional PDFs were quite different between the two bandwidths, with the smaller bandwidth associated with a larger number of clusters (Figs. 3.6c,d). Preference for one bandwidth over the other is not obvious here; however, the initial criteria for event consideration when ranking and classifying the outbreaks (see Sections 6b and 6d) would eliminate these clusters *in either case*. The number of reports within each cluster is small using the bandwidth of 1, and the density ratio of each cluster is small using the bandwidth of 1.5.

On 18 July 2006, numerous severe reports were observed over much of the eastern US, with a small, separate cluster in Alabama and Mississippi (Fig. 3.4d). Widely dispersed reports were observed in the plains, and a small cluster of reports was observed in southern Arizona. The separate cluster in Alabama and Mississippi was identified using a bandwidth of 1, as well as the cluster in Arizona (Fig. 3.6e). This was not the case for the bandwidth of 1.5, unless the PDF threshold was increased for the Southeast cluster and decreased for the Arizona cluster (Fig. 3.6f). Cases like these appear to be handled somewhat more appropriately by the lower bandwidth, resulting in its selection for the latitude-longitude map projection with 1° grid spacing.

Using a different map projection required modifications to bandwidth and PDF threshold selection, if the distance quantities were defined in terms of grid points (rather than latitudes and longitudes, or distances between the grid point and the location of the severe report). For example, a Lambert conformal projection with 54-km grid spacing, using the grid point method, also was conducted to compute KDE estimates of severe weather clusters for each day in the 49-yr period (Fig. 3.7). If the bandwidth and PDF threshold were identical (in magnitude) to that of the latitude-longitude projection, the results were markedly different (cf. Figs. 3.5a and 3.7a). These differences were

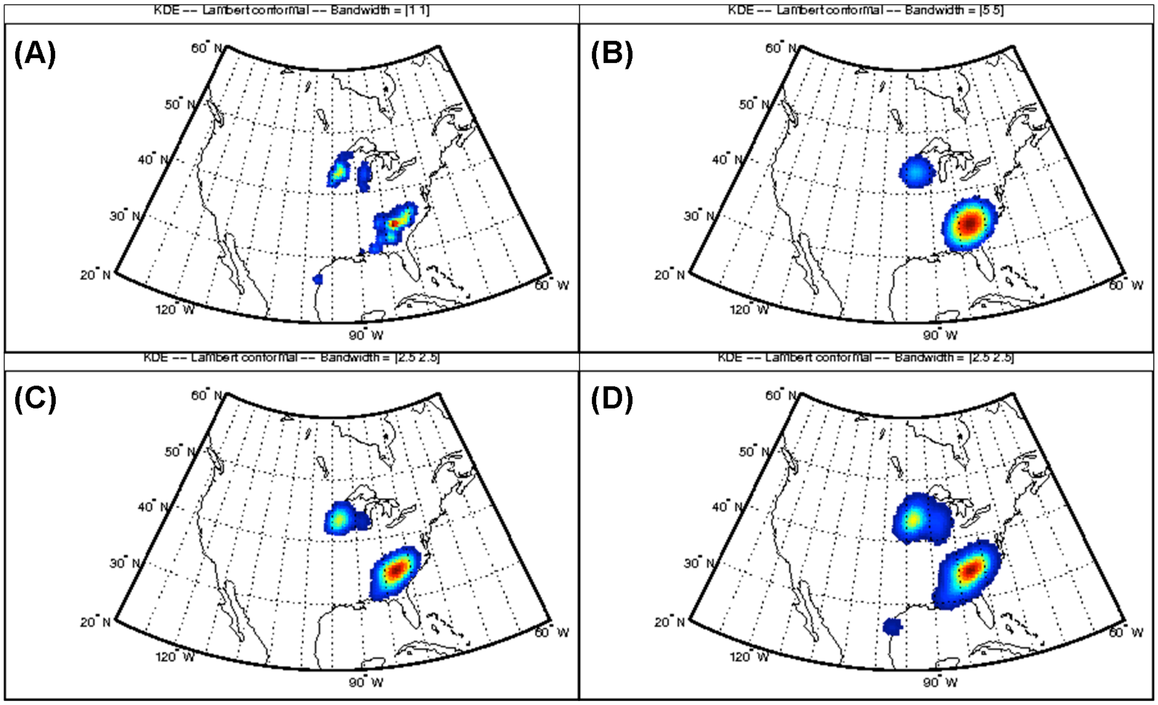

KDE -- Lambert conformal -- Bandwidth = [1 1]
KDE -- Lambert conformal -- Bandwidth = [5 5]
(A)
(B)
KDE -- Lambert conformal -- Bandwidth = [2.5 2.5]
KDE -- Lambert conformal -- Bandwidth = [2.5 2.5]
(C)
(D)
60° N
50° N
40° N
30° N
20° N
120° W
90° W
60° W

Fig. 3.7. As in Fig. 3.5, using the grid point method and a Lambert conformal projection with 54-km grid spacing, with a bandwidth of (a) 1, (b) 5, and (c) 2.5. Shading begins with a threshold of 0.001. (d) As in (c), with shading beginning with a threshold of 0.00025.

anticipated, as the grid boxes and the grid spacing were different for the two projections. Essentially, the Lambert conformal projection (with lower grid spacing) required a substantially higher bandwidth and a lower PDF threshold to replicate the characteristics of the latitude-longitude projection (Fig. 3.7d).

Using relatively high bandwidths for the Lambert conformal projection resulted in characteristic shapes that did not match the reports well (cf. Figs. 3.4a and 3.7b), indicative of too much smoothing. Using a bandwidth of 2.5 for the Lambert conformal projection seemed to replicate the characteristic shapes of the two biggest clusters for the latitude-longitude map projection well (cf. Figs. 3.5a and 3.7c), but the PDF threshold of 0.001 did not capture the reports in Deep South Texas. Lowering the threshold to 0.00025 (Fig. 3.7d) solved this problem, and the coverage for each of the three clusters was quite similar to that of the latitude-longitude projection with a bandwidth of 1 and PDF threshold of 0.001. Note that these changes to the bandwidth are similar to the changes in grid spacing between the two map projections. A 1° latitude-longitude projection is generally on the order of 100–150-km grid spacing. This is approximately 2-2.5 times the grid spacing of the Lambert conformal projection. As the magnitude of the bandwidth for the Lambert conformal projection is increased by a factor of 2-2.5, the PDF threshold is approximately 1/4-1/6 of its value for the latitude-longitude projection. *This finding indicates that the selection of alternative map projections should not change the results substantially, if the bandwidth and PDF threshold are changed accordingly.*

Finally, the differences between initially converting the severe reports to grid point values and computing the distances between these grid point values and each of the grid points to the distances (in km) between the grid point coordinates of a selected map projection to the actual point locations of each severe report (the distance method) are also quite minor (Fig. 3.8). *Thus, initially converting the severe reports to a grid did not alter the areal coverage of a severe weather cluster substantially, as long as the bandwidth and PDF threshold were modified accordingly.* Because of this finding, the results for the grid point method, using the latitude-longitude map projection, will be discussed for the rest of the paper.

The selection of the bandwidth and PDF thresholds clearly is subjective. However, the objective of reproducibility for ranking and classifying severe weather outbreaks is met, provided the same bandwidth and PDF thresholds are used. Furthermore, selection of slightly different values would not alter the results substantially for the most significant severe weather events. For example, if a bandwidth of 1 or 1.5 is

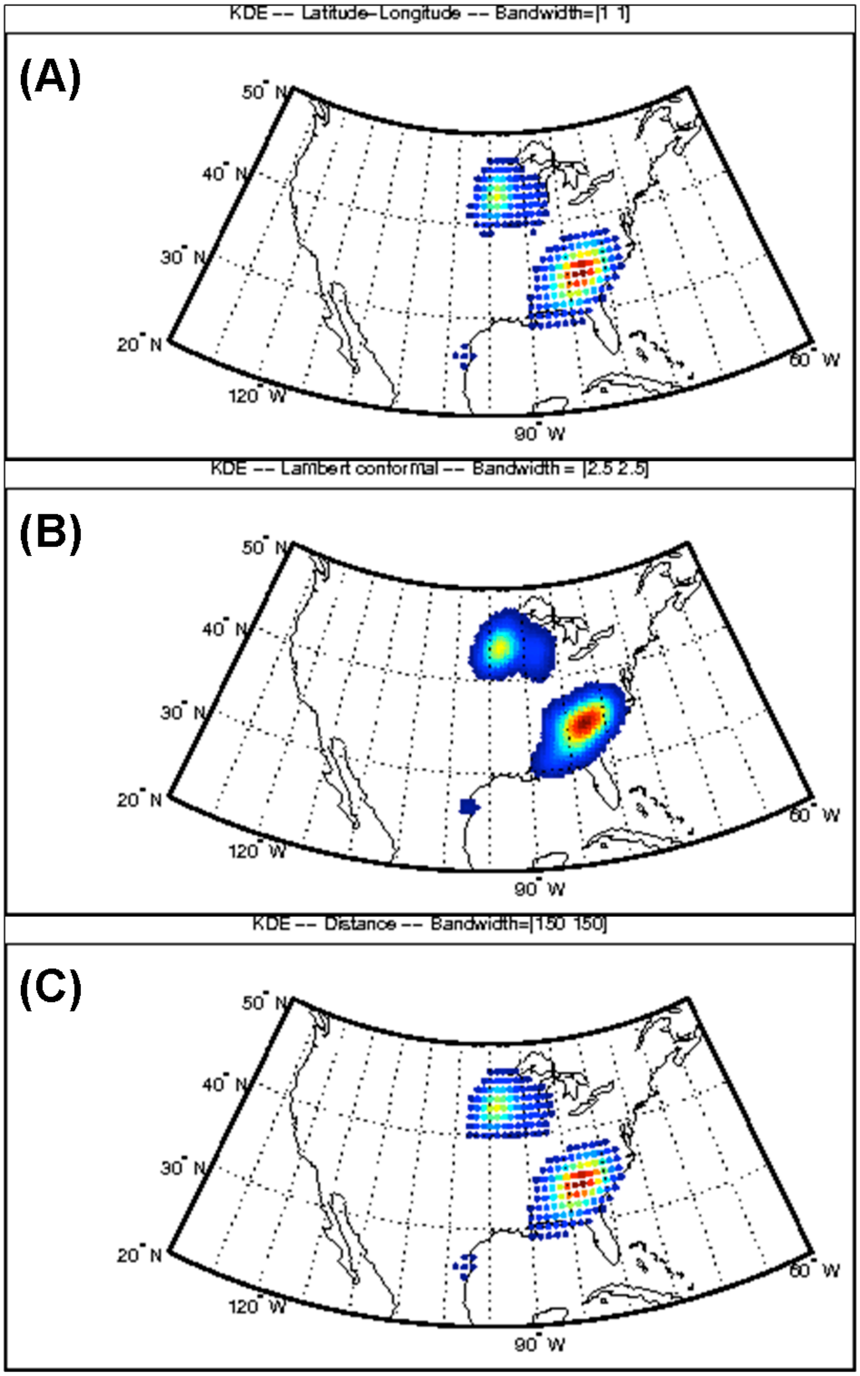

KDE -- Latitude-Longitude -- Bandwidth=[1 1]
(A)
50° N
40° N
30° N
20° N
120° W
90° W
60° W
KDE -- Lambert conformal -- Bandwidth = [2.5 2.5]
(B)
50° N
40° N
30° N
20° N
120° W
90° W
60° W
KDE -- Distance -- Bandwidth=[150 150]
(C)
50° N
40° N
30° N
20° N
120° W
90° W
60° W

Fig. 3.8. Scatter plots showing the grid points that exceed a specified probability density function threshold using kernel density estimation of the severe reports from 1200 UTC 29 April 1991 to 1159 UTC 30 April 1991. In (a), the grid point method (see relevant text) is used, with a latitude-longitude projection with 1° grid spacing, a bandwidth of 1, and a PDF threshold of 0.001. In (b), the grid point method is used, with a Lambert conformal projection, 54-km grid spacing, a bandwidth of 2.5, and a PDF threshold of 0.0003. In (c), the distance method (see relevant text) is used, with a latitude-longitude projection, 1° grid spacing, a bandwidth of 150 km, and a PDF threshold of 0.00014.

chosen using the latitude-longitude projection, the areas enclosed by the 0.001 PDF threshold are quite similar for the two most significant events on 28 July 2006 (e.g., see Figs. 3.6a,b). The results generally are reasonably robust to modifications of the bandwidth and PDF threshold selections, so long as these selections avoid the tendencies shown in Figs. 3.5c,d.

The final step in the KDE analysis is to eliminate the cases that would not be classified readily as a severe weather event or outbreak. The criteria for such elimination were selected somewhat arbitrarily, but the objective was to remove cases with relatively sparse coverage of reports within a cluster or with relatively few reports within a cluster. These cases, in addition to not qualifying as significant severe weather events, also tended to be handled more variably by the KDE scheme (e.g., cf. Figs. 3.6c,d).

Any cluster in which the number of reports within the region was less than the detrended annual mean number of reports for a cluster or the density ratio was less than the detrended annual mean for a cluster was removed from consideration (refer to Fig. 3.3). The detrended annual mean number of reports for a cluster was small (from ~5 in 1960 to ~43 in 2008), as desired, to ensure that as many cases considered to be significant severe weather events as possible were included. For the latitude-longitude map projection with 1° grid spacing, using a bandwidth of 1 and a PDF threshold of 0.001, a total of 6072 cases were retained.

d. Outbreak rankings

After removal of the report clusters consisting of few reports or sparse coverage, annual means of the variables used in the linear-weighted multivariate indices to rank the outbreaks were computed (i.e., the average value per cluster for a particular year).

Variables with secular trends in the annual means were detrended (e.g., Fig. 3.3). Each variable (detrended or otherwise) was standardized as in Eqs. (3.1)-(3.3), and the scores for each cluster were computed as in Eq. (3.4). The relative weights of the variables were altered to develop 26 indices, with the weights equivalent to those of SD10a (their Fig. 4), with the same notation. As Section 3b discussed, the middle-50% parameter was replaced by the density ratio, and the density ratio was given a weight of 3 for each of the 26 indices.

As explained in SD10a, there are essentially two sets of indices. The first set, which includes indices N0-N16 and N20, gives nonzero weights for all of the tornado variables.[13] These will be referred to as the "all-tornado indices". The second set, which includes N17-N19 and N21-N25, gives nonzero weights to only two of the tornado variables (which are changed among the indices). These will be referred to as the "two-tornado indices". The reasons for developing these two sets of indices include an investigation of the volatility of the rankings when a large number of the variables are removed, to counteract a negative bias for severe report clusters with a large number of significant nontornadic events, and to determine the additional explanatory power of highly-correlated tornado variables. Within the two sets of indices, modifications to the weights investigated the preference toward particular tornado variables and the changes in the rankings when giving significant nontornadic reports relatively high weights. The reader is referred to SD10a for more explanation regarding the choice of variables and their weights.

The scores for each of the 6072 cases obtained using the grid point method for the latitude-longitude projection were computed for all 26 indices (Fig. 3.9). There were

[13] The N0 index is the control, in which each variable is given equal weight.

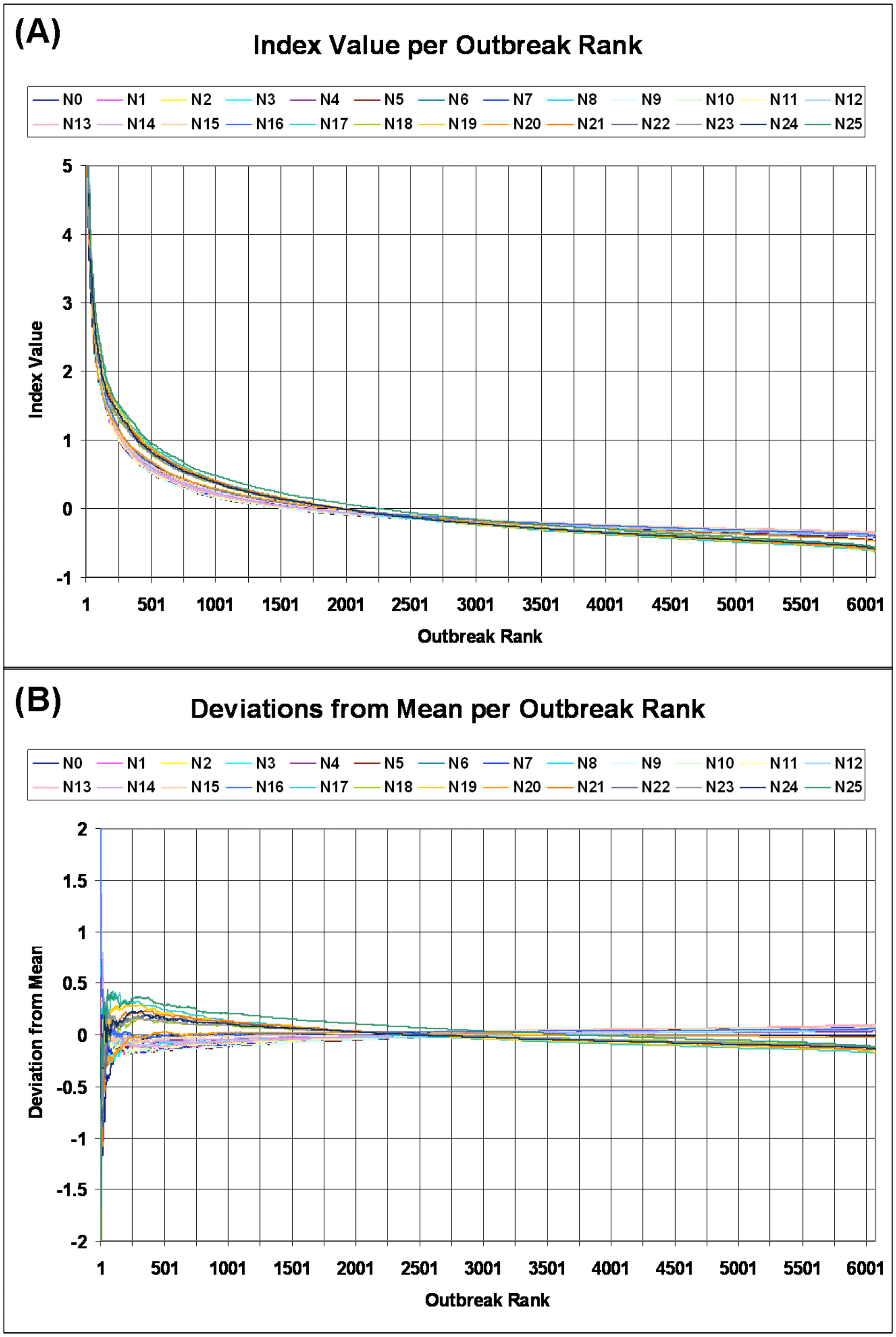


Fig. 3.9. (a) The index scores (*y*-axis) and rankings (*x*-axis) of each of the 6072 cases for each of the indices in the study (labeled). (b) The deviations (*y*-axis) of each of the indices (labeled) from the mean score of all the indices for a particular rank (*x*-axis).

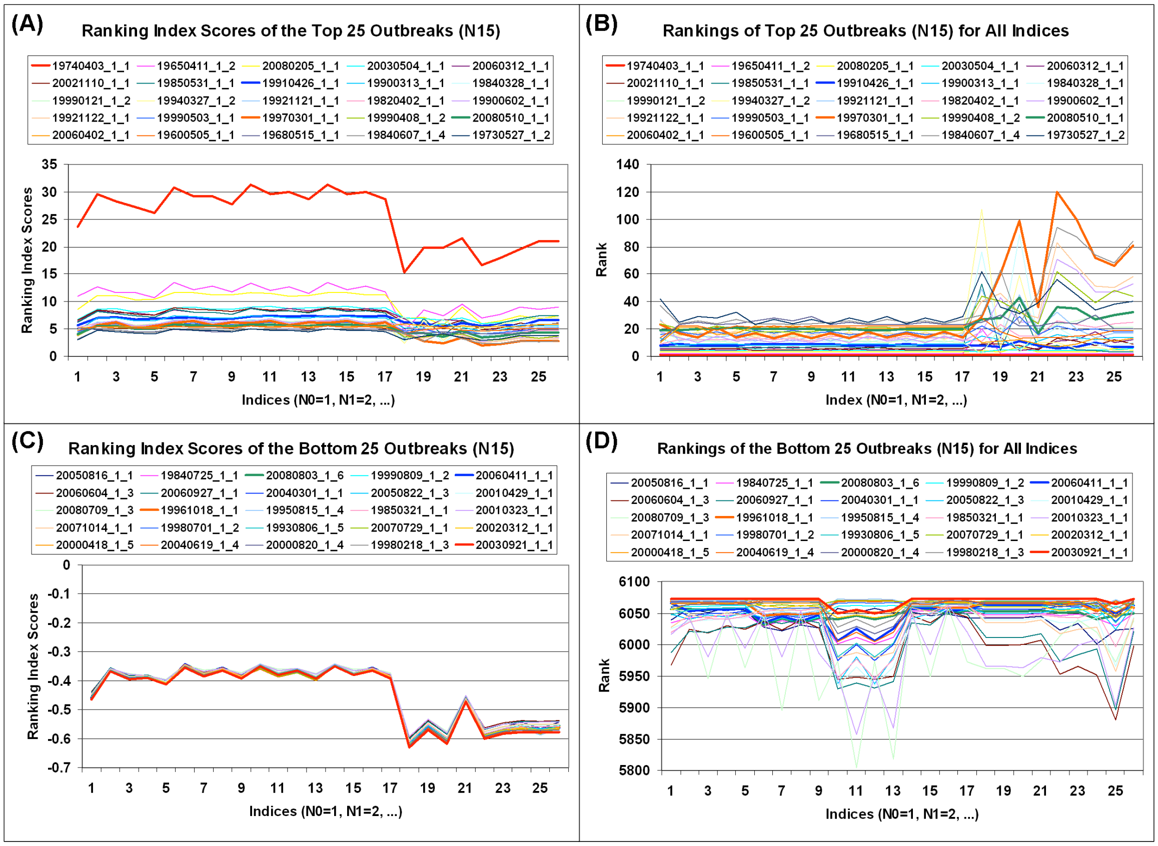


Fig. 3.10. (a) The ranking index scores (*y*-axis) for each of the 26 indices (*x*-axis; N0=1, N1=2, etc.) for the top 25 outbreaks based on the rankings of the N15 index. (b) The rankings for each of the 26 indices [as indicated in (a)] of the top 25 outbreaks based on the ranking of the N15 index. (c) As in (a), for the bottom 25 cases based on the N15 index. (d) As in (b), using the bottom 25 cases based on the N15 index. Four cases of each type are in bold for convenience.

three main findings: (1) The highest-scored cases (approximately 250 of them) have a very steep negative slope, analogous to the first ~200 cases in SD10a and Section 3 herein (cf. Fig. 3.6). The next ~500 cases have a smaller but relatively substantial negative slope. The final 5250 cases have very small to nearly neutral slopes (analogous to the middle ~1000 cases in SD10a). The deviations of the individual index scores from the mean score of all the indices for each rank (Fig. 3.9b) indicate substantial noise and relatively large deviations for the top 200-250 cases, gradually less noise and smaller deviations for the next 500 cases, and virtually no noise and small deviations for the final

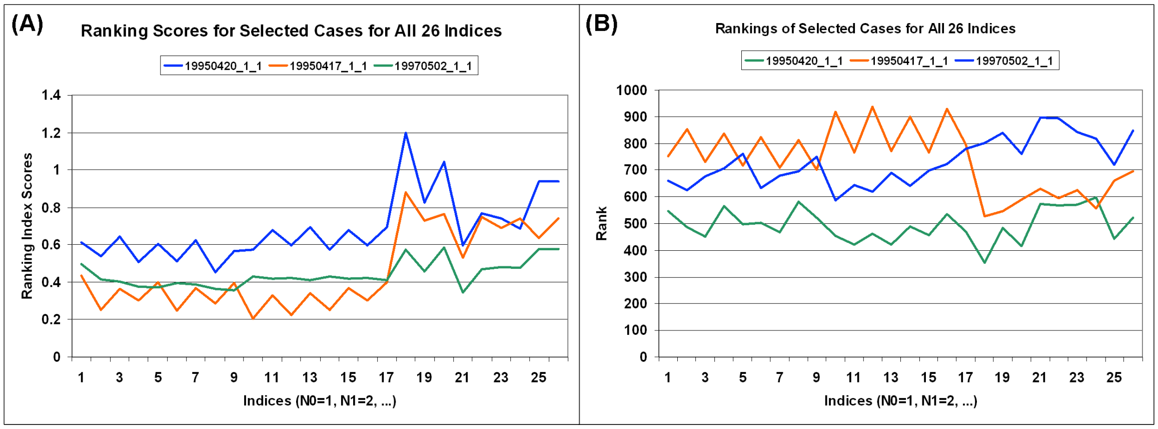


Fig. 3.11. (a) As in Fig. 3.10a, for three specific events (17 April 1995, 20 April 1995, 2 May 1997) as referenced in the text. (b) As in Fig. 3.10b, for the cases in (a).

5000 cases. These tendencies indicated that the rankings of the top cases were relatively consistent no matter what index was used (e.g., Figs. 3.10a,b), whereas the rankings were relatively volatile for the lower cases (e.g., Fig. 3.11 – see also SD10a for more details). (2) None of the curves exhibit a second steep negative-sloped section analogous to the final ~200 cases in SD10a. This was an intentional outcome of the study. That is, the cases with substantial geographic scatter and/or relatively few reports for a given event have been excluded successfully from consideration. (3) Two distinct groups of curves are present. The group of curves with a steeper slope for the top cases and a more neutral slope for the remaining cases consists of the all-tornado indices. The other group of curves comprised the two-tornado indices. These differences are more noticeable than in SD10a, and are likely a result of the strong correlations among the tornado variables (their Section 3a). Additionally, the two groups of curves intersect twice. The first intersection occurs in the high-ranked portion of the cases, and the second occurs at around the 2500-3000 ranks. These results suggest that the modifications of the weights within the two groups of indices do not affect the rankings substantially, but removing a

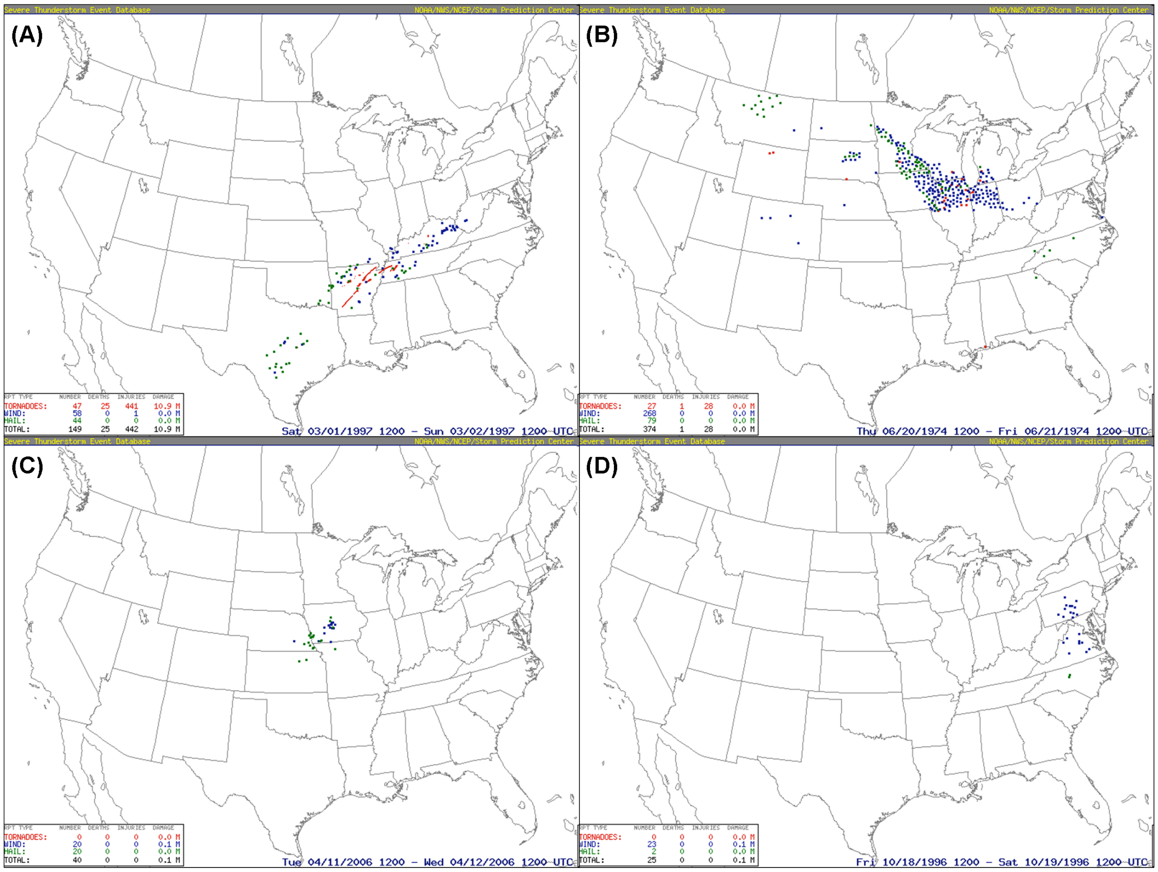

Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(A)
RPT TYPE NUMBER DEATHS INJURIES DAMAGE
TORNADOES: 47 25 441 10.9 M
WIND: 58 0 1 0.0 M
HAIL: 44 0 0 0.0 M
TOTAL: 149 25 442 10.9 M
Sat 03/01/1997 1200 – Sun 03/02/1997 1200 UTC
(B)
RPT TYPE NUMBER DEATHS INJURIES DAMAGE
TORNADOES: 27 1 28 0.0 M
WIND: 268 0 0 0.0 M
HAIL: 79 0 0 0.0 M
TOTAL: 374 1 28 0.0 M
Thu 06/20/1974 1200 – Fri 06/21/1974 1200 UTC
(C)
RPT TYPE NUMBER DEATHS INJURIES DAMAGE
TORNADOES: 0 0 0 0.0 M
WIND: 20 0 0 0.1 M
HAIL: 20 0 0 0.0 M
TOTAL: 40 0 0 0.1 M
Tue 04/11/2006 1200 – Wed 04/12/2006 1200 UTC
(D)
RPT TYPE NUMBER DEATHS INJURIES DAMAGE
TORNADOES: 0 0 0 0.0 M
WIND: 23 0 0 0.1 M
HAIL: 2 0 0 0.0 M
TOTAL: 25 0 0 0.1 M
Fri 10/18/1996 1200 – Sat 10/19/1996 1200 UTC

Fig. 3.12. As in Fig. 3.1a, for (a) 1 March 1997, (b) 20 June 1974, (c) 11 April 2006, and (d) 18 October 1996.

subset of the tornado variables from the indices can affect the rankings of the cases noticeably.

This last point is illustrated with the cases highlighted in Fig. 3.10. The absence of the 1 March 1997 tornado outbreak (Fig. 3.12a) from the top 25 outbreaks in the two-tornado indices (N17-N19 and N21-N25) is a result of the relative lack of nontornadic reports on that day. Some other cases exhibit this behavior (e.g., 15 November 2005), indicating a potential drawback of removing several tornado variables from consideration. Interestingly, these days commonly were excluded from consideration in SD10a because the total number of severe reports was below the top 30 days for that year (true for both 1 March 1997 and 15 November 2005). The new scheme presented herein allows for such days to be considered while simultaneously excluding cases with excessive geographic scatter.

On the other hand, the presence of the 20 June 1974 severe weather outbreak (not shown in Fig. 3.10; reports in Fig. 3.12b) in the two-tornado indices and its absence from the remaining indices (except N0, the control) was a result of a relative lack of tornadoes but an anomalously large number of wind reports observed on that day. This was considered herein to be a desirable characteristic of the ranking indices with few tornado variables included; however, this comes at the cost of placing cases with a large number of strong tornadoes with relatively few nontornadic reports lower (e.g., the N25 index places 1 March 1997 as 81st). The selection of the "best" ranking index is dependent on research goals. If the task is to identify tornado outbreaks and discriminate from all other events, the indices that include a larger number of tornado variables should be selected. If the task is to identify significant severe weather outbreaks of any type, emphasizing the total number of reports and significant nontornadic reports, the selection of indices with fewer tornado variables is reasonable. Nevertheless, most of the highest-ranked cases were major tornado outbreaks, no matter which index was used to rank the cases, another result considered to be desirable.

The volatility of rankings increases for cases below the steep portion of the curves (Fig. 3.11). For the 20 April 1995, 17 April 1995, and 2 May 1997 severe weather events (SD10a; their Fig. 10), the rankings are variable within the two groups of indices (e.g., the 17 April 1995 cluster had a range of 174 for rankings among indices N13-N16 and 171 for rankings among indices N17-N19, N22, and N25). However, the range widens substantially between the two groups of indices [e.g., the 17 April 1995 cluster had a high ranking of 526 (N17) and a low ranking of 930 (N15) – a range of slightly greater than 400]. Thus, the volatility of rankings between the two groups of indices is more substantial than within the two groups. However, the cases falling within the strongly-sloped section of the curves (hereafter, the major severe weather outbreaks) tended to remain in that section no matter what index within the same group of indices (i.e., the all-tornado or two-tornado indices) was used. As in SD10a, this finding suggests that diagnosis or prognosis of the ranking (or index score) of an outbreak is challenging, whereas the diagnosis or prognosis of an outbreak's severity based on general location within the curves (of the scores or rankings) may be more feasible.

Interestingly, the lowest-ranked cases were reasonably consistent no matter what index was used (Figs. 11c,d). In SD10a, the effectiveness of the middle-50% parameter was determined to be the reason behind the consistency of the lowest-ranked cases. These cases are not considered in this study. Instead, the rankings were relatively consistent because these events had relatively few reports, relatively limited coverage, and/or a relative lack of significant severe weather (e.g., Figs. 3.12c,d). Also noticeable are the dates of these cases. The majority of these cases occur after 1990, as a result of the relative lack of reporting with similar events prior to this time. As noted earlier, nonmeteorological artifacts have not been removed completely.

e. Classifying outbreaks

Because of the relative volatility of severe weather event rankings outside of the extremes, classification of all of these cases based on the characteristics of the severe reports is appropriate and potentially beneficial for operational forecasters. As in SD10a, a cluster analysis is performed on the four-dimensional decomposition of the indices. All of the variables associated with tornadoes are included in the tornado component, all of the variables associated with wind are included in the wind component, and all of the variables associated with hail are included in the hail component. The fourth component includes the remaining variables (the total number of severe reports of all types, and the density ratio) and is referred to hereafter as the "miscellaneous" component.

After analyzing several types of cluster analyses, two of the most appropriate methods were the *k*-means (sample seeding) cluster analysis (Gong and Richman 1995) and the Ward's hierarchical technique (Ward 1963). Analysis of the decomposition using three-dimensional scatter plots, in which one of the four components is eliminated from the analysis, allows for simple interpretation of the results. As in SD10a, the N3 and N22 indices will be presented, as the N3 (N22) index is one that incorporates all (a subset) of the tornado variables. Results of the cluster analysis within the two groups of indices (i.e., the all-tornado indices and the two-tornado indices) were not substantially different (not shown).

Analysis of silhouette plots of the *k*-means cluster analyses (Kaufman and Rousseeuw 1990) for 2-15 clusters indicates that a small number of even-numbered clusters was favored (i.e., 2, 4, and 6; statistical significance not observed) for the N3 index (Fig. 3.11). The 2-cluster analysis suggests that significant severe weather outbreaks (in total 872) were clustered separately from the remaining cases (5200). Although major tornado outbreaks were a substantial proportion of the cases in this cluster (Fig. 3.13a), significant severe weather of any type was included (Fig. 3.13b). The 3 April 1974 tornado outbreak is an outlier in Fig. 3.13a (reports in Fig. 3.14a). The 21 April 1996 hail-dominant outbreak (Fig. 3.14b) and the 1 July 1994 and 30 May 1998

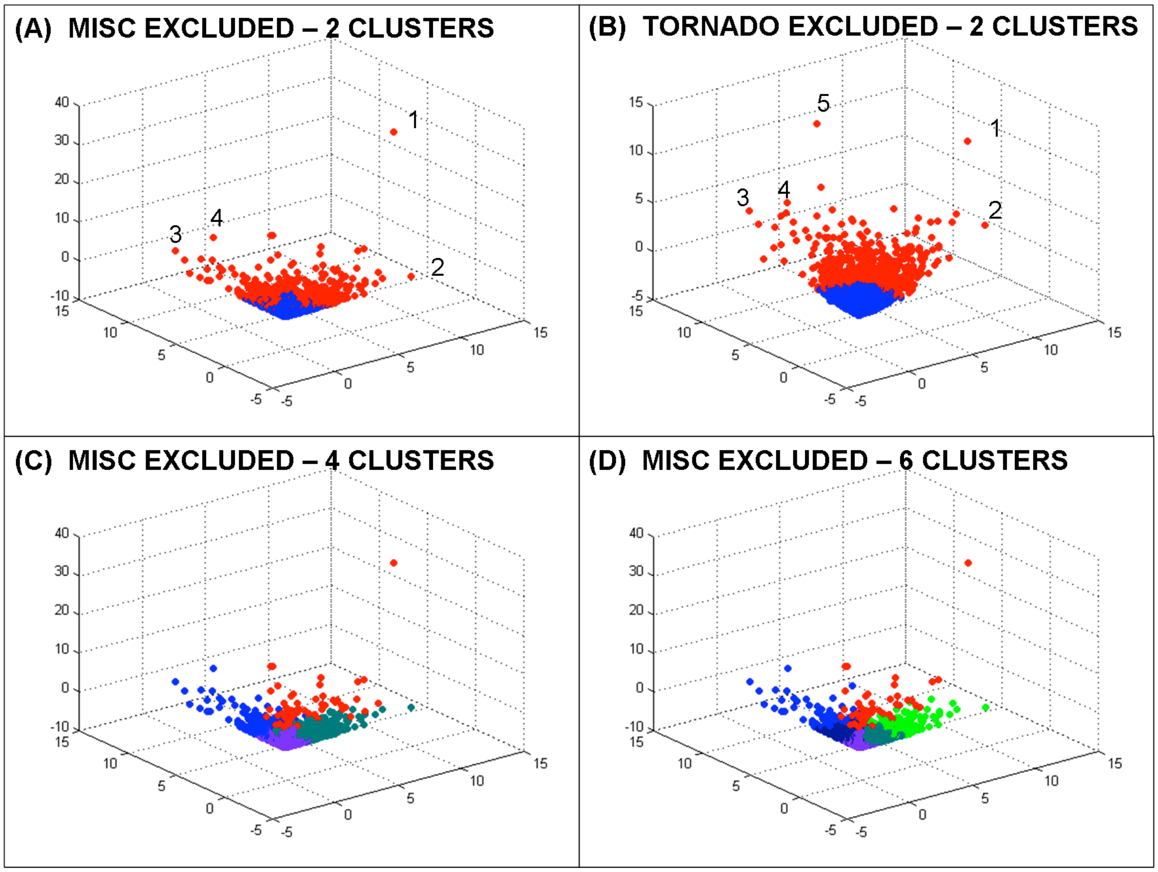


Fig. 3.13. Clusters obtained using the four-dimensional decomposition of the N3 index and k-means cluster analysis. Clusters identified by color, and excluded components of the analysis are labeled. Cases identified include (1) 3 April 1974, (2) 21 April 1996, (3) 1 July 1994, (4) 30 May 1998, and (5) 20 June 1974. In (c) and (d), hail-dominant events are shown in shades of green, wind-dominant events are shown in shades of blue, major tornado outbreaks are shown in red, and mixed-mode events are shown in purple.

wind-dominant outbreaks (Figs. 3.14c,d) are distinct outliers in the four-dimensional decomposition as well. The 20 June 1974 outbreak (refer to Fig. 3.12b) is a noticeable outlier when the miscellaneous component is analyzed (Fig. 3.13b), as this component includes the total number of reports of any type – which is anomalously large for this case. All of these events are included in the significant severe weather outbreak category.

The 4-cluster analysis indicates the existence of outbreaks that are dominated by one type of severe weather event (Fig. 3.13c). The major tornado outbreaks (red; 57 cases), hail-dominant outbreaks (green; 887 cases), and wind-dominant outbreaks (blue;

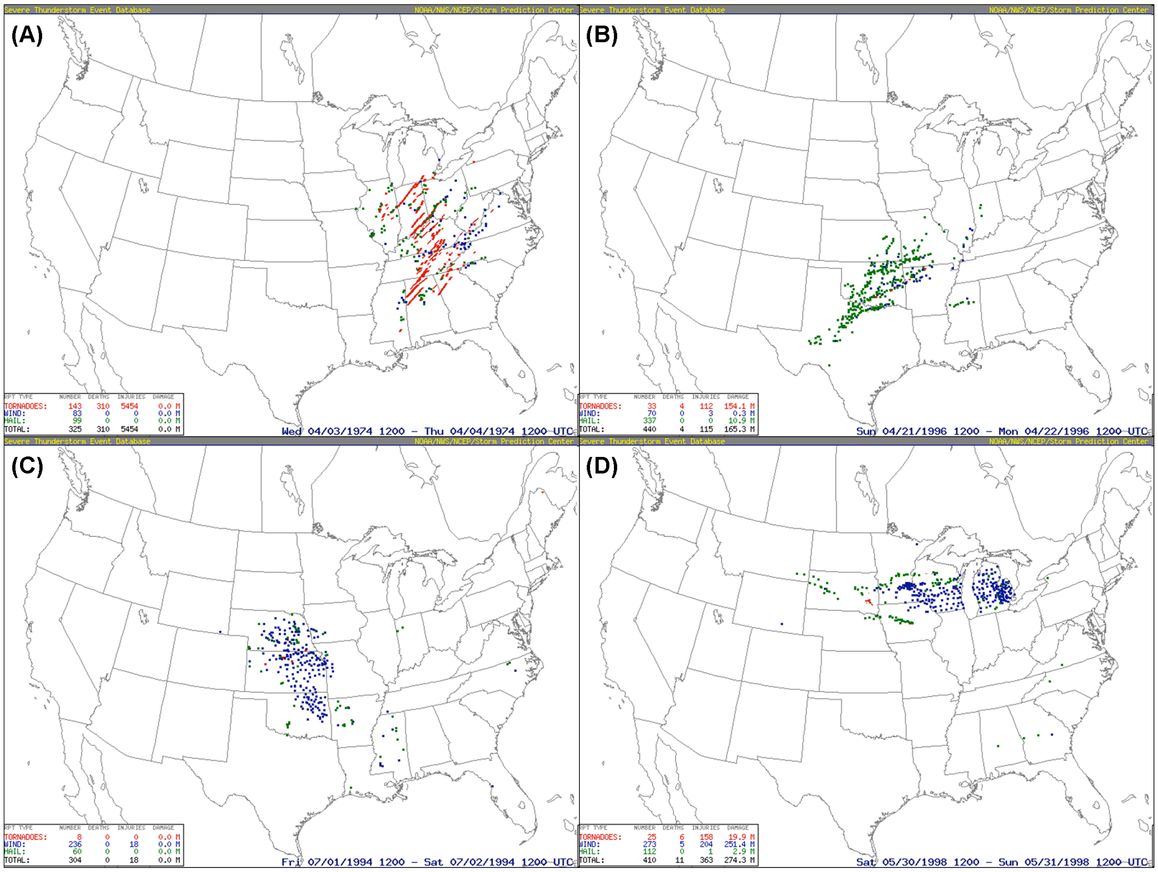

Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(A)
RPT TYPE NUMBER DEATHS INJURIES DAMAGE
TORNADOES: 143 310 5454 0.0 M
WIND: 83 0 0 0.0 M
HAIL: 99 0 0 0.0 M
TOTAL: 325 310 5454 0.0 M
Wed 04/03/1974 1200 – Thu 04/04/1974 1200 UTC
Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(B)
RPT TYPE NUMBER DEATHS INJURIES DAMAGE
TORNADOES: 33 4 112 154.1 M
WIND: 70 0 3 0.3 M
HAIL: 337 0 0 10.9 M
TOTAL: 440 4 115 165.3 M
Sun 04/21/1996 1200 – Mon 04/22/1996 1200 UTC
Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(C)
RPT TYPE NUMBER DEATHS INJURIES DAMAGE
TORNADOES: 8 0 0 0.0 M
WIND: 236 0 18 0.0 M
HAIL: 60 0 0 0.0 M
TOTAL: 304 0 18 0.0 M
Fri 07/01/1994 1200 – Sat 07/02/1994 1200 UTC
Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(D)
RPT TYPE NUMBER DEATHS INJURIES DAMAGE
TORNADOES: 25 6 158 19.9 M
WIND: 273 5 204 251.4 M
HAIL: 112 0 1 2.9 M
TOTAL: 410 11 363 274.3 M
Sat 05/30/1998 1200 – Sun 05/31/1998 1200 UTC

Fig. 3.14. As in Fig. 3.1a, for (a) 3 April 1974, (b) 21 April 1996, (c) 1 July 1994, and (d) 30 May 1998.

340 cases) are similar to the cluster analysis findings in SD10a (their Fig. 11). The remaining cases are the relatively minor “mixed-mode” events (purple; 4788 cases), in which little preference for any type of severe report is noted. The four days specified in Figs. 3.13a and 3.14 are in the classes one would expect in the 4-cluster analysis.

The 6-cluster analysis separates the hail-dominant and wind-dominant groups into two classes each (Fig. 3.13d). The major events are labeled in bright green (bright blue), whereas the minor events are identified by darker shades of the respective colors. In this analysis, there were 47 major tornado outbreaks, 262 major hail-dominant clusters, 104 major wind-dominant clusters, 1002 minor hail-dominant clusters, 806 minor wind-dominant clusters, and 3851 minor mixed-mode events. The major events (413 cases) primarily make up the steep portion of the characteristic curves in Fig. 3.9a.

The *k*-means cluster analysis of the N22 four-dimensional decomposition is quite similar (not shown), with the same interpretations of the various clusters for the 2-cluster, 4-cluster, and 6-cluster analyses. Additionally, the number of significant severe weather outbreaks in the 2-cluster analysis is 889, only 17 more cases than the N3 analysis. Severe weather events commonly were placed in the same categories no matter which index was used.

Ward's hierarchical technique also was found to be relatively reasonable in categorizing groups of cases into particular types. However, this technique appeared to group major tornado outbreaks and significant wind events (primarily derechos; see Johns and Hirt 1987) together (cf. Figs. 3.13c and 3.15c), which is undesirable. Distinguishing tornadoes from derechos has been a focus of several past studies (e.g., Stensrud et al. 1997; Doswell and Evans 2003), as the societal impacts of these cases typically are quite different. The 6-class analyses are relatively similar for the *k*-means and Ward's techniques (cf. Figs. 3.13d and 3.15d; 78% of cases classified in the same manner), though the minor hail-dominant and wind-dominant classes for each technique are quite different (37% and 31% of cases classified identically for these classes respectively). Several other linkage techniques (e.g., "average" and "single") were susceptible to classifying outliers (e.g., 3 April 1974; 11 April 1965; 5 February 2008) and were considered inappropriate for the purposes of this study. Because of these findings, the *k*-means cluster analysis was the preferred technique for classification of severe weather events based on the characteristics of the severe reports.

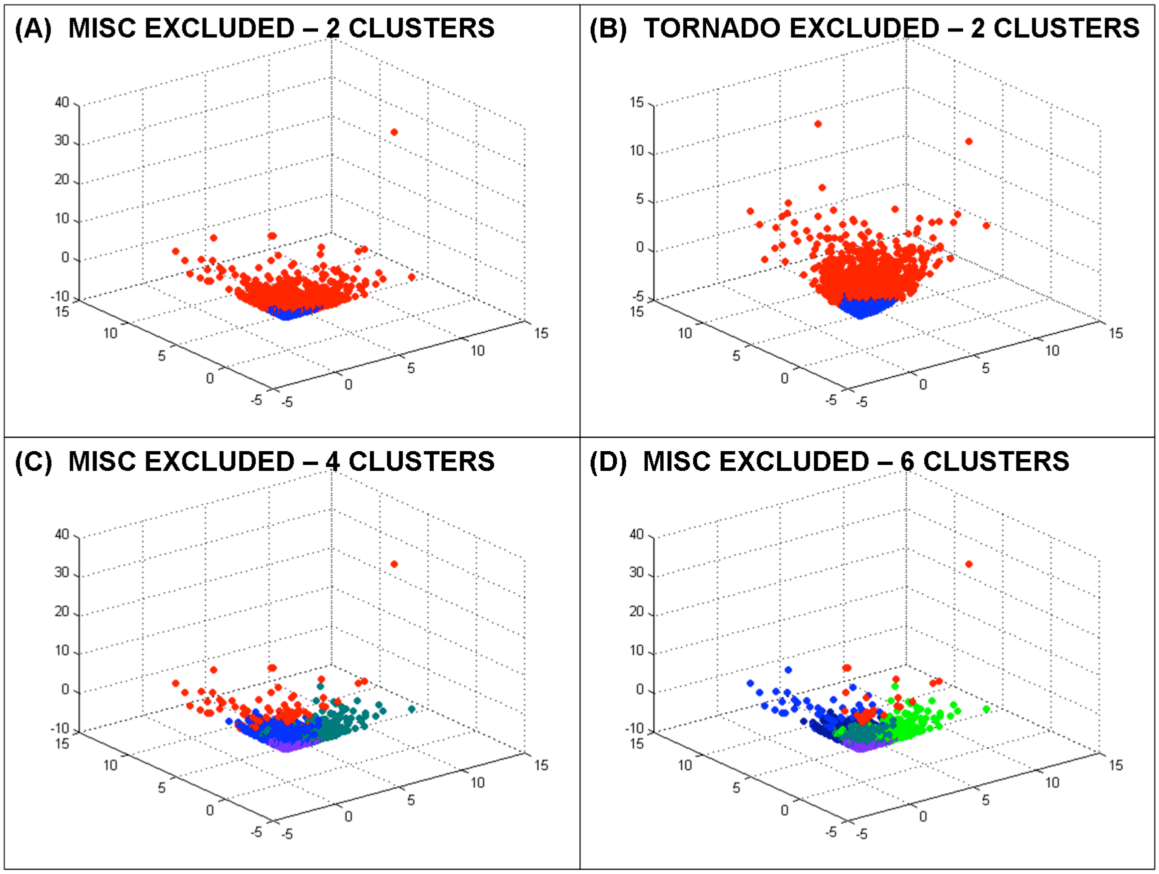


Fig. 3.15. As in Fig. 3.13, using the Ward's hierarchical clustering technique.

Cluster analysis also was performed on the total one-dimensional scores, for guidance on possible categorization of severe weather events based on relative severity. This was performed by taking the average scores of all of the indices for each rank from 1 to 6072 (the same computation used to create Fig. 3.9b). Ward's hierarchical and *k*-means cluster analyses were conducted on these scores. Once again, the other hierarchical techniques were susceptible to categorizing the outlier cases separately and generally were discounted. Analyses of silhouette plots and dendrograms (not shown) suggested a low number of clusters were favored. The resulting *k*-means and Ward's cluster analyses (not shown) provided little guidance as to preferential grouping of the events based on relative severity. The *k*-means cluster analysis showed high inter-cluster variability, and the Ward's technique identified events ranked higher and lower than the two regions where the all-tornado and two-tornado ranking indices intersected. The clusters of the Ward's technique were indicative of the characteristics of the indices developed and not of the outbreaks themselves.

Clear distinctions among various groups of severe weather events based on their relative severity were not found, suggesting that the severity of outbreaks is reminiscent of a spectrum rather than of separate bins. However, categorical distinction of these events would be beneficial from a forecasting standpoint and appears to be more feasible than predicting the index values for these events (Section 4). Thresholds distinguishing various categories of severity could be determined by testing various index values using diagnostic or prognostic meteorological variables [such as the energy helicity index (EHI; Hart and Korotky 1991), the significant tornado parameter (STP; Thompson et al. 2003), or other individual or combined meteorological parameters] and identifying which thresholds seem to perform optimally based on predetermined accuracy and/or skill criteria.

f. Summary and conclusions

This study is a follow-up to that of SD10a, in which a new technique to account for cases with large geographic scatter or multiple clusters of regionally separated severe weather reports was introduced, with the goal of developing a way to rank and classify severe weather events of any type. The technique proposed uses kernel density estimation to identify regions associated with a particular cluster of severe reports, rather than the use of the middle-50% parameter introduced in D06. By tuning the bandwidth and the threshold of the density estimation's approximation of the two-dimensional probability density function, severe reports within these regions were associated with the cluster. Cases in which the number of reports within the cluster, or the ratio of severe reports to grid points (based on a specified map projection) within the cluster, was lower than the detrended mean value on a given year were excluded from consideration. This process effectively excludes cases that feature large geographic scatter, but includes as separate events cases in which regionally-separated clusters exist on a given day, which was a limitation of the work by SD10a.

The selection of map projection and grid spacing should not result in substantial differences in the regions associated with a particular cluster, as long as modifications to the bandwidth and probability density function threshold are taken into account. This permits the use of relatively coarse grid spacing, though values well above 150 km are likely unwise to implement based on the magnitude of the selected bandwidths. Also, converting the severe reports to a grid (as in Brooks et al. 1998) versus using the point values does not change the results of the work in a substantial way.

After the severe weather clusters were identified, the procedure to rank and classify these events, in terms of severity and the characteristics of the severe reports respectively, was essentially identical to that of SD10a. The results were also similar, as major tornado outbreaks were the highest-ranked events. Up to 250 severe weather events appeared to be distinct from the remaining cases, as evidenced by a very steep slope for the scores of the indices for these cases (versus a gradual slope for the remaining cases). The rankings for the highest-ranked cases were relatively similar no matter which index was used, though important modifications were observed when a subset of the tornado variables were removed from the indices used to rank these events.

As expected (and desired), such removal permitted several severe weather events with few or no tornadoes to be included among the highest-ranked cases. As in SD10a, no index can be justified as optimal. Objectives of future research investigating severe weather outbreaks should dictate the selection of a specific index. For example, if the goal of a research project is to study differences of tornado outbreaks from all other types, use of indices that include a large number of tornado variables appears to be appropriate. However, if the goal is to distinguish major severe weather outbreaks from minor events no matter what type of event the case may be categorized as, use of indices with fewer tornado variables may be appropriate.

The rankings of the cases below the top 250 are much more volatile, as observed in SD10a. Subjective investigation of these events found that a large number of these cases are qualitatively similar in terms of the numbers and types of severe reports. Thus, the prediction of a severe weather event's rank is likely formidable, whereas predicting the categorical relative severity of an event is more feasible.

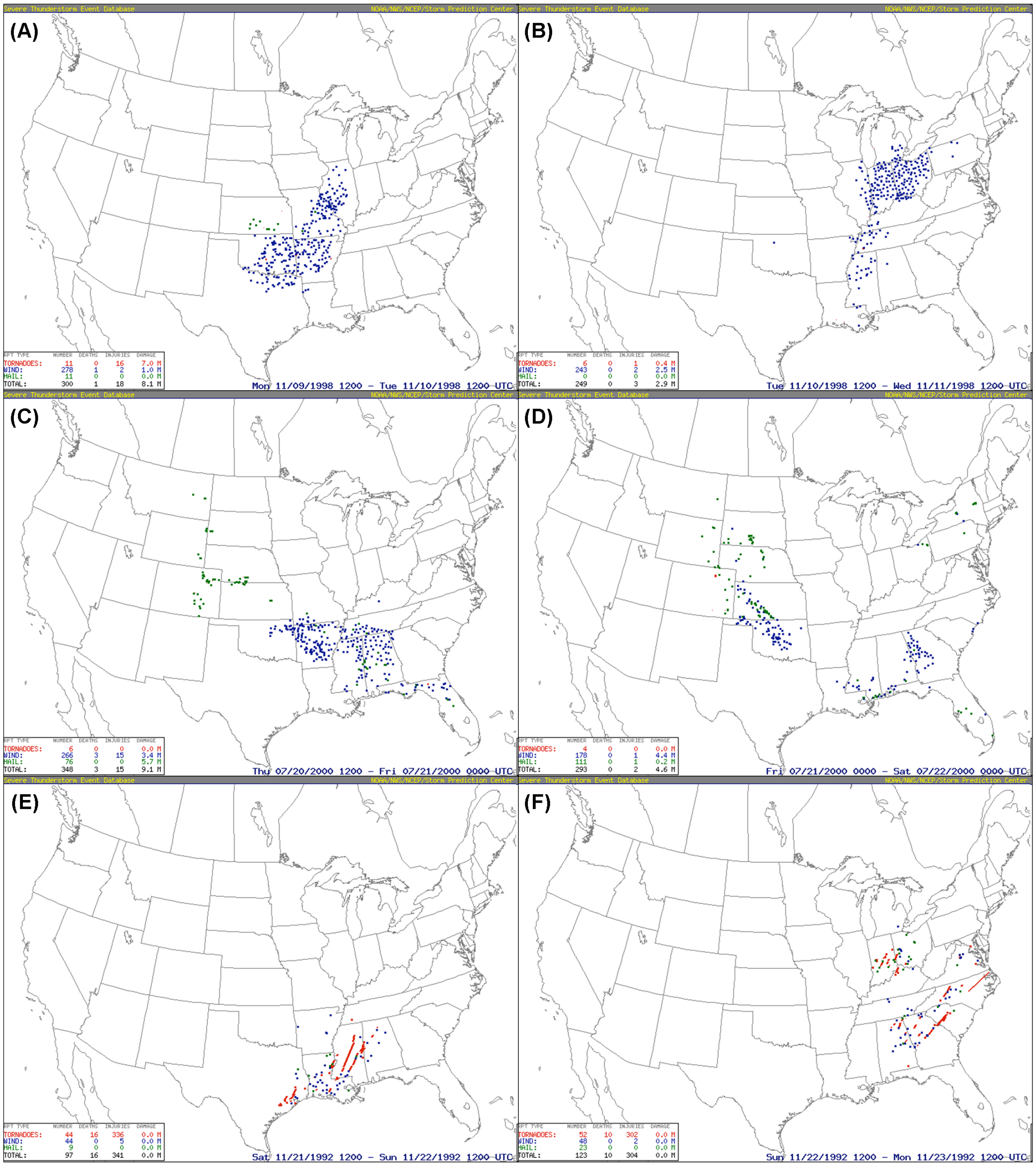
Binary classification of events based on their relative severity was quite similar for the *k*-means and Ward's hierarchical cluster analyses; however, differences between the techniques become substantial as the number of classes increases. Thus, the separation of the highest-ranked cases (approximately 200-250) from the remaining cases (~5900) is a recommended starting point for future work. Determination of an optimal threshold, based on the ability of meteorological covariates (see Brown and Murphy 1996) to distinguish these events, would also be appropriate.

Classification of severe weather events into various types based on the characteristics of the severe reports also resulted in categories similar to those found in

SD10a. In general, events could be classified as major tornado, hail-dominant, wind-dominant, or minor (mixed-mode) outbreak cases. Differences among the indices were very minor, whereas differences among various types of cluster analyses were more substantial. However, the 6-class categorization of events between the *k*-means and Ward's hierarchical cluster analyses were reasonably similar, in which the two additional classes roughly could be described as minor wind-dominant and minor hail-dominant events.

Although the kernel density estimation method appears to be an effective means of accounting for days with large geographic scatter and days with multiple clusters of severe reports, some limitations of the ranking techniques remain. For example, the selection of a 24-h period for which to analyze events independently leads to the possibility of misrepresenting events that occur at the end of one period and the beginning of another (e.g., Figs. 3.16a,b). The current method would identify such a circumstance as two separate events and would likely underestimate the severity of the event. Though these examples are quite rare in the dataset, future work is planned to try to account for these events.

Furthermore, multiple events can occur in the same region within a 24-h period (e.g., Figs. 3.16c,d). The current method would consider this a single event and would overestimate its severity. These two limitations suggest that a time dimension should be added to the density estimation technique; however, its inclusion presents challenges that require substantial investigation before implementation. Additionally, objective identification, ranking, and classification of multi-day events associated with a single synoptic-scale system (e.g., Figs. 3.16e,f) are prudent and would provide a valuable


Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(A)
(B)
(C)
(D)
(E)
(F)
Mon 11/09/1998 1200 – Tue 11/10/1998 1200 UTC
Tue 11/10/1998 1200 – Wed 11/11/1998 1200 UTC
Thu 07/20/2000 1200 – Fri 07/21/2000 0000 UTC
Fri 07/21/2000 0000 – Sat 07/22/2000 0000 UTC
Sat 11/21/1992 1200 – Sun 11/22/1992 1200 UTC
Sun 11/22/1992 1200 – Mon 11/23/1992 1200 UTC

(A)

RPT TYPE	NUMBER	DEATHS	INJURIES	DAMAGE
TORNADOES:	11	0	16	7.0 M
WIND:	278	1	2	1.0 M
HAIL:	11	0	0	0.0 M
TOTAL:	300	1	18	8.1 M

(B)

RPT TYPE	NUMBER	DEATHS	INJURIES	DAMAGE
TORNADOES:	6	0	1	0.4 M
WIND:	243	0	2	2.5 M
HAIL:	0	0	0	0.0 M
TOTAL:	249	0	3	2.9 M

(C)

RPT TYPE	NUMBER	DEATHS	INJURIES	DAMAGE
TORNADOES:	6	0	0	0.0 M
WIND:	266	3	15	3.4 M
HAIL:	76	0	0	5.7 M
TOTAL:	348	3	15	9.1 M

(D)

RPT TYPE	NUMBER	DEATHS	INJURIES	DAMAGE
TORNADOES:	4	0	0	0.0 M
WIND:	178	0	1	4.4 M
HAIL:	111	0	1	0.2 M
TOTAL:	293	0	2	4.6 M

(E)

RPT TYPE	NUMBER	DEATHS	INJURIES	DAMAGE
TORNADOES:	44	16	336	0.0 M
WIND:	44	0	5	0.0 M
HAIL:	9	0	0	0.0 M
TOTAL:	97	16	341	0.0 M

(F)

RPT TYPE	NUMBER	DEATHS	INJURIES	DAMAGE
TORNADOES:	52	10	302	0.0 M
WIND:	48	0	2	0.0 M
HAIL:	23	0	0	0.0 M
TOTAL:	123	10	304	0.0 M

Fig. 3.16. As in Fig. 3.1a, for (a) 9 November 1998, (b) 10 November 1998, (c) 1200 UTC 20 July 2000 to 0000 UTC 21 July 2000, (d) 0000-1200 UTC 21 July 2000, (e) 11 November 1992, and (f) 12 November 1992.

resource for research investigating these events. This work is beyond the scope of the current study, however.

The technique presented herein is not the only way in which severe weather events (outbreaks) could be identified, ranked, and classified. The choices made in this study were subjective, but were designed to be (1) reproducible, (2) simple to implement and interpret, (3) effective in reducing the impact of nonmeteorological artifacts in the dataset, and (4) capable of identifying geographically clustered events and eliminating the other cases. Although the scheme is not shown herein to be optimal, the results of this study indicate that these four criteria have been met, and the method can be modified according to the objectives of future research investigating severe weather outbreaks or can be implemented in other meteorological research (such as flash floods, winter storms, hurricanes, etc.).

4. Using areal coverage to identify the major severe weather outbreaks and comparing to SPC categorical outlooks

a. Introduction

The revised ranking technique discussed in Section 3 allows for many more cases to be included in discrimination investigations. Instead of the 840 outbreak cases that were analyzed using the areal coverage and principal component analysis techniques (Section 2) from the period 1979-2006, a set of 4057 outbreak cases is available (from the period 1979-2008). Furthermore, the techniques incorporated in Section 3 preclude cases with excessive geographic scatter in the reports or multiple spatially distinct clusters of reports associated with multiple synoptic- or subsynoptic-scale systems. In short, the new case list is a substantial improvement in terms of sample size and relevant events to examine.

An additional benefit of the expanded case list is the ability to compare objective techniques to operational categorical forecasts of severe weather outbreaks. Specifically, the Storm Prediction Center issues categorical risks of severe weather multiple times a day for the same period of time considered when ranking the cases (that is, 1200 UTC on the nominal date to the same time the following day[14]). The categorical risks are high, moderate, slight, "see text", and "no organized severe weather expected". These outlooks are defined in terms of probabilities of various types of severe weather occurring within a certain distance (~40 km) from a particular point (see http://www.spc.noaa.gov/faq/ and http://www.spc.noaa.gov/misc/prob_to_cat_day1_seetext.jpg for more information). Thus, evaluation of SPC categorical outlooks using the ranking scheme as developed in

[14] However, the time period changes for Day-1 categorical outlooks issued after the valid starting period. For example, the 1630 UTC Day-1 outlook is only valid from 1630 UTC on the nominal date to 1200 UTC the following day.

Section 3 is not based on the strict definitions of the categorical outlooks. However, higher rankings clearly are tied to events in which the probabilities of severe weather occurring within a certain distance of a point increases, based on the implementation of the kernel density estimation technique (see Brooks et al. 1998).

One goal of the research conducted thus far is to provide a new means of guidance for operational forecasters in future potential outbreak situations. Though the research completed to this point only completes the first steps toward this implementation, a clear method of evaluation is to determine if any objective techniques developed with the purpose of discriminating the relative severity of the outbreaks perform as well as or better than current operational forecasts of these events. Because of disparities between the outbreak ranking scheme and the categorical outlook criteria, this task is somewhat complicated. Rather than verifying the objective techniques proposed herein using the specific probability criteria defined by the SPC for categorical outlooks, the two modes of diagnosis/prognosis will be evaluated based on the *scores used to rank the events*. Therefore, it is appropriate to note that the SPC categorical outlooks were not designed for this specific task, whereas the objective techniques introduced herein were developed with this type of discrimination as the primary goal.

Perhaps a more formidable complication is the fact that the atmosphere does not produce severe weather outbreaks that are easily classifiable. That is, severe weather outbreaks occur as a spectrum of events rather than as distinct bins. This makes defining a major severe weather outbreak difficult and open to controversy. As a result, a primary objective of this study is to investigate what threshold values of outbreak ranking scores, as developed in Section 3, SPC categorical outlooks and objective techniques seem to be most accurate and/or skillful.

Additionally, analyses will be used to evaluate the objective techniques rather than model forecast fields. This decision was based on two characteristics of the research project. (1) Because of the relatively large sample of cases for these types of studies, it was deemed appropriate to determine if characteristics of the *analyses* were capable of distinguishing outbreak type before running model simulations of such a large number of cases. (2) Several of the Day 1 convective outlooks issued by the SPC are issued just before or during the event, making such outlooks short term forecasts (i.e., shorter in length than the latest available model forecasts). Future research, pending the results of the objective technique diagnoses, then would assess model simulated fields of environmental variables and compare to longer-term categorical outlooks.

This study not only serves as a comparison between new proposed objective techniques and current operational forecasts, but it also serves as a current "state of the science" in diagnosing the severity of outbreaks based on the fields of severe weather parameters typically analyzed to predict these events. As I am not aware of any previous research that has developed objective techniques to diagnose the relative severity of outbreaks, besides those of M09, S09, S10a, S10b, and Shafer et al. (2010c), the work discussed in this section will provide one of the first comprehensive evaluations of our ability to diagnose and/or forecast the nature of severe weather outbreaks.

In Section 4b, the data and methods for comparing the SPC categorical outlooks with the areal coverage technique (the objective technique for which this particular study is focused) will be described. Section 4c examines the areal coverage technique using all of the cases available from 1979-2008. Section 4d analyzes the cases available from 2003-2008, comparing the areal coverage technique and the SPC categorical outlooks, and compares the results to those using the whole dataset. Section 4e discusses the implications of the findings.

b. Data and methods

As in S10a, the NARR dataset was used to determine the areal coverage associated with a particular severe weather event. The valid time for each event is 1200 UTC on the nominal date to 1159 UTC the following day, as in previous studies. The valid time of each case was determined by selecting the first available time before the median time of the reports associated with a particular case. As Section 3 illustrated, each case was associated with a region (as in Fig. 4.1a), in which each report falling within that region was associated with that event during the 24-h period. As NARR data are available from 1979 onward, each event from 1 January 1979 to 31 December 2008 using the 1° latitude-by-longitude map projection and grid point method, as described in Section 3, was considered (4057 total).

SPC categorical outlooks were obtained from their public website (http://www.spc.noaa.gov/products/outlook/) for the period 23 January 2003 to 31 December 2008 and were directly compared with areal coverage diagnoses for events occurring within this time frame. As multiple events can occur on the same day, only the highest-ranked events were considered for comparison (with the assumption that the most severe categorical outlook was associated with the highest-ranked case on a given day). For this analysis, a total of 727 of the 4057 cases qualified. Categorical outlooks of moderate and high were considered to be indicative of forecast major severe weather

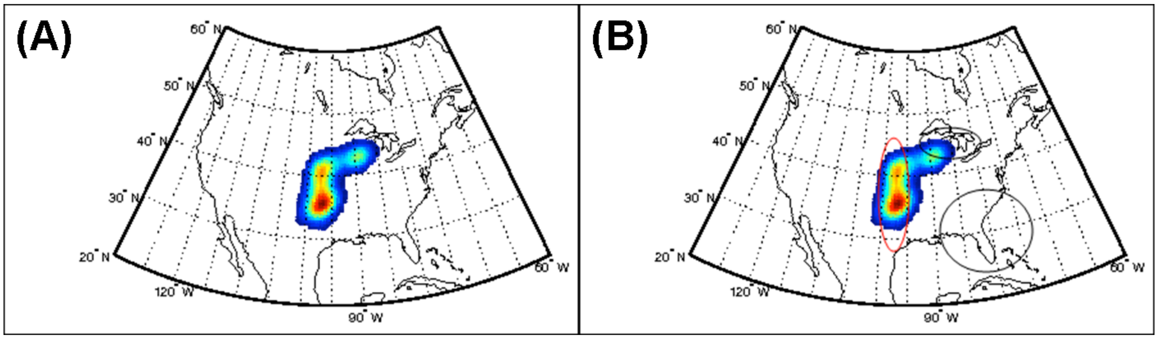


Fig. 4.1. (a) Region identified by the KDE method associated with the 13 March 1990 severe weather outbreak, using a 1° latitude-by-longitude map projection and the grid point method (see Section 3 for details). (b) As in (a), with three hypothetical regions indicating where a selected severe weather parameter exceeds a threshold. The area within the red oval is the largest area of the three that intersects the region identified by the KDE method. The intersect method would select the region in the red oval as the region associated with the severe weather outbreak.

outbreaks, whereas slight, "see text", and no-risk days were considered forecasts of less significant (null) events. However, comparison of the areal coverage technique to the SPC categorical outlooks using slight and high risk days as the threshold forecast of major severe weather outbreaks also was conducted and will be shown in Section 4c.

Several methods of computing areal coverage were considered, but only two will be shown in this paper. The first method is referred to hereafter as the *kernel density estimation (KDE) method*. For this method, a severe weather parameter is calculated at each grid point in an 18-km 300x200 CONUS-centered domain (the same domain used in S10a). The average, median, or sum value of the severe weather parameter considered is computed, using each grid point associated with the severe weather event, as described in Section 3. For example, each grid point shaded in Fig. 4.1a would be considered for the computation of average, median, or sum values of the severe weather parameter selected to discriminate the outbreaks. The final value would be used as a diagnosis of the severity of the outbreak, with a predetermined threshold used as the criterion for diagnosis. If the average, median, or sum value exceeds the threshold, the case would be diagnosed as a major severe weather outbreak.

The second method hereafter is referred to as the *intersect method*. For this method, a threshold value of a severe weather parameter is selected. All contiguous regions in which grid points exceed this value were determined, and the largest region in which the threshold is exceeded that intersects at least one point of the region identified by the KDE technique as associated with the event was selected as the region of interest. Once again, the value of the severe weather parameter at each grid point within this "intersect region" (e.g., the region within the red oval in Fig. 4.1b) was computed, and the average, median, or sum value for all of the grid points considered was obtained. This final value was used as the diagnosis of the severity of the outbreak.

The development of the intersect method, in addition to the KDE method, was necessary because forecasters do not know the exact region associated with the outbreak before the event occurs. In essence, the KDE method is strictly diagnostic in that it cannot be directly applied to forecast guidance. The intersect method, although diagnostic for the results shown herein, easily can be implemented in an operational setting – as contiguous regions of severe weather parameters exceeding a specified value in a region of interest can be identified readily using model forecast fields. The hypothesis driving the research shown in subsequent sections is that the KDE method should verify better than the intersect method, but the intersect method more likely would be implemented as guidance in an operational environment. Recent research investigating the utility of various model fields to identify regions associated with severe weather reports (the so-called surrogate severe reports technique) may permit the use of a technique similar to the KDE method in future studies, however (see Schwartz et al. 2010; Sobash et al. 2010).

An obvious weakness of the intersect method is that the region selected may not be an area associated with the event. However, any method developed is subject to some limitations. For example, regions in which a severe weather parameter exceeds a specified threshold may not be associated with *any* observed severe weather. Thus, considering every contiguous region in which a severe weather parameter exceeds a threshold is limited in this manner. These "limitations" are daily decisions by severe weather forecasters and should be considered inherent in any objective technique developed.

Although multiple indices were developed to rank the outbreaks, the results herein only consider the N15 index, as described in Section 3 (and Shafer and Doswell 2010b). As with the studies by S10a and Shafer et al. (2010c), the selection of N15 was essentially arbitrary, though the index was susceptible to ranking tornado outbreaks highest and did not negatively bias events with a predominant nontornadic report type. Sample curves for the rankings, including the N15 index, are shown in Fig. 3.9a. Although this chart considers all 6072 cases from the 1960-2008 period, the curves for the 4057 cases considered when using the NARR data fit a similar curve (not shown). Several subsequent figures also will illustrate the general nature of the scores used to rank the indices as well (Sections 4c,d).

The parameters analyzed were the same as those in past studies (S09, M09, S10a, S10b). Emphasis herein will be on the supercell composite parameter (SCP) and significant tornado parameter (STP). A discussion regarding the correlations among the variables (Section 4e) will address the lack of emphasis on using multiple variables in this study. The areal coverage technique and the SPC categorical outlooks were evaluated using binary contingency statistics, as in S10a and Section 2 herein. In addition to the emphasis on selecting values of areal coverage that score best when selecting a particular value of the outbreak ranking index score used to classify events as "major" or "null" events, the results presented in this study will focus on iteratively selecting ranking index thresholds using a specific threshold of areal coverage (or SPC categorical outlook) to determine the outbreak ranking index threshold with which these methods perform best. These two approaches should provide a (more) comprehensive analysis of the ability of areal coverage and current operational forecasts to identify the most significant severe weather outbreaks.

c. Areal coverage results 1979-2008

Diagnoses of major severe weather outbreaks using specific areal coverage thresholds when using the sum values of STP and SCP using the KDE and intersect methods (Figs. 4.2 and 4.3) indicate a tendency for the most significant severe weather outbreaks to be identified correctly. Specifically, most of the outbreaks on the upper portion of Fig. 4.2 and on the right portion of Fig. 4.3 were diagnosed correctly as major severe weather outbreaks (red dots). However, many events are diagnosed as major severe weather outbreaks that score similarly to less significant events (as indicated by the number of red dots in the vicinity of the blue dots in Fig. 4.2, and the number of red dots on the left portion of each panel in Fig. 4.3). This is a clear signal of a false alarm problem, as previous sections have discussed.[15]

[15] The selection of the specific thresholds of areal coverage in Figs. 4.2 and 4.3 will be discussed in Section 4d.

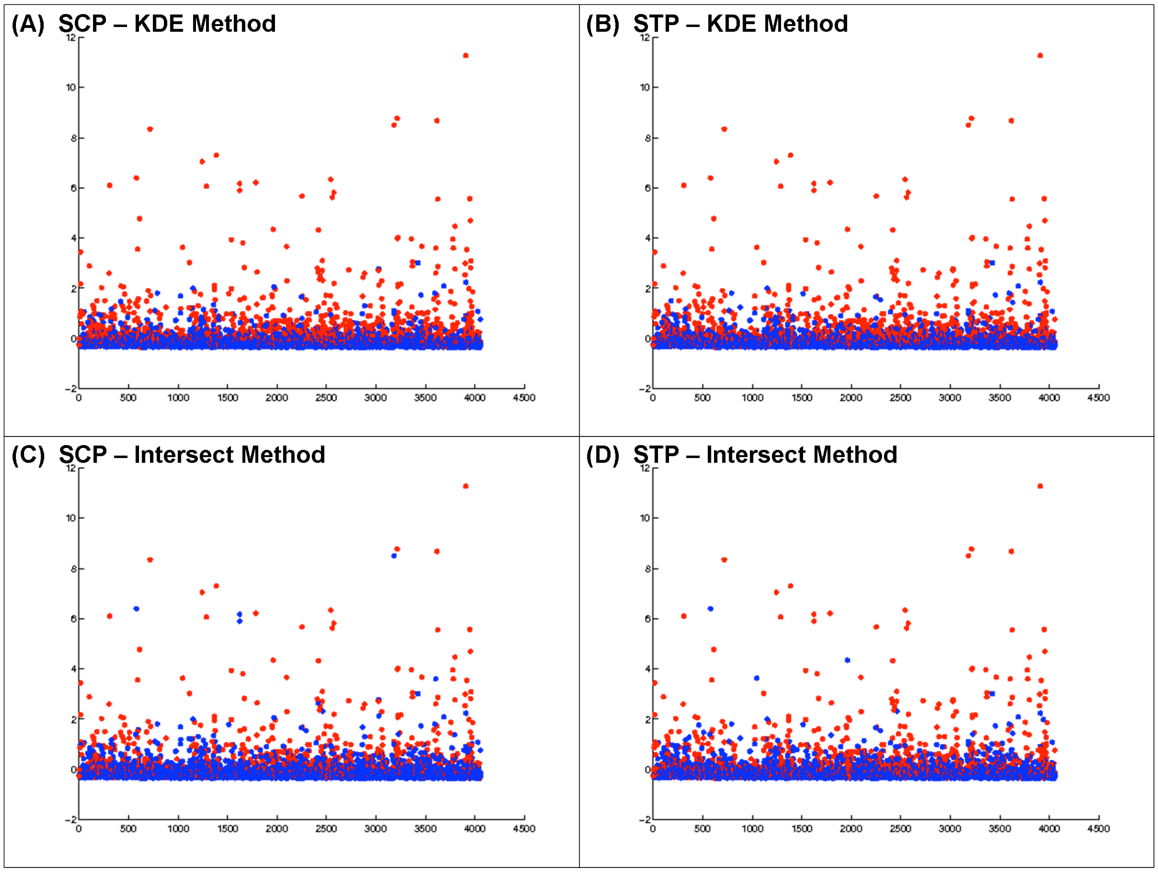


Fig. 4.2. Scatter plots showing diagnoses of major severe weather outbreaks (red dots) or null cases (blue dots) using areal coverage thresholds, with the outbreak ranking index scores on the *y*-axis and the case identification number (chronological order) on the *x*-axis. In (a), the SCP using the KDE method is used, with an areal coverage threshold of 7500. In (b), the STP using the KDE method is used, with an areal coverage threshold of 1500. (c) As in (a), using the intersect method, SCP threshold of 1, and an areal coverage threshold of 15 000. (d) As in (b), using the intersect method, STP threshold of 1, and an areal coverage threshold of 1500.

Additionally, the intersect method is susceptible to an increased number of misses with major severe weather outbreaks. There are a few cases, using SCP or STP as the diagnostic variable, in which the intersect method diagnoses null events for cases with N15 scores above the value of 2. This contrasts with the KDE method diagnoses for these cases, which are almost entirely red dots above index values of 2. These cases appear to be examples in which the SCP and STP contiguous regions may not be most associated with the event itself, or may indicate that the selected threshold (unity for Figs.

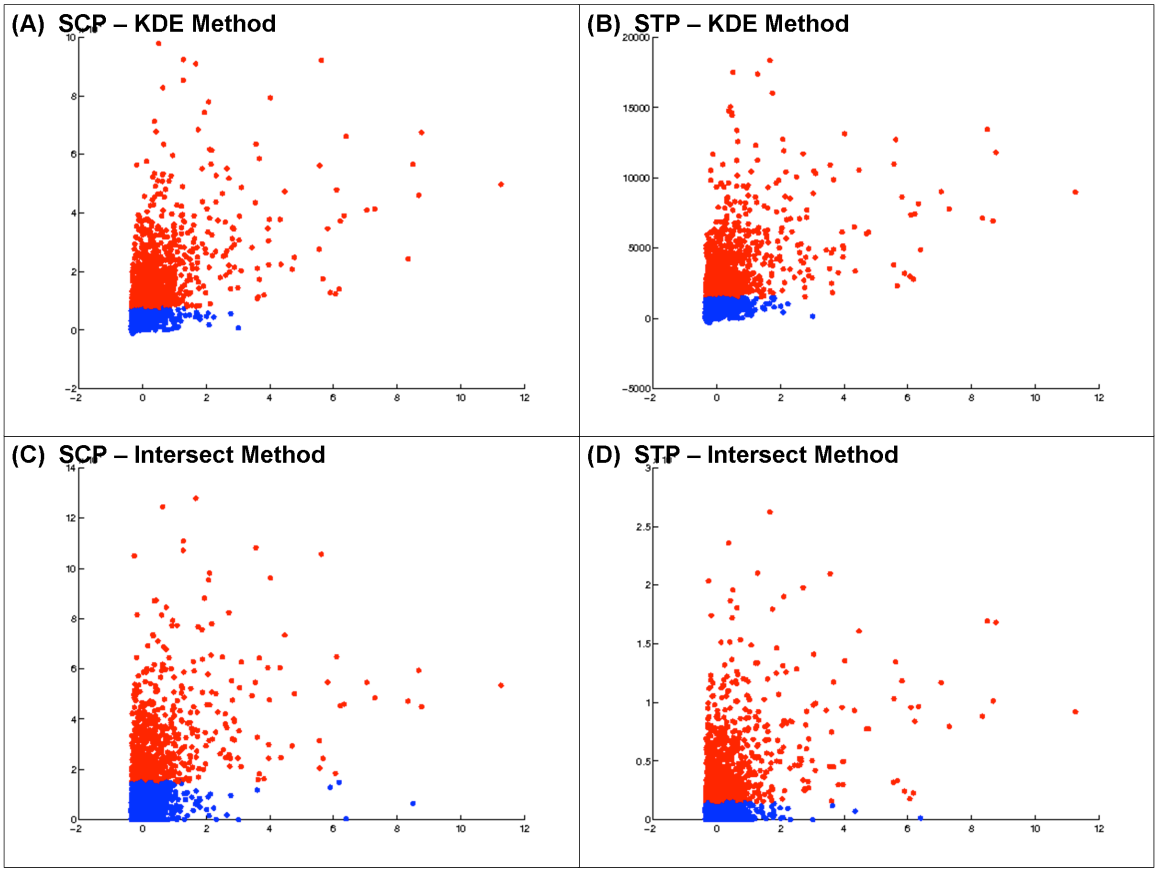


Fig. 4.3. As in Fig. 4.2, with the areal coverage on the y-axis and the outbreak ranking index score on the *x*-axis. Major severe weather outbreaks are on the right side of each figure (i.e., higher ranking index scores).

4.2c,d; 4.3c,d) is too high for correct classification of these events. Indeed, a lower threshold results in correct diagnoses for these cases, but at the expense of an increased number of false alarms (not shown).

Contingency statistics (Figs. 4.4-4.6; with abbreviations as in Section 2) suggest that the threshold values selected for diagnoses of outbreak type tend to work best when the outbreaks are classified as major events with index scores above values of around 0. Scores of zero are around the mean value of the ranking index scores for the 6072 cases considered from 1960-2008. Section 4e will discuss the implications of this in more detail. Additionally, characteristics of cases with various scores of the N15 index will be

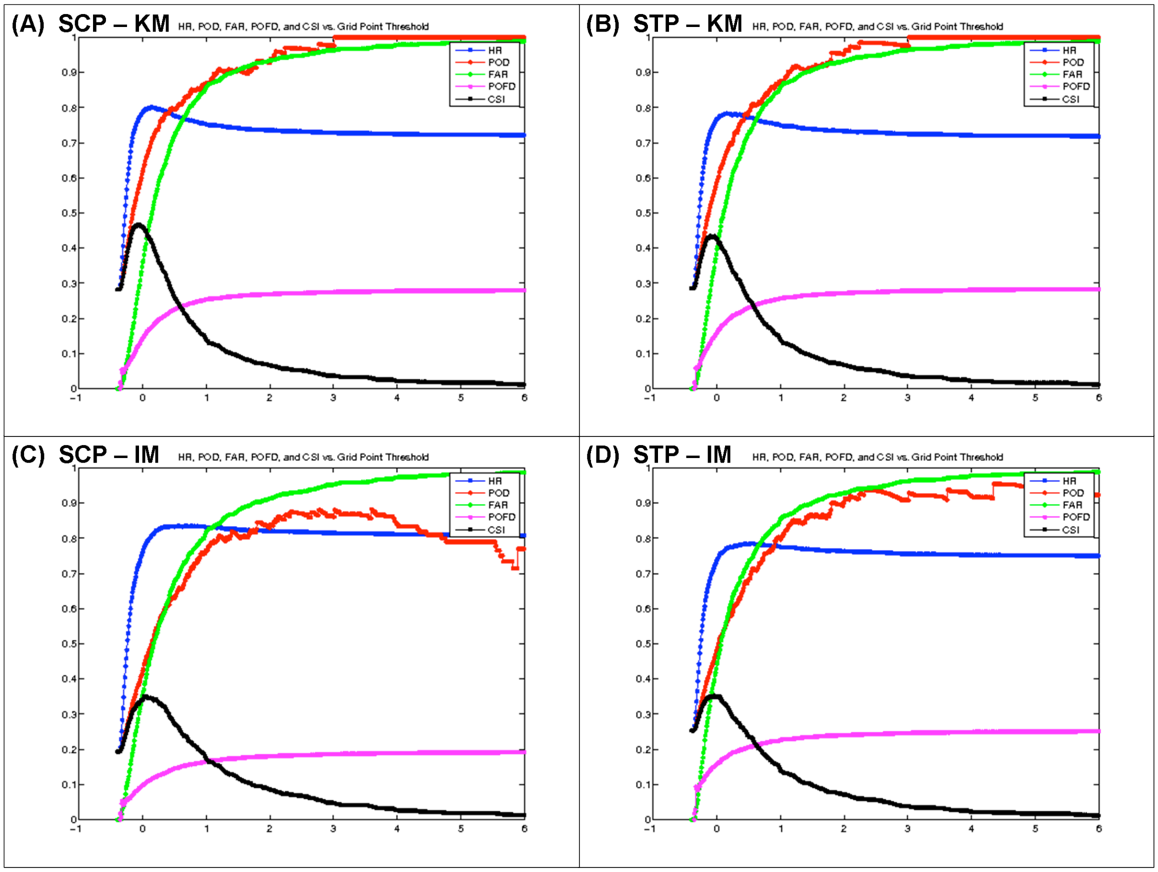


Fig. 4.4. Plots of hit rate (HR – blue), probability of detection (POD – red), false alarm ratio (FAR – green), probability of false detection (POFD – magenta), and critical success index (CSI – black) as a function of outbreak ranking index score classification thresholds (-0.4 to 6 in increments of 0.01, where outbreaks scoring above the threshold are considered major severe weather outbreaks), with the same variables and methods (KM is KDE method; IM is intersect method) as in Fig. 4.2.

examined, to determine what type of cases seem to be near the "best" thresholds for major versus less significant severe weather outbreaks.

With the four specific examples of severe weather parameters and areal coverage methods in Figs. 4.4-4.6, the accuracy of the diagnoses tend to be highest at index thresholds around the value of 0. This is generally where the CSI is highest, the skill scores (besides PSS) peak, and the bias tends to be around 1. As Fig. 4.4 indicates, the FAR is very high for index thresholds much greater than zero, with the FAR generally higher than the POD when the intersect method is used. The Roebber (2009) diagrams

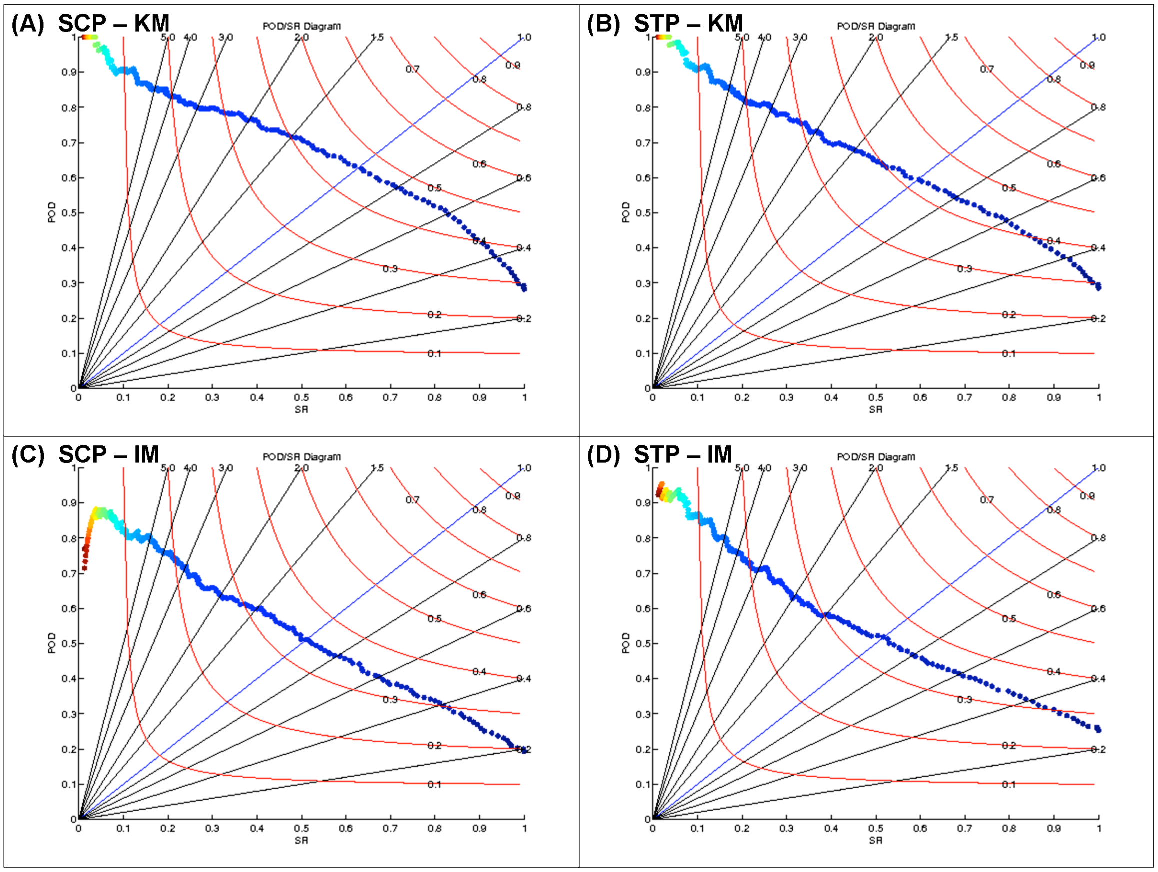


Fig. 4.5. Roebber (2009) diagrams for the same variables and areal coverage methods as in Fig. 4.2. Each dot on the diagrams represents the values of POD, success ratio (SR; 1 – FAR), bias, and CSI for a specific outbreak ranking index score threshold. The darkest blue dot uses a threshold of -0.4, whereas the darkest red dot represents a threshold of 6. Thresholds are iterated by a factor of 0.01.

(Fig. 4.5) and skill score charts (Fig. 4.6) clearly suggest the superior performance of the KDE method, as expected, with higher CSIs, a bias near unity, and skill scores generally at or above 50% around index thresholds of zero. For these same thresholds, the skill scores are lower (near or less than 40%) using the intersect method. The relatively poor statistics for the intersect method appear to be primarily because of a reduction of the POD, as indicated by the larger number of blue dots at higher thresholds in Figs. 4.2 and 4.3 and by the lower POD curves in Figs. 4.4c,d compared to Figs. 4.4a,b.

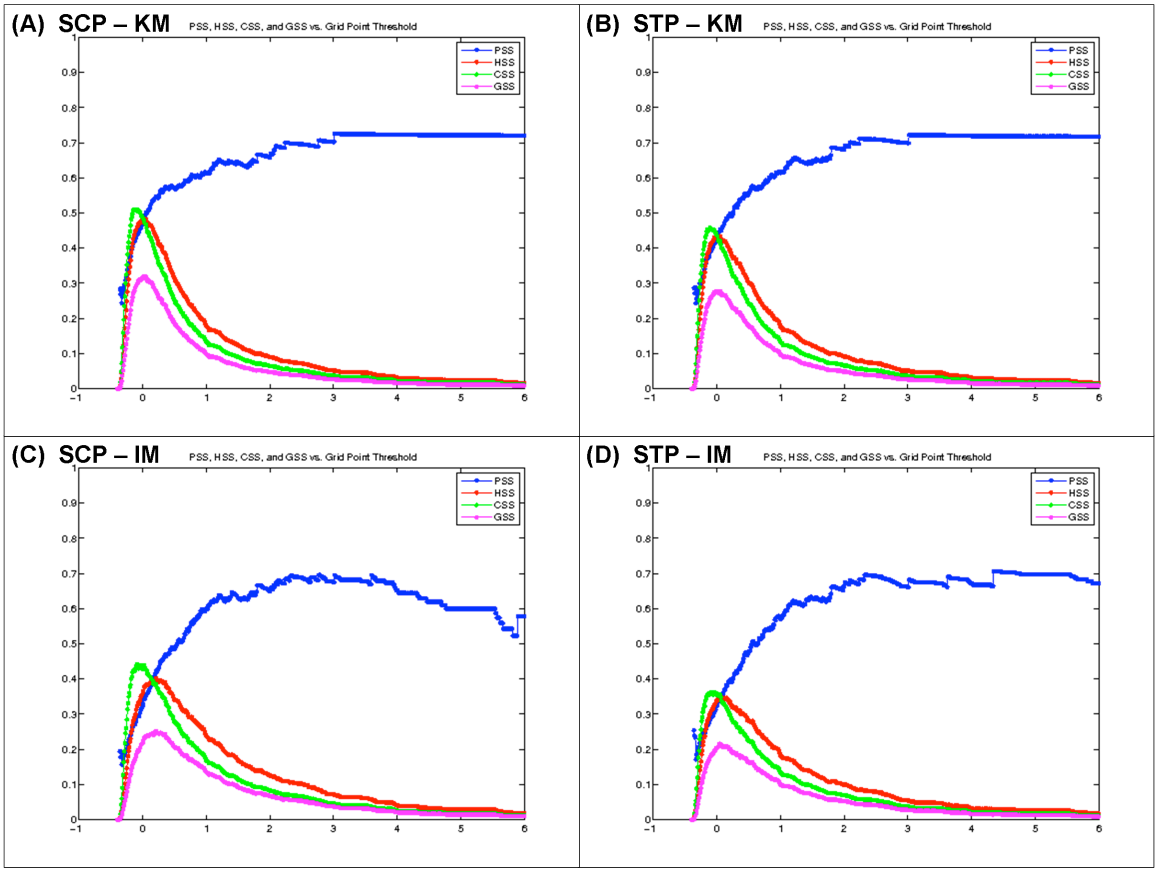


Fig. 4.6. As in Fig. 4.4, except for the Peirce skill score (PSS – blue), the Heidke skill score (HSS – red), the Clayton skill score (CSS – green), and the Gilbert skill score (GSS – magenta).

Figure 4.6 illustrates the undesirable characteristic of the PSS in a rare-events dataset. If major severe weather outbreaks are considered to be cases exceeding very high ranking score thresholds, the PSS improves substantially. This is as a result of the propensity for these cases to be classified correctly most of the time (i.e., high PODs), and the number of correct nulls is very large (i.e., low POFDs – see Fig. 4.4).

If the outbreak ranking index score of 0 is used as the threshold for classifying outbreaks as major or null events, the optimal areal coverage threshold can be determined for each severe weather parameter and areal coverage technique tested (as in Figs. 4.7-4.9). In general, the thresholds selected for Figs. 4.2-4.6 are near the maximum CSI (for

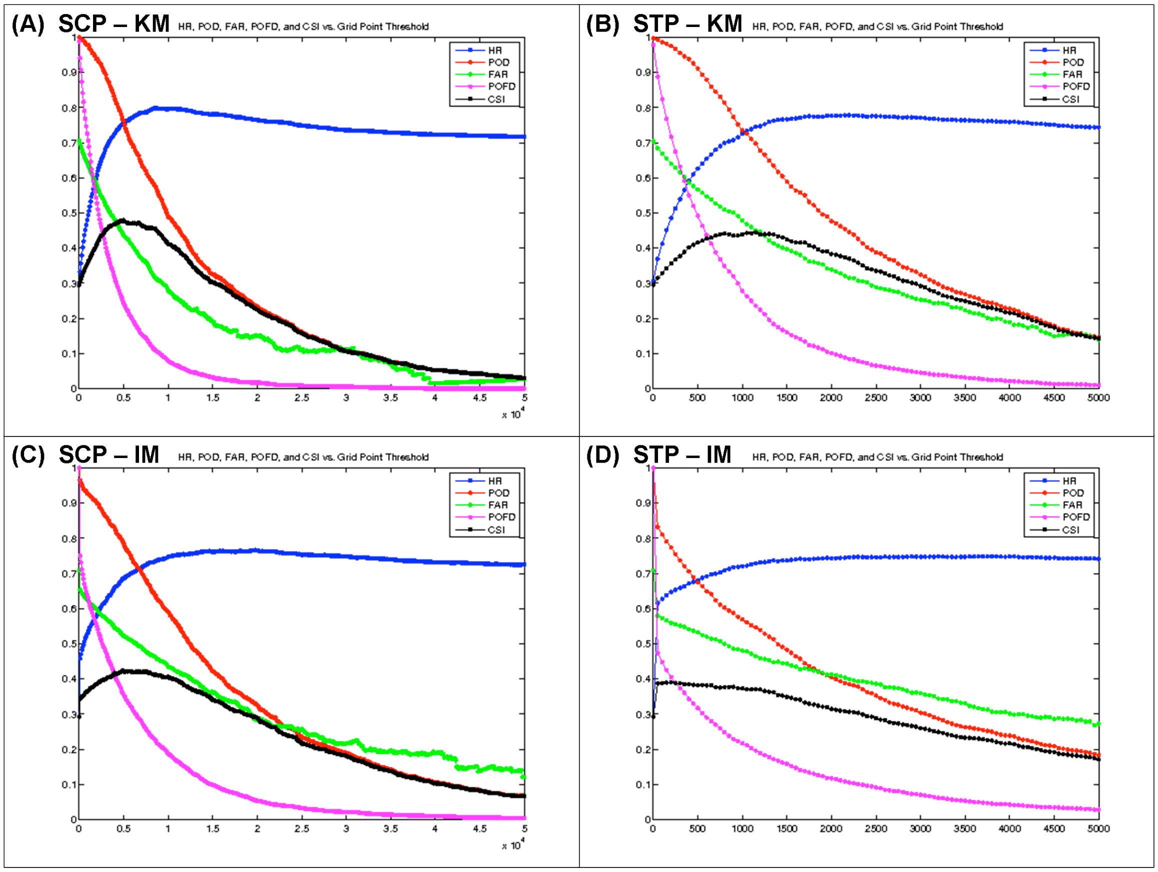


Fig. 4.7. As in Fig. 4.4, except the contingency statistics are shown as a function of the areal coverage for the outbreak ranking index threshold of 0.

a bias of unity) and skill scores for the KDE method. However, the thresholds for the intersect method may prove to verify somewhat better with lower thresholds (~10 000 for SCP, 1250 for STP). In Section 4d, these thresholds will be used for comparison with SPC categorical outlooks. As previously discussed, results clearly worsen when the intersect method is used versus the KDE method.

The trends in the statistics with increased areal coverage threshold differ from those with increased outbreak ranking score threshold. Specifically, the POD, POFD, and FAR decrease with increased threshold (because the outbreaks are classified consistently by using the index score of 0 as the threshold criterion). As higher areal coverage occurs with fewer and fewer cases, the number of misses (false alarms) should

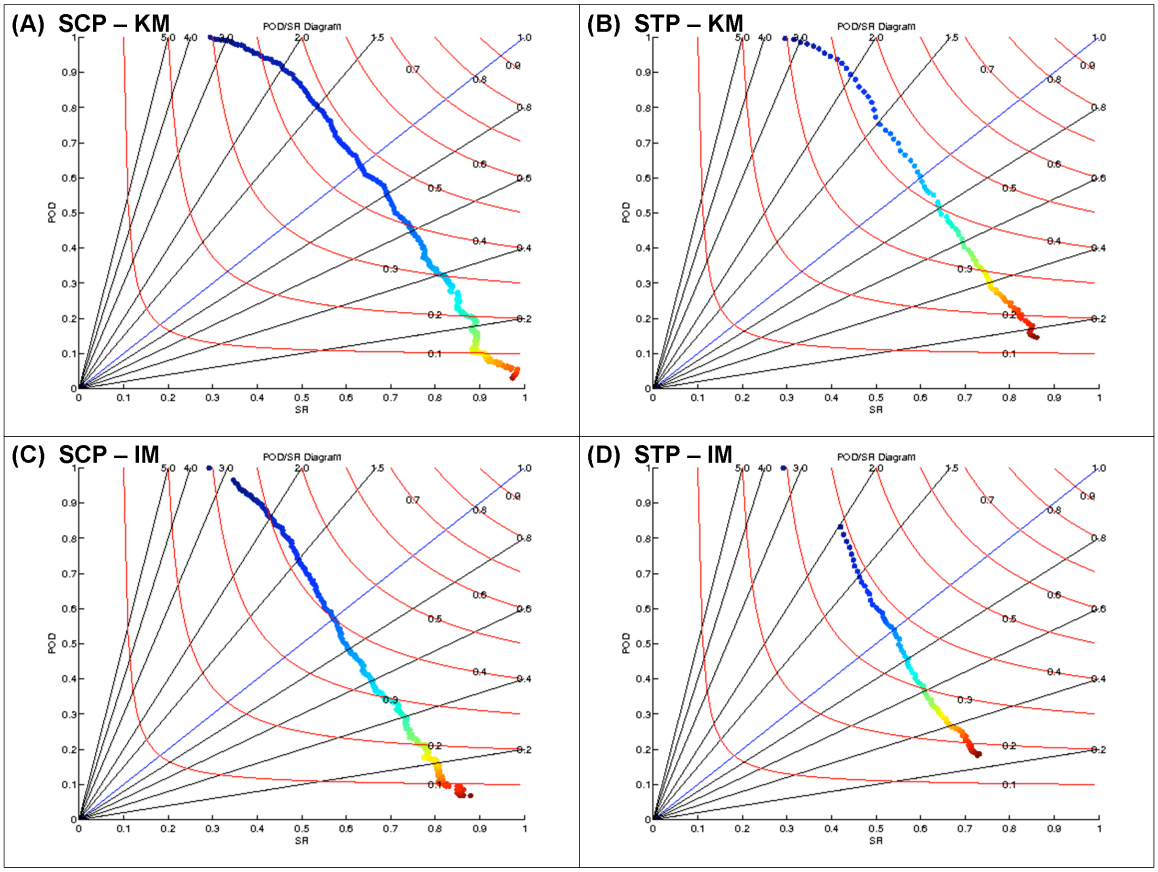


Fig. 4.8. As in Fig. 4.5, except each dot represents an areal coverage threshold. For SCP, the thresholds are incremented by 100. For STP, the thresholds are incremented by 50. The range of thresholds is 0 – 50 000 for SCP and 0 – 5000 for STP, as in Fig. 4.7.

be expected to increase (decrease), resulting in the observed trends. Moreover, the PSS tends to be higher at the lowest areal coverage thresholds (compared to HSS and GSS), whereas the CSS exhibits a similar trend to the PSS when the skill scores are analyzed compared to the incremental increase in outbreak ranking score threshold. Thus, the PSS probably should be avoided when determining the “best” threshold for which outbreaks are classified based on a selected value of areal coverage, whereas the CSS probably should be neglected when determining the “best” areal coverage threshold based on a selected value of outbreak ranking score threshold.

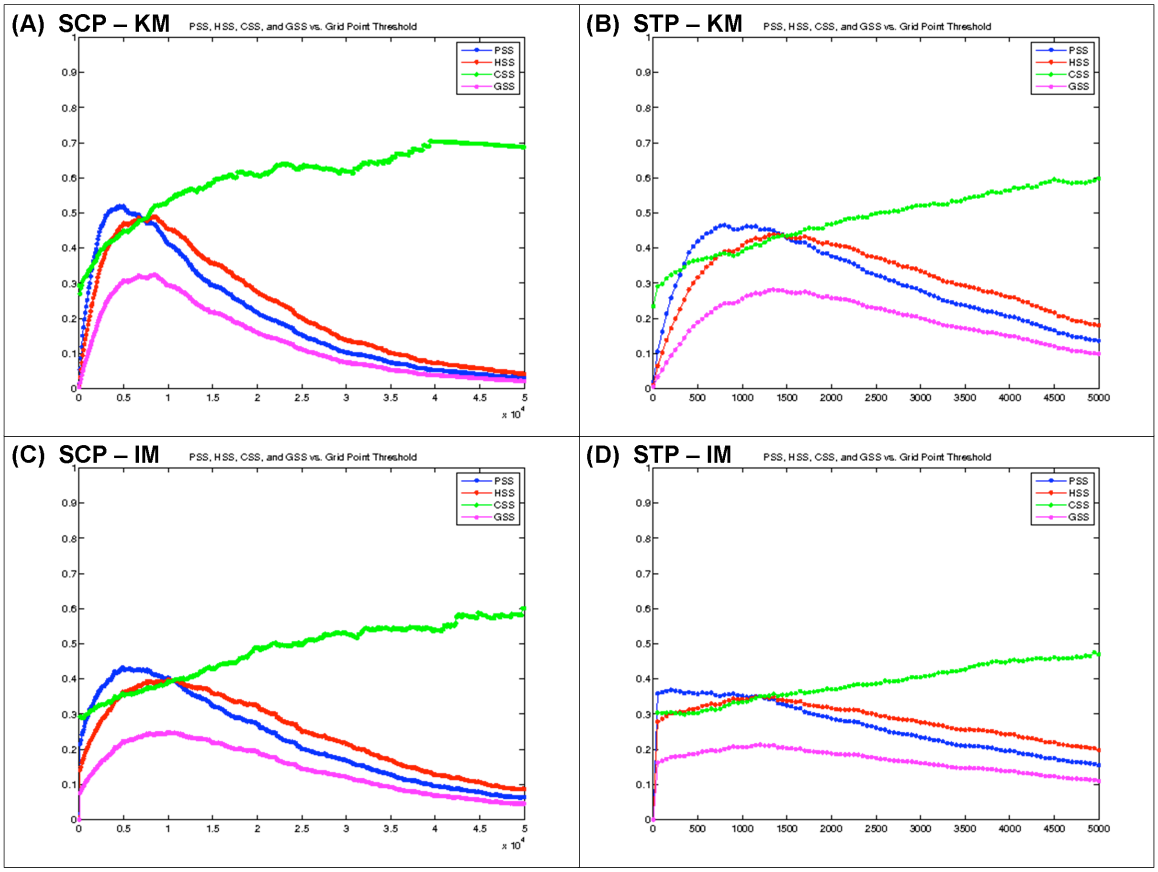


Fig. 4.9. As in Fig. 4.6, except the contingency statistics are shown as a function of the areal coverage for the outbreak ranking index threshold of 0.

Note that in Figs. 4.6 and 4.9, the PSS equals the CSS at the approximate thresholds in which the HSS and GSS peak. The PSS and CSS are equal when the number of correct hits (*a*) equals the number of correct nulls (*d*). Using the equation for HSS:

$$HSS = \frac{(a+d) - E_c}{N - E_c} \tag{4.1}$$

where

$$E_c = \frac{1}{N}((a+c)(a+b)+(c+d)(b+d)), \tag{4.2}$$

b is the number of false alarms, c is the number of misses, and N is the total number of observations (or forecasts). Note that HSS is a maximum when ($a + d$) is large (i.e., the accuracy is large). For a constant ($a + d$) – as approximately demonstrated for relatively large values of outbreak ranking score thresholds or areal coverage thresholds (see Figs. 4.4 and 4.7) – the product ad is a maximum when $a = d$. Equation (4.1) can be modified to show that the numerator of the HSS is $2(ad – bc)$, so if ($a + d$) is near a maximum and ($b + c$) is near a minimum (as $a + b + c + d$ is a constant), the numerator is largest when $a = d$. Thus, the tendency for HSS to be a maximum where PSS and CSS are equivalent should be expected. A similar argument can be used for the GSS, using the minimum ($b + c$) assumption when $a = d$.

Various outbreak ranking scores can be tested to determine “optimal” values of areal coverage (based on the best accuracy and/or skill) for each major severe weather outbreak threshold (as in Fig. 4.10). If the severe weather parameter selected is SCP using the KDE method, the skill scores (other than PSS) tend to peak at smaller values with increased outbreak ranking score thresholds (cf. Fig. 4.9a and Fig. 4.10). The maximum PSS increases with increased outbreak ranking score threshold because of the exceedingly rare occurrences of highly-scored outbreaks, as previously discussed. As expected, the maximum in skill scores occurs with increased areal coverage as the outbreak ranking score thresholds are increased.

d. Comparing areal coverage to SPC categorical outlooks

Categorical outlooks of moderate- or high-risk days by the SPC from the period 23 January 2003 to 31 December 2008 are associated with days featuring large regions of severe weather parameters favorable for significant severe weather during this same

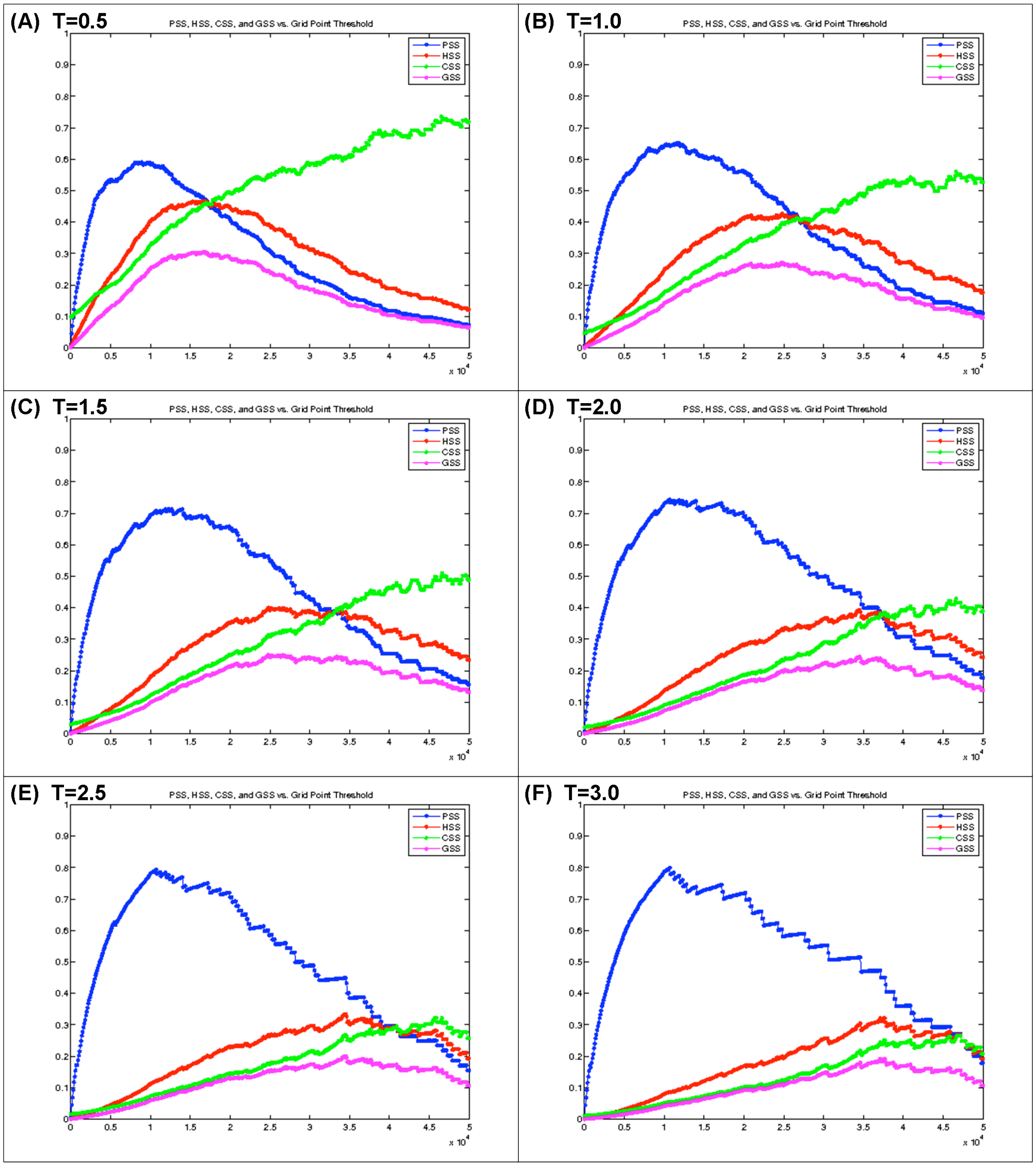


Fig. 4.10. As in Fig. 4.9a, for outbreak ranking score thresholds of (a) 0.5, (b) 1.0, (c) 1.5, (d), 2.0, (e) 2.5, and (f) 3.0.

period (Fig. 4.11). The thresholds indicated on each of the variables and areal coverage techniques shown in Fig. 4.11 are associated with the approximate values in which the SPC categorical outlooks and areal coverage diagnoses are most often the same (Fig. 4.12). (In this analysis, moderate- and high-risk days are assumed to be forecasting

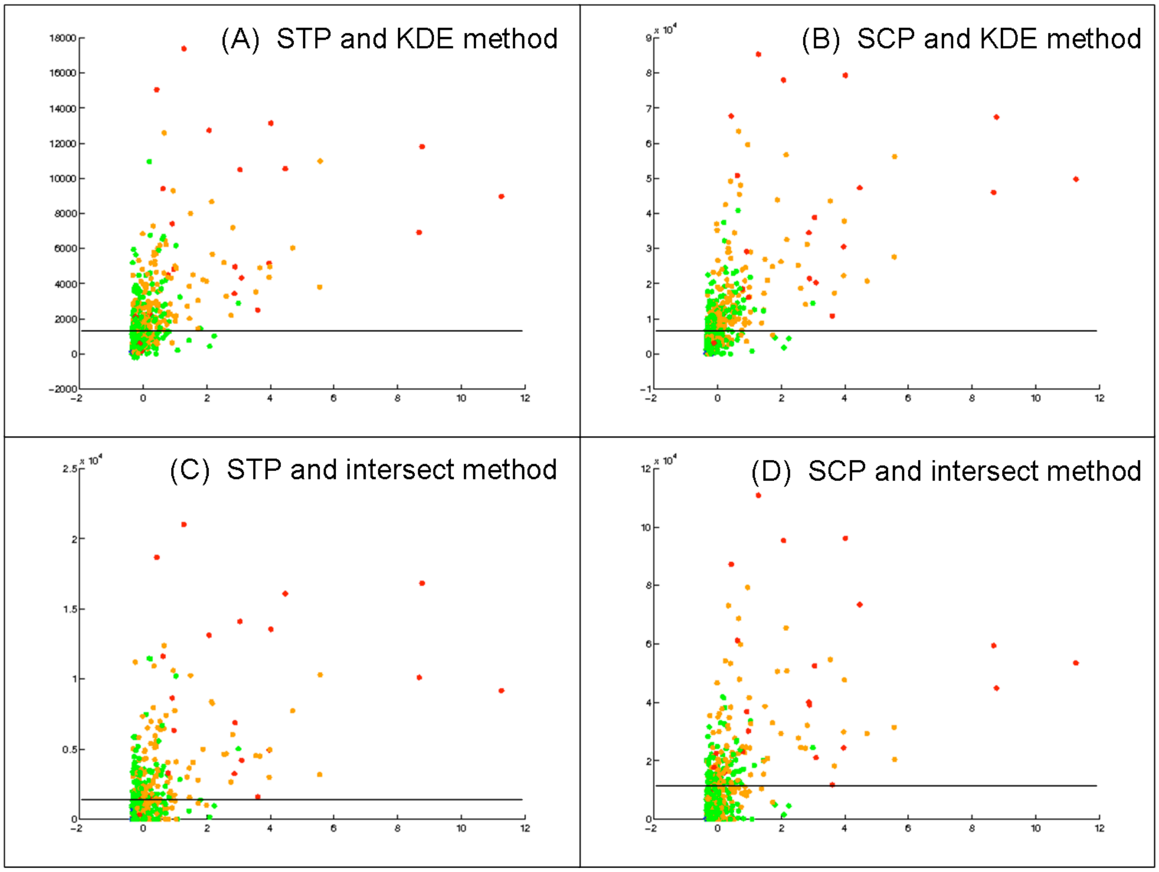
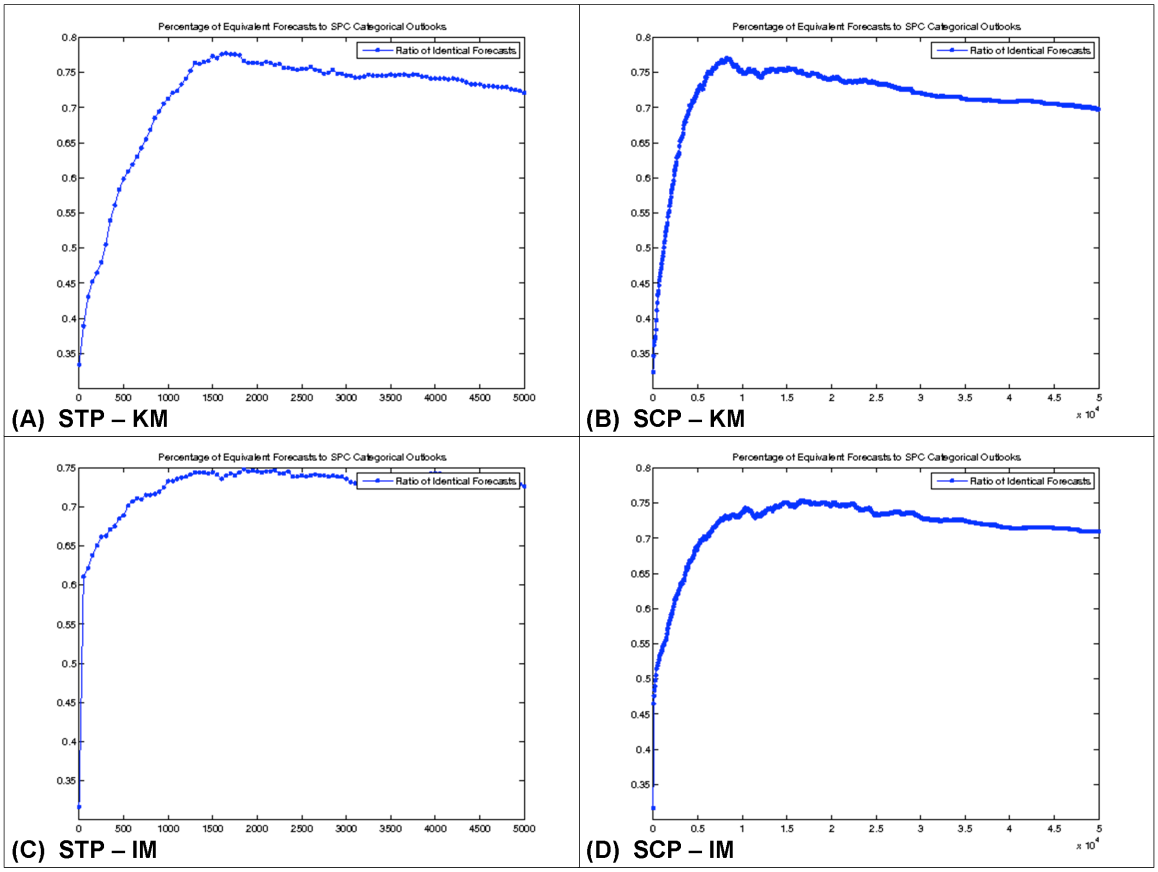


Fig. 4.11. Plot of SPC categorical outlooks (high risk – red, moderate risk – orange, slight risk – green, less than slight risk – blue) compared to areal coverage values (*y*-axis, with variables and methods indicated) for the period 23 January 2003 – 31 December 2008. Outbreak ranking scores are indicated on the *x*-axis. The black horizontal lines on each panel represent the approximate value in which the SPC categorical outlooks match the areal coverage diagnoses the most (see Fig. 4.12).

major severe weather outbreaks. This should be assumed for the rest of this section, unless otherwise stated.) In general, for the values of areal coverage thresholds that agree most with SPC categorical outlooks, the diagnoses are the same ~75% of the time.

Although the primary focus in the rest of this section will be on SPC categorical outlooks issued at 1630 UTC (which is after the beginning of the 24-h period associated with each outbreak case), outlooks issued at 0600, 1300, and 2000 UTC also were analyzed. Contingency statistics for the four outlook times (Figs. 4.13 and 4.14) show



Percentage of Equivalent Forecasts to SPC Categorical Outlooks

(A) STP – KM

Percentage of Equivalent Forecasts to SPC Categorical Outlooks

(B) SCP – KM

Percentage of Equivalent Forecasts to SPC Categorical Outlooks

(C) STP – IM

Percentage of Equivalent Forecasts to SPC Categorical Outlooks

(D) SCP – IM

Fig. 4.12. The ratio of SPC categorical outlooks that agree with areal coverage diagnoses (*y*-axis, assuming moderate- and high-risk outlooks forecast major severe weather outbreaks), as a function of areal coverage (*x*-axis) for the variables and areal coverage techniques indicated.

subtle improvement over the course of the Day 1 outlooks. Most of the improvement appears to occur between the 0600 and 1300 UTC outlooks, with CSI increasing with bias values near unity. Slight improvement is noted in later outlooks, particularly at relatively high thresholds – possibly as a result of increased confidence in significant severe weather and upgraded outlooks.

If the categorical risk were modified such that a slight or higher risk was designated as a major severe weather outbreak, or if only a high risk was indicative of a major severe weather outbreak, the Roebber diagrams change substantially (Fig. 4.15a). For slight risk thresholds, the POD for any threshold is near or equal to 1, as most of the

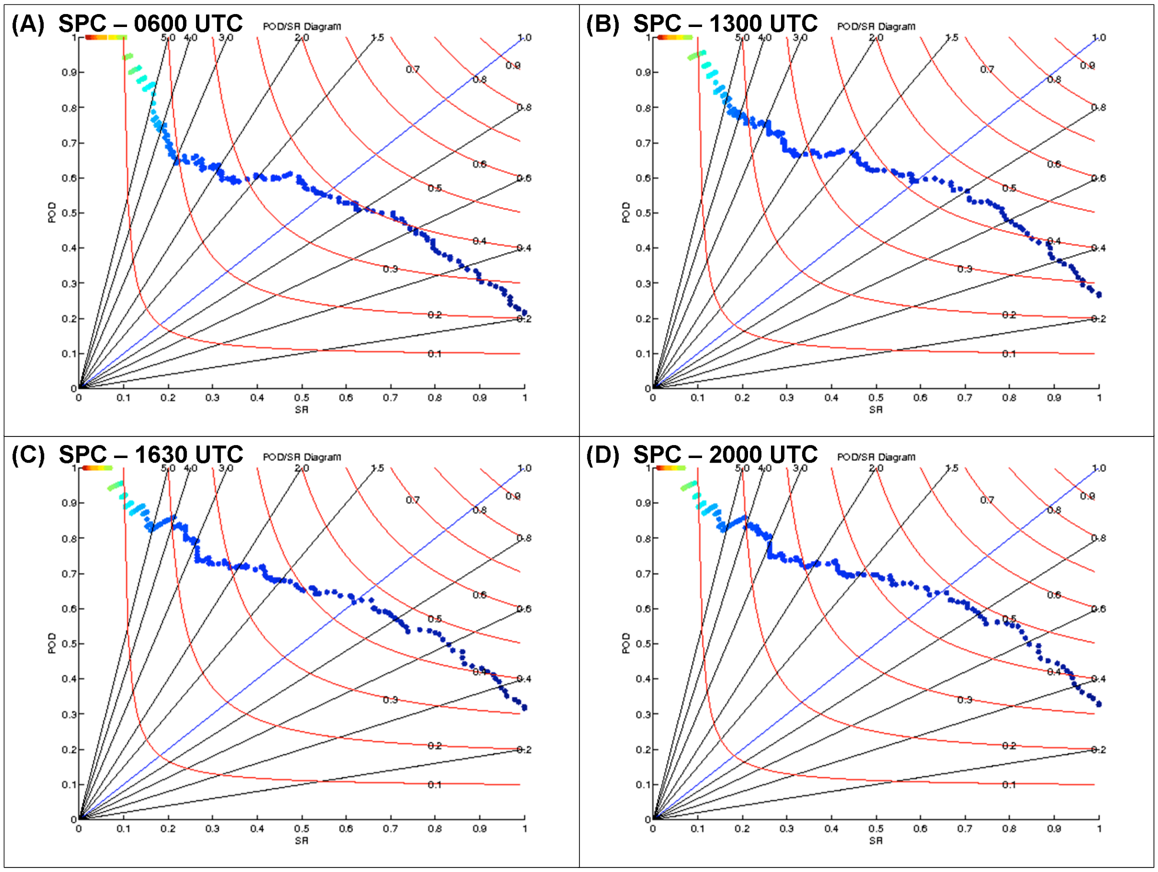
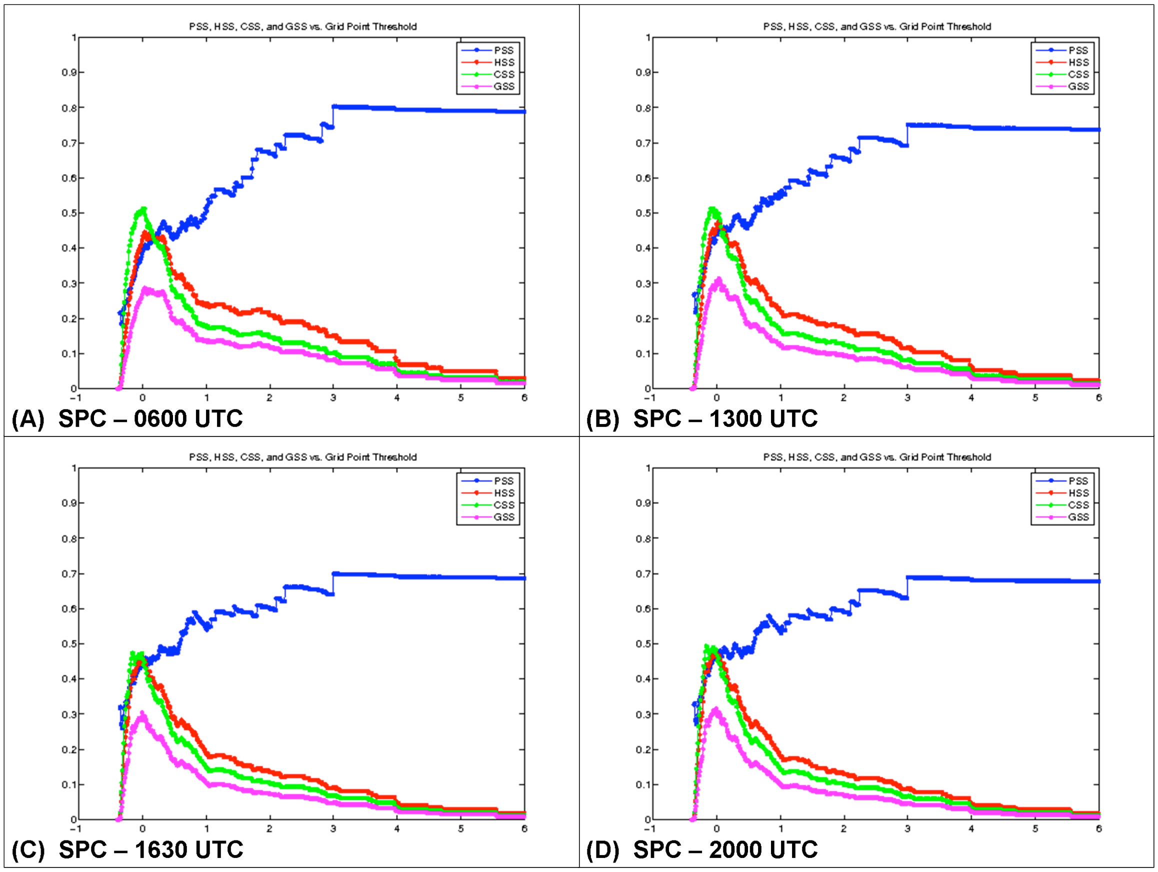


Fig. 4.13. Roebber diagrams of SPC categorical outlooks issued at (a) 0600 UTC, (b) 1300 UTC, (c) 1630 UTC, and (d) 2000 UTC on the nominal date of the outbreaks. Each dot on the diagrams represents the outbreak ranking index score threshold used to classify events as major outbreaks (in increments of 0.01), with the darkest blue color associated with a score of -0.4 and the darkest red color indicating a score of 6.

727 cases considered were designated as slight- or higher-risk days (refer to Fig. 4.11). However, the success ratio (SR) decreases rapidly with increased outbreak ranking index threshold, as the number of false alarms becomes very large. On the other hand, high risk thresholds resulted in lower PODs for equivalent SRs (compared to moderate risks), with CSIs reduced for biases near unity. The noisy appearance of the high-risk statistics suggests the rare issuance of high-risk outlooks, and the susceptibility of the statistics to volatility when the outbreak ranking index threshold increases (eliminating a case with a high-risk outlook from the major severe weather outbreak classification).



(A) SPC – 0600 UTC

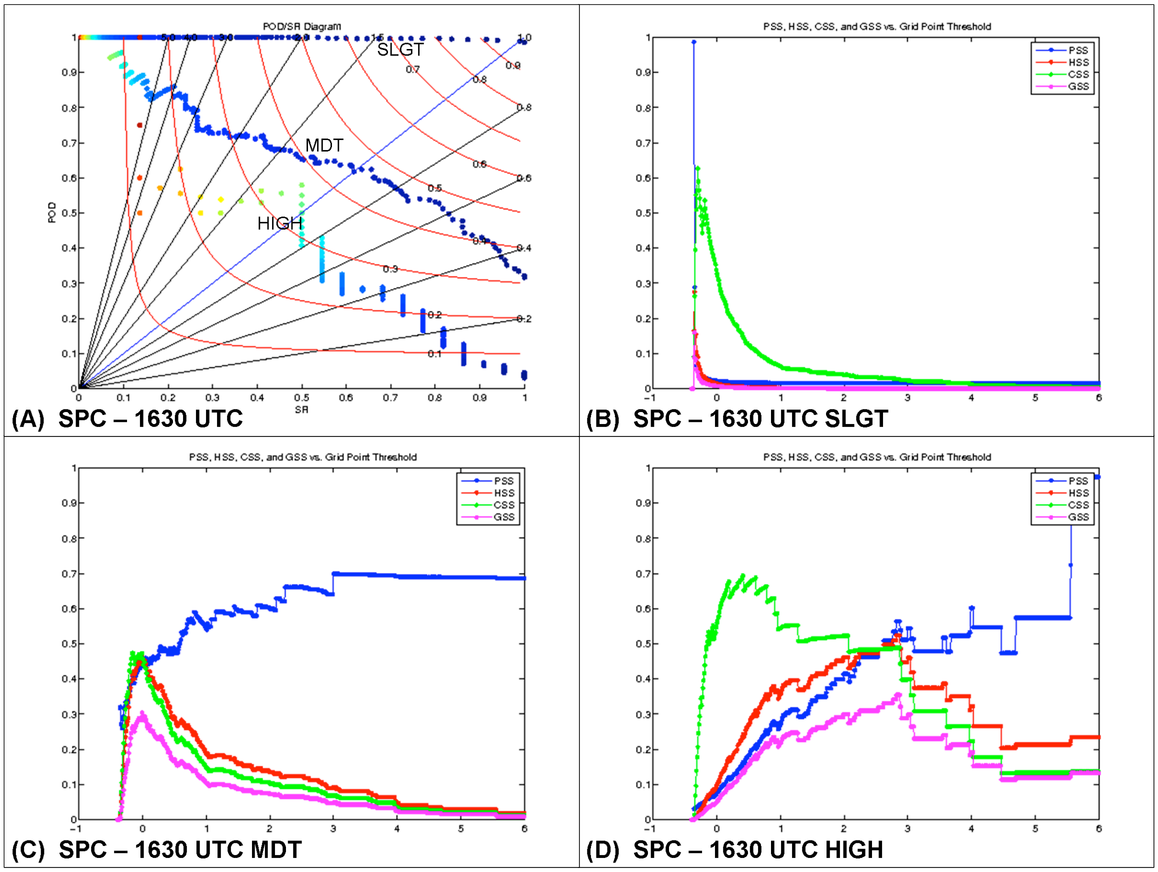
(B) SPC – 1300 UTC

(C) SPC – 1630 UTC

(D) SPC – 2000 UTC

Fig. 4.14. As in Fig. 4.13, except showing skill scores (labeled) as a function of outbreak ranking index threshold (*x*-axis).

The skill scores when using different categorical outlooks as thresholds for forecasting major severe weather outbreaks noticeably change. For the slight-risk threshold, the skill scores are at their maximum at or near the lowest outbreak ranking index threshold. This agrees with subjective notions of these events, in which the lowest-ranked cases still feature (small) clusters of severe weather reports (Section 4e). The designation of these days as major severe weather outbreaks, however, may not be appropriate. The moderate-risk threshold results in skill score maxima near the index threshold of 0, and the high-risk threshold features skill score maxima just below the value of 3. These findings are desirable, as higher-risk categorical outlooks should perform better using higher outbreak classification thresholds.



(A) SPC – 1630 UTC

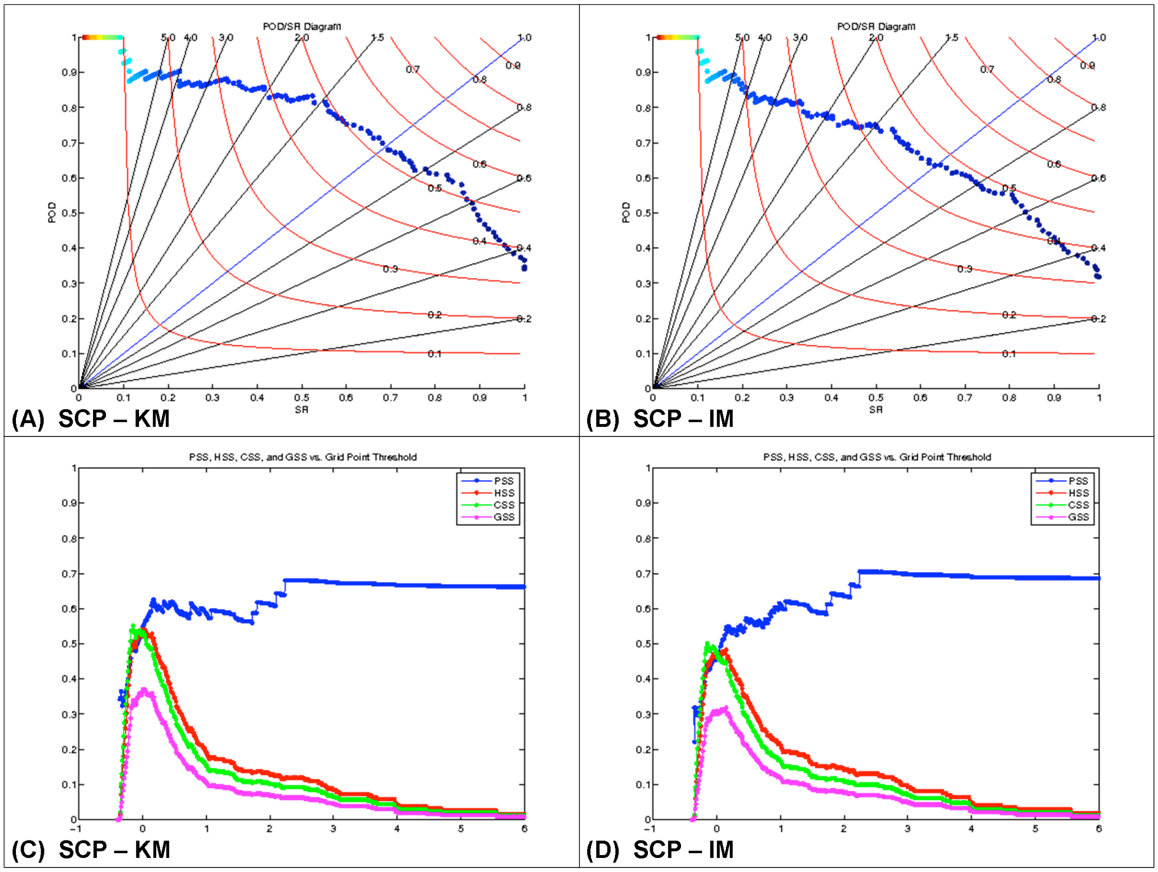
(B) SPC – 1630 UTC SLGT

(C) SPC – 1630 UTC MDT

(D) SPC – 1630 UTC HIGH

Fig. 4.15. (a) Roebber diagram of SPC categorical outlooks issued at 1630 UTC on the nominal date of the outbreaks, using the slight-, moderate-, and high-risk thresholds as forecasts of major severe weather outbreaks. (b) Skill scores as a function of outbreak ranking index threshold for the SPC categorical outlooks issued at 1630 UTC on the nominal date of the outbreaks, using the slight-risk category as the threshold for forecast major severe weather outbreaks. (c) As in (b), using the moderate-risk category as the threshold for forecast major severe weather outbreaks. (d) As in (b), using the high-risk category as the threshold for forecast major severe weather outbreaks.

As with the areal coverage thresholds when analyzing the 1979-2008 period, the SPC categorical outlooks of moderate- and high-risk days tend to perform best around outbreak ranking index thresholds of around 0. Based on the percentage of areal coverage diagnoses and SPC categorical outlooks that agree (Fig. 4.12), one would expect the contingency statistics of the areal coverage diagnoses to be quite similar to those of the SPC categorical outlooks. This is exactly what is observed when looking at the areal coverage diagnoses for the same 727 cases (cf. Figs. 4.13, 4.14, and 4.16). As



(A) SCP – KM

(B) SCP – IM

(C) SCP – KM

(D) SCP – IM

Fig. 4.16. (a) Roebber diagram of the areal coverage diagnoses for the 23 January 2003 – 31 December 2008 period using SCP as the severe weather parameter and the KDE method as the areal coverage technique. A threshold of 7500 was used to make the diagnoses. (b) As in (a), using the intersect method and an areal coverage threshold of 10 000 (as shown in Fig. 4.10c). (c) Skill scores as a function of outbreak ranking index threshold (*x*-axis), using the variable and areal coverage method as indicated in (a). (d) As in (c), using the variable and areal coverage method as indicated in (b).

observed when analyzing the 1979-2008 dataset, the intersect method is slightly worse than the KDE method for diagnosing major severe weather outbreaks. With HSSs near 0.5 at outbreak ranking index thresholds of around 0 using the areal coverage method, these results are quite comparable to the 1630 and 2000 UTC SPC outlooks.

A final comparison should focus on finding areal coverage diagnoses that are comparable to a high-risk threshold for forecasts of major severe weather outbreaks. Using SCP as the severe weather parameter and either the KDE or intersect methods (as

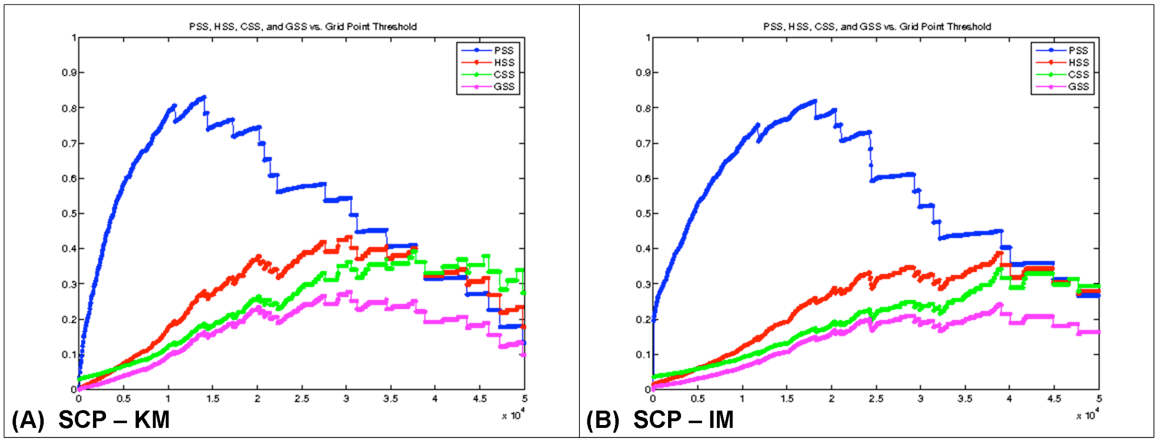
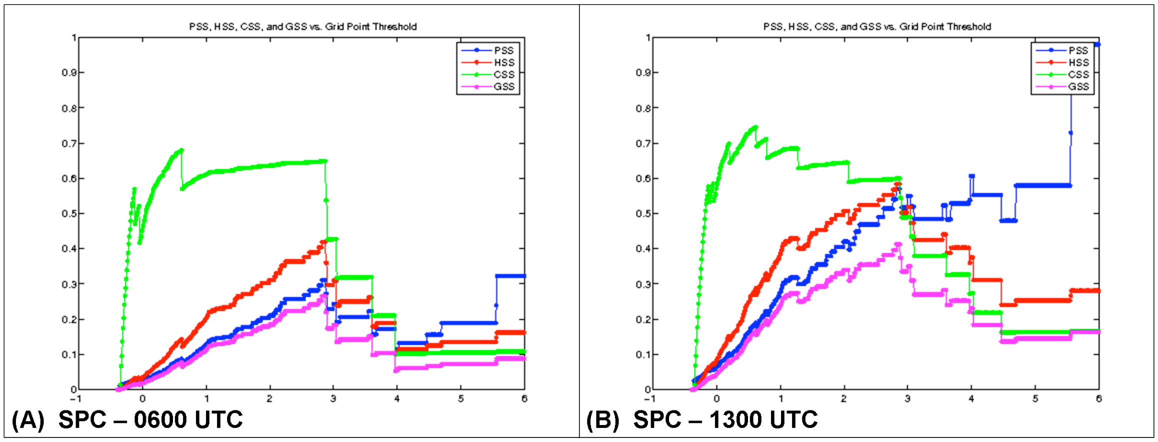


Fig. 4.17. Skill scores as a function of areal coverage threshold, using an outbreak ranking index threshold of 2.75 to differentiate major severe weather outbreaks from less significant events (as suggested in Fig. 4.15d), using (a) SCP and the KDE method and (b) SCP and the intersect method.

in Fig. 4.17), the maximum skill scores tend to be lower compared to the SPC categorical outlooks (cf. Fig. 4.15d). This is true in general when using a single variable for areal coverage diagnoses. The improved performance of SPC categorical outlooks of high risk suggest several possibilities: (1) The areal coverage technique does not take into account convective initiation, and inclusion of convective inhibition is problematic because of reanalysis datasets that only coarsely resolve the boundary layer. (2) Areal coverage diagnoses are susceptible to false alarms in cases for which a linear convective mode occurs (see S10a). (3) SPC forecasters rely on more than one variable to predict the occurrence of the most significant severe weather outbreaks. Inclusion of multiple variables in areal coverage diagnoses may improve such diagnoses. However, as Section 4e will show, many of the variables available are highly correlated, indicating little additional information may be provided when using multivariate areal coverage.[16] (4) Many of the severe weather outbreaks already begin to occur by 1630 UTC, suggesting that human forecasters are capable of distinguishing these events more easily. This last

[16] The use of multiple variables in other techniques, however, may be more beneficial.

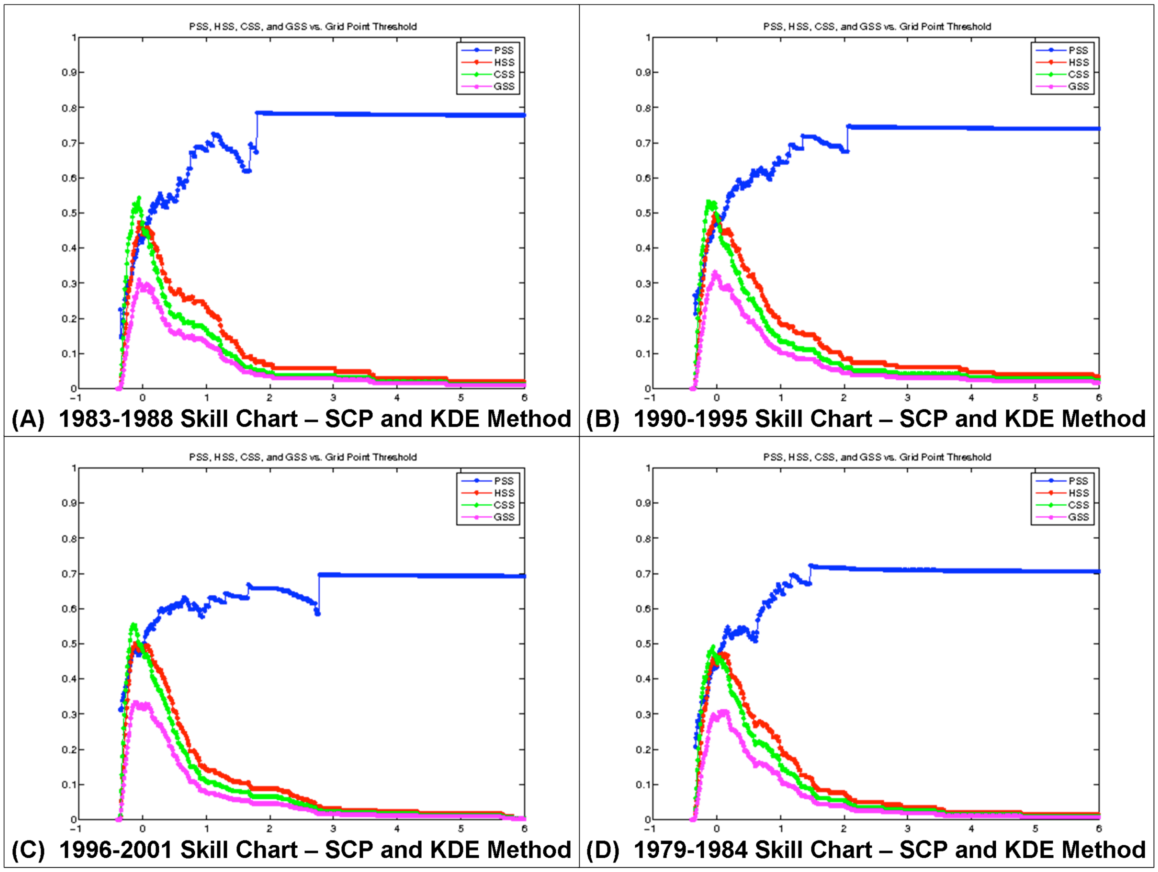


(A) SPC – 0600 UTC **(B) SPC – 1300 UTC**

Fig. 4.18. As in Fig. 4.17, using the SPC categorical outlooks (high-risk threshold) issued at (a) 0600 UTC and (b) 1300 UTC.

point is illustrated by the demonstrably lower skill scores for high-risk thresholds in outlooks issued prior to 1200 UTC (Fig. 4.18). (5) Pattern recognition (Johns and Doswell 1992) and forecaster experience may also play a (potentially substantial) role in the markedly improved performance.

One important observation when analyzing the 2003-2008 cases is that the contingency statistics for the highest-ranked cases do not agree well with the scores when analyzing the 1979-2008 cases (cf. Figs. 4.10e,f and 4.17). This is highly suggestive of sample size sensitivities, which are unsurprising given the exceedingly rare occurrence of major tornado outbreaks. As a result, making conclusions from a sample of cases as small as the outbreak events available from the 2003-2008 period is questionable (Doswell 2007b) and susceptible to error when expanding the dataset. Consider four randomly selected six-year periods in the 1979-2008 suite of cases. A plot of skill scores as a function of the outbreak ranking index score thresholds (Fig. 4.19) shows very little variation in the statistics for relatively low thresholds, with more noticeable differences at higher (>0.5) thresholds. As the outbreak ranking index score increases, the number of major severe weather outbreaks is decreased, reducing the sample of these cases.



(A) 1983-1988 Skill Chart – SCP and KDE Method **(B) 1990-1995 Skill Chart – SCP and KDE Method**

(C) 1996-2001 Skill Chart – SCP and KDE Method **(D) 1979-1984 Skill Chart – SCP and KDE Method**

Fig. 4.19. Plot of skill scores as a function of outbreak ranking index score threshold, using SCP and the KDE method, for cases from (a) 1983-1988, (b) 1990-1995, (c) 1996-2001, and (d) 1979-1984.

Expanding the case list to 6072 events (1960-2008) only resulted in approximately 500 cases with outbreak ranking index scores above the value of 1 and fewer than 200 above the value of 2. This represents ~8.33% and ~3.33% of the events, respectively. As posited in Section 2, such a small sample of cases of major severe weather outbreaks makes their distinction from the less significant events very challenging, particularly given our limited understanding of these events (Section 5).

e. Implications of the findings

Besides the somewhat arbitrary nature of many severe weather parameters (indices) used in operational forecasting today, e.g. in terms of how variables are

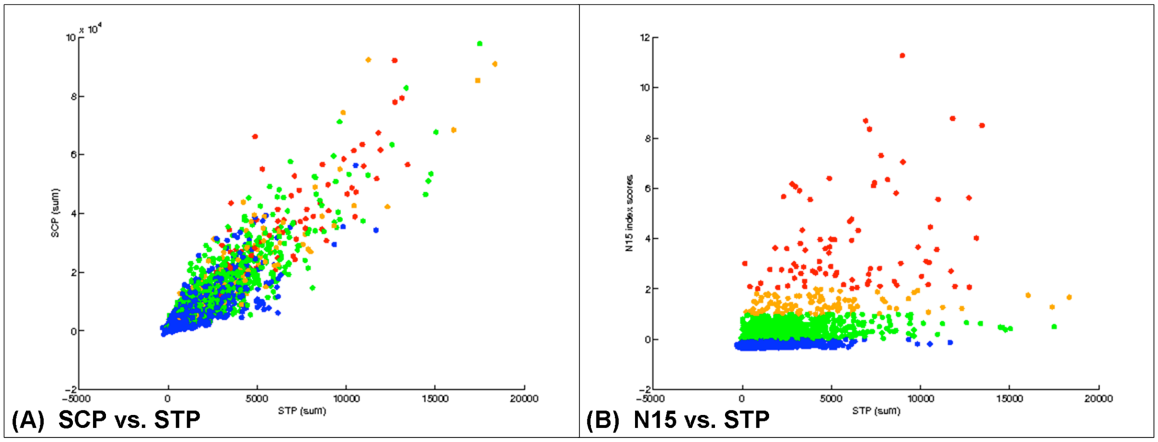


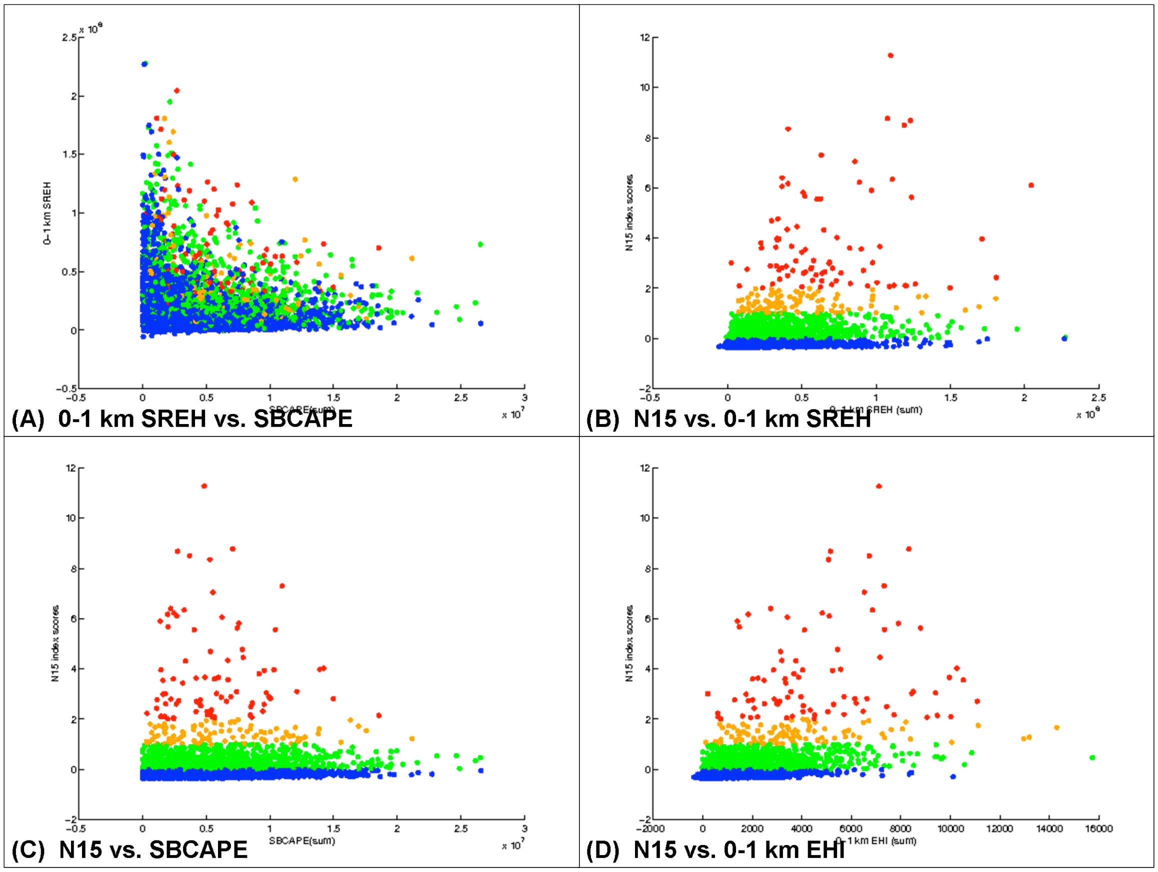
Fig. 4.20. (a) Scatter plot of the sum of STP using the KDE method (*x*-axis) and the sum of SCP using the KDE method (*y*-axis) for each of the 4057 cases from 1979-2008. Red dots indicate cases with N15 scores equal to or exceeding 2, orange dots indicate scores between 1 (inclusive) and 2 (not inclusive), green dots indicate scores between 0 (inclusive) and 1 (not inclusive), and blue dots indicate scores below 0. (b) As in (a), except the N15 index scores are indicated on the *y*-axis.

combined (see Doswell and Schultz 2006), one drawback to these variables is becoming more obvious with each severe weather discrimination study. Specifically, most variables are very highly correlated to each other and not nearly as correlated with specific severe weather events (e.g., major severe weather outbreaks). Using the areal coverage of the sum of SCP and the sum of STP with the KDE method, these two variables have a correlation of 0.9087 for the 4057 cases from 1979-2008, whereas the sum of STP with the KDE method has a correlation of 0.5438 with the N15 index (used to rank the outbreaks). This is illustrated by the one-dimensional nature of the scatter when analyzing a two-dimensional scatter plot of the SCP and STP variables (Fig. 4.20a) compared to the two-dimensional scatter plot of STP and the N15 index (Fig. 4.20b). Unfortunately, adding variables does not separate the highly ranked cases from the less significant events using the areal coverage method, even with increased dimensionality, because of these strong correlations.

This suggests the need for rigorous, systematic testing of newly proposed variables that elucidate *additional* information regarding the distinction of severe weather events. If it is shown that a newly proposed variable discriminates events nearly as well as already demonstrated variables, the new variable would serve little additional purpose if the variable is highly correlated to the variables already in use. However, if the variable is not well correlated with these other variables, then their combination may provide substantial additional information regarding event discrimination. This is demonstrated using the variables 0-1 km SREH and SBCAPE (Fig. 4.21a) compared to using each variable alone (Figs. 4.21b,c). The correlation of the KDE method sum of SREH with the N15 index is 0.3849, the KDE method sum of SBCAPE with the N15 index is 0.1288, and the correlation of the two variables is -0.1796. As Fig. 4.21a suggests, the combination of the two variables may provide additional means of discrimination of severe weather outbreaks. The variable 0-1 km EHI, which is simply a product of the two variables (using SBCAPE versus other forms of CAPE), has a correlation with the N15 index of 0.5034 (Fig. 4.21d). This improvement demonstrates added utility of EHI over each of the individual variables, which was possible because of poorly correlated variables both demonstrating weak to moderate correlation with the severity of the outbreaks.[17]

On the other hand, demonstrating that SCP or STP is preferred over 0-1 km EHI is more questionable. For example, the correlation of STP and 0-1 km EHI (KDE method) is 0.9323, and the STP and N15 index have a correlation of 0.5438 – not a substantial improvement over 0-1 km EHI. Additionally, because these indices are

[17] Later in this section, results from combinations of variables will be shown to illustrate the limited improvement when using multiple severe weather variables.



(A) 0-1 km SREH vs. SBCAPE

(B) N15 vs. 0-1 km SREH

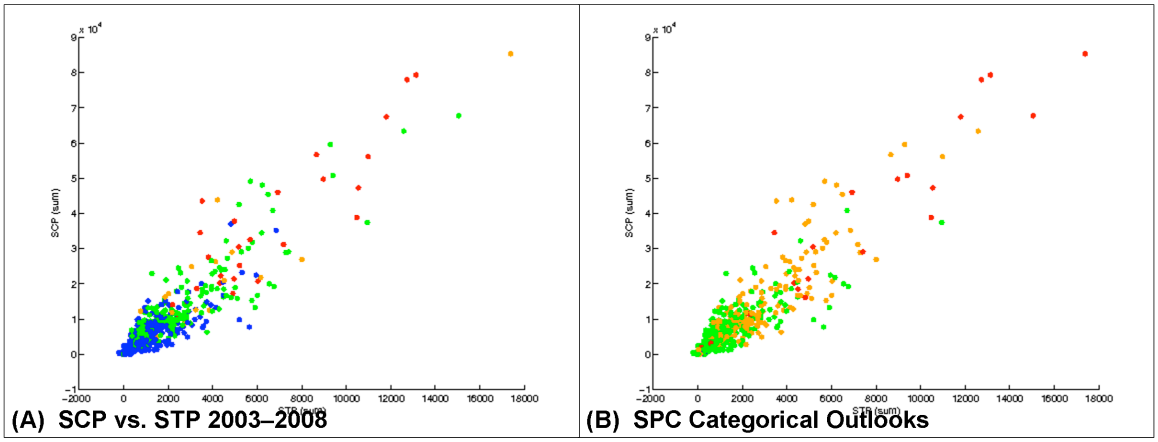
(C) N15 vs. SBCAPE

(D) N15 vs. 0-1 km EHI

Fig. 4.21. (a) As in Fig. 4.20a, with SBCAPE on the *x*-axis and 0-1 km SREH on the *y*-axis. (b) As in Fig. 4.20b, with 0-1 km SREH on the *x*-axis. (c) As in (b), with SBCAPE on the *x*-axis. (d) As in (b), with 0-1 km EHI on the *x*-axis.

combinations of severe weather parameters, relatively high correlations exist between them. For example, STP and 0-1 km SREH (KDE method) have a correlation of 0.5527. Taking the (arbitrary) product of the two variables results in a correlation with the N15 index of 0.5143, which is no better than using the STP itself.

The two-dimensional multivariate scatter plots and the SPC categorical outlooks (1630 UTC) look quite similar, in terms of the identification of the most significant severe weather outbreaks (e.g., Fig. 4.22). High-risk outlooks tend to be issued when areal coverage of STP and SCP are high, with lower risks as STP and SCP both approach small values. Clearly, high-risk outlooks are not *always* issued when areal coverage of



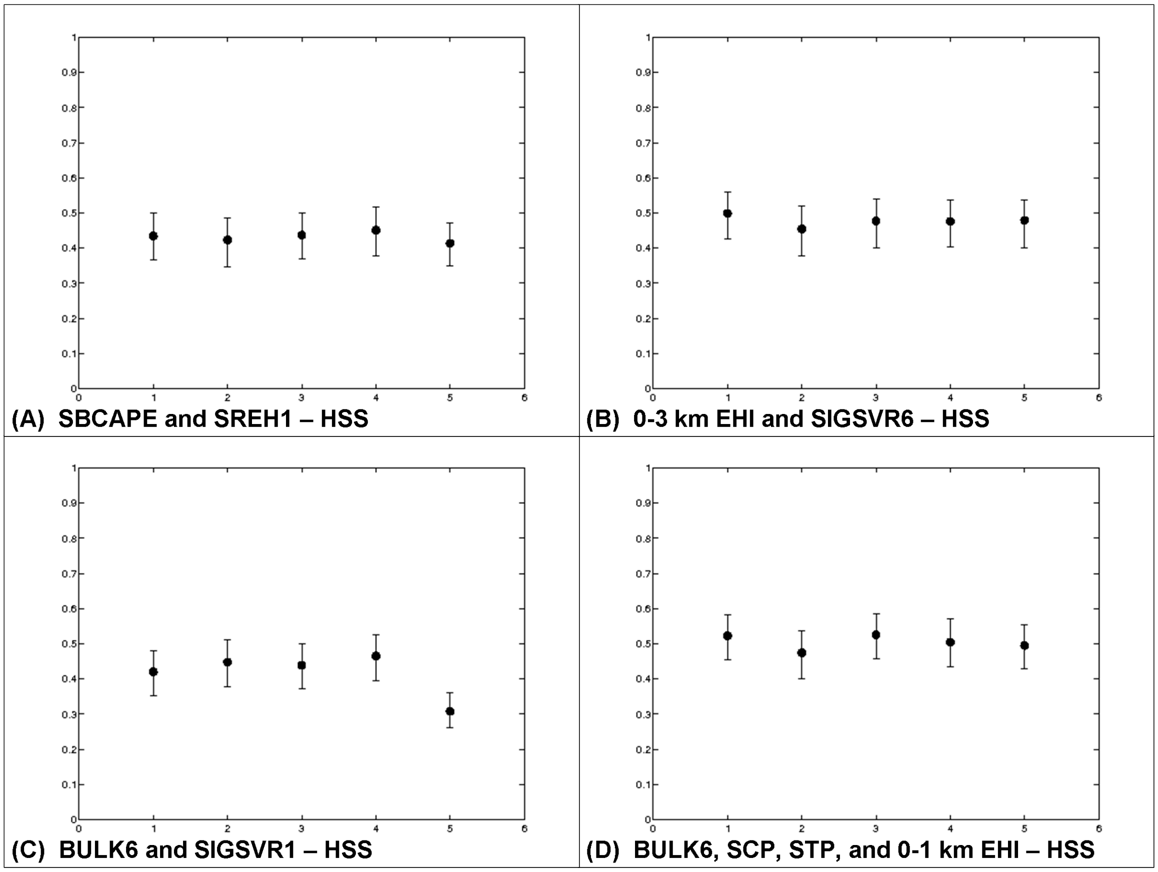
(A) SCP vs. STP 2003–2008

(B) SPC Categorical Outlooks

Fig. 4.22. (a) As in Fig. 4.20a, for the 23 January 2003 – 31 December 2008 period. (b) As in (a), with SPC categorical outlooks indicated (red – high risk; orange – moderate risk; green – slight risk; blue – less than slight risk). *Note that the colors in (a) should not be directly compared to the colors in (b).*

STP and SCP are high, as other factors are involved in the occurrence or absence of numerous tornadic supercells on a given day (e.g., convective mode; questions regarding convective initiation; uncertainties with location) that are not accounted for entirely by the areal coverage technique.

Testing combinations of variables was conducted using linear and quadratic discriminant analysis and decision tress (algorithms 1-5 in Table 2.1). Cases from 1979-2002 were used for training the statistical models, and they were tested on the cases used for comparing SPC categorical outlooks with areal coverage diagnoses (i.e., the 727 cases from 23 January 2003 – 31 December 2008, as discussed in Section 4d). Bootstrapping was performed to develop 95% confidence intervals for the results. Various combinations of severe weather parameters were tested (e.g., Fig. 4.23) using multiple thresholds of outbreak ranking index for event classification (Fig. 4.24). As expected, little improvement was observed when adding variables compared to single variable statistics. Because the number of major severe weather outbreaks decreases as the outbreak ranking index threshold is increased, the confidence intervals become larger.



(A) SBCAPE and SREH1 – HSS

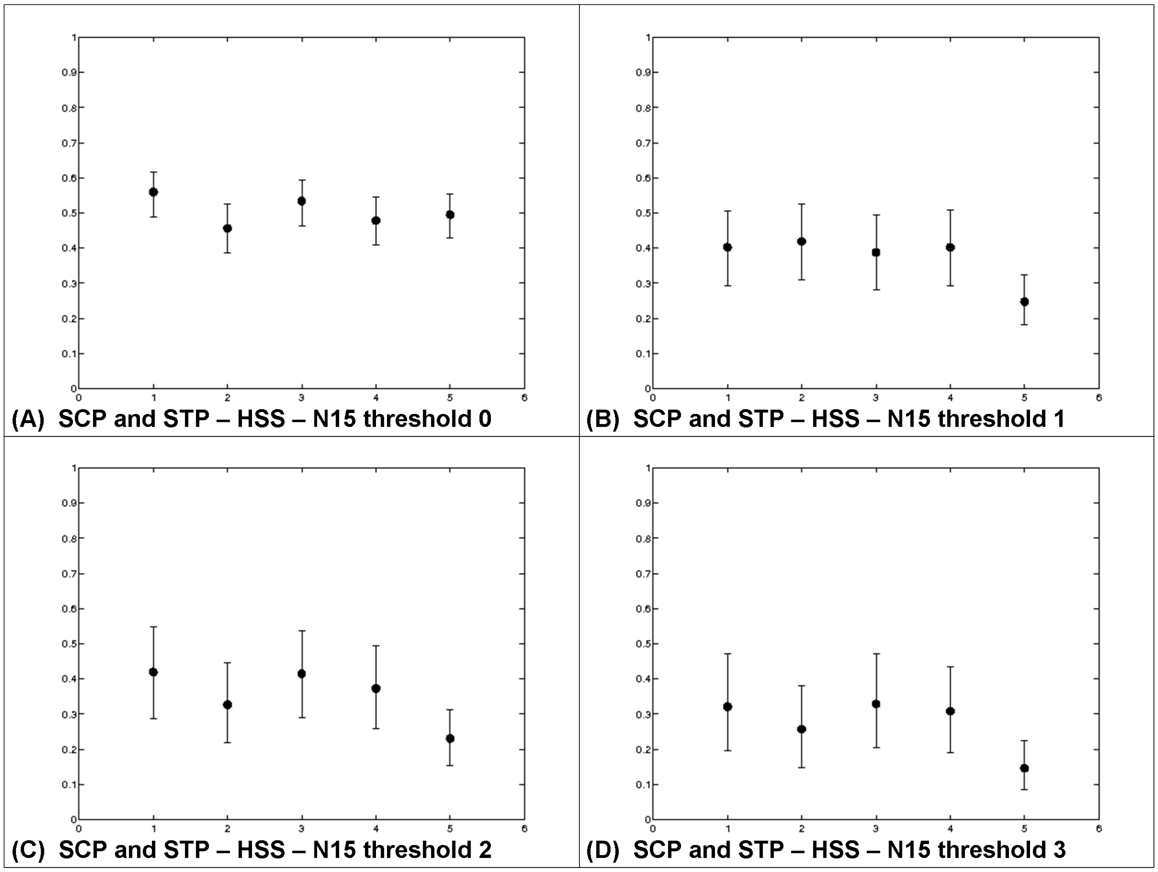
(B) 0-3 km EHI and SIGSVR6 – HSS

(C) BULK6 and SIGSVR1 – HSS

(D) BULK6, SCP, STP, and 0-1 km EHI – HSS

Fig. 4.23. Bootstrapped 95% confidence intervals of the Heidke skill scores using areal coverage of multiple variables (KDE method) on the testing data (2003-2008). The black dots are the medians of the confidence intervals. Statistical algorithms identified on the *x*-axis and labeled as in Table 2.1. In (a), SBCAPE and 0-1 km SREH (SREH1) are used. In (b), 0-3 km EHI and the product of SBCAPE and 0-6 km bulk shear (SIGSVR6) are used. In (c), 0-6 km bulk shear (BULK6) and the product of SBCAPE and 0-1 km bulk shear (SIGSVR1) are used. In (d), BULK6, SCP, STP, and 0-1 km EHI are used. The N15 index threshold of 0 is used to differentiate major severe weather outbreaks from null events.

Inspecting the cases near the thresholds resulting in most skillful discrimination, using either the areal coverage method or the SPC categorical outlooks, can provide insight as to what cases to expect when diagnosing or predicting a major severe weather outbreak in the future. Recall that cases are considered for ranking using the KDE technique (Section 3) based on two criteria: (1) The number of reports for a given event must exceed the detrended event-averaged number of reports for a given year. (2) The



(A) SCP and STP – HSS – N15 threshold 0

(B) SCP and STP – HSS – N15 threshold 1

(C) SCP and STP – HSS – N15 threshold 2

(D) SCP and STP – HSS – N15 threshold 3

Fig. 4.24. As in Fig. 4.23, using SCP and STP areal coverage (KDE method), with the N15 index threshold of (a) 0, (b) 1, (c) 2, and (d) 3 to differentiate major severe weather outbreaks from null events.

density of the reports within an event region must exceed the detrended event-averaged density of reports for a given year. Because of these two criteria, events with few or no reports or widely dispersed severe reports were automatically excluded from consideration. Analyzing the SPC slight-risk outlooks, the optimal thresholds for which prediction of major severe weather outbreaks occur are very near the bottom-ranked cases. It is at this point in which some of the days feature categorical outlooks without a slight or higher risk for severe weather (e.g., refer to Figs. 4.15a,b). This observation is encouraging, as it seems to provide at least *some* evidence that the two criteria are reasonable for designation of cases as severe weather events. It would be beneficial to

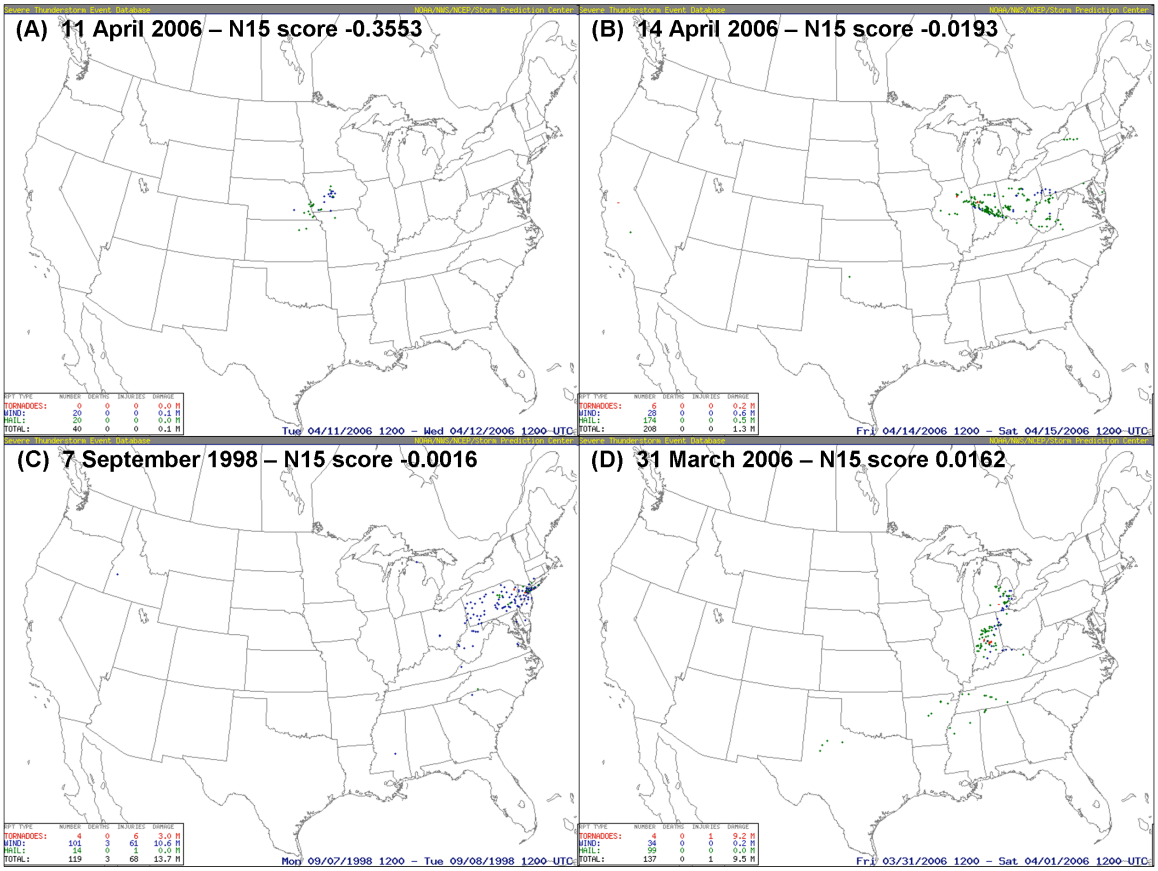
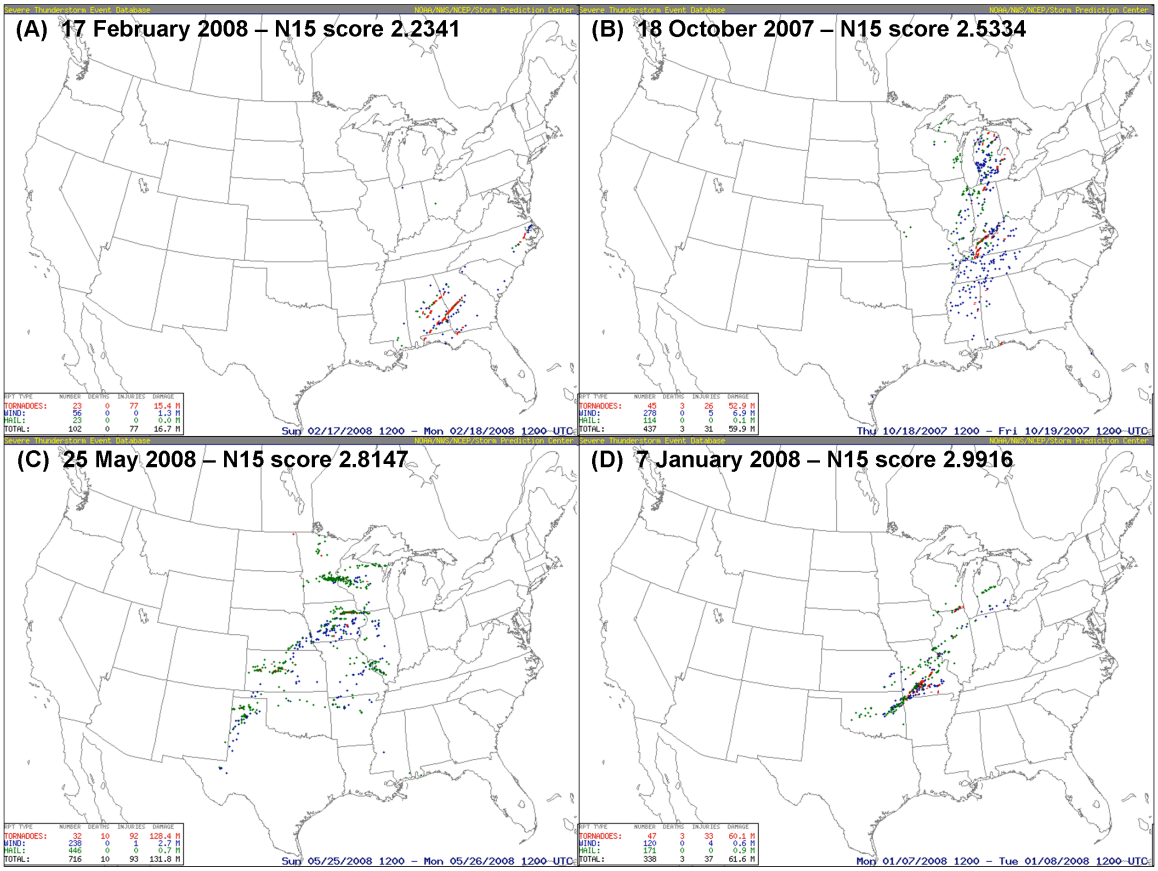

Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(A) 11 April 2006 – N15 score -0.3553
RPT TYPE NUMBER DEATHS INJURIES DAMAGE
TORNADOES: 0 0 0 0.0 M
WIND: 20 0 0 0.1 M
HAIL: 20 0 0 0.0 M
TOTAL: 40 0 0 0.1 M
Tue 04/11/2006 1200 – Wed 04/12/2006 1200 UTC
Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(B) 14 April 2006 – N15 score -0.0193
RPT TYPE NUMBER DEATHS INJURIES DAMAGE
TORNADOES: 6 0 0 0.2 M
WIND: 28 0 0 0.6 M
HAIL: 174 0 0 0.5 M
TOTAL: 208 0 0 1.3 M
Fri 04/14/2006 1200 – Sat 04/15/2006 1200 UTC
Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(C) 7 September 1998 – N15 score -0.0016
RPT TYPE NUMBER DEATHS INJURIES DAMAGE
TORNADOES: 4 0 6 3.0 M
WIND: 101 3 61 10.6 M
HAIL: 14 0 1 0.0 M
TOTAL: 119 3 68 13.7 M
Mon 09/07/1998 1200 – Tue 09/08/1998 1200 UTC
Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(D) 31 March 2006 – N15 score 0.0162
RPT TYPE NUMBER DEATHS INJURIES DAMAGE
TORNADOES: 4 0 1 9.2 M
WIND: 34 0 0 0.2 M
HAIL: 99 0 0 0.0 M
TOTAL: 137 0 1 9.5 M
Fri 03/31/2006 1200 – Sat 04/01/2006 1200 UTC

Fig. 4.25. Severe reports from 1200 UTC on the nominal date to 1200 UTC the following day, with tornadoes in red, hail in green, and wind in blue, for (a) 11 April 2006, (b) 14 April 2006, (c) 7 September 1998, and (d) 31 March 2006.

investigate days that do not meet these criteria, to determine if these criteria are sufficient for event/null categorization in future studies (see Section 5). Cases near the lowest rankings tend to have a small number of nontornadic reports clustered in a very small region (e.g., Fig. 4.25a), which agrees with subjective notions regarding the relative severity of outbreaks (i.e., that these cases are relatively minor compared to days with more events, higher density of reports, more tornadoes, etc.).

Cases with scores around the index value of 0, which is near the best threshold in which areal coverage diagnoses and SPC categorical outlooks of moderate or high risks distinguish major severe weather outbreaks from less significant events, tend to have a larger number of reports with high spatial densities (Figs. 4.25b-d). The reports generally remain nontornadic, though several of these reports can be significant (≥ 65 kt wind gusts; ≥ 5 cm diameter hail). As scores continue above zero (as in Fig. 4.25d), the number of tornadoes and/or their significance begin to increase. For example, the 31 March 2006 event featured only four tornado reports, but two received F2 ratings. In contrast, the 7 September 1998 event featured only one significant tornado, and the 14 April 2006 event featured no significant tornadoes (even though this event featured more total tornadoes than the higher-ranked cases in Fig. 4.25c,d). *In general, cases above the ranking index value of 0 exhibited a higher frequency of* ***multiple significant*** *tornadoes.* Additionally, cases featuring widespread significant wind gusts (derechos), widespread significant hail, or both were common above scores of 0 and generally nonexistent below this threshold. These tendencies should be expected, as cases above the mean score of the 6072 cases (for N15, around the value of 0) should feature standardized values of the report variables that were positive, particularly for the higher-weighted tornado and significant nontornadic variables. This is perhaps why there is preference toward threshold values around the mean (which is near 0). The ranking scheme was developed based on subjective notions that major severe weather outbreaks are comprised of a (relatively) large number of *significant* severe weather events.

High-risk outlooks issued by the SPC have “optimal” thresholds around ranking index scores of 2.75. Cases surrounding this value (Fig. 4.26) are primarily major tornado outbreaks, as are most events with scores above 2. Although there is a tendency for cases ranked around or higher than 2.75 to feature widespread nontornadic reports (cf. Figs. 4.26a,c), this is by no means seen in all cases (see Figs. 4.26b,d). Although cases


Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(A) 17 February 2008 – N15 score 2.2341
Sun 02/17/2008 1200 – Mon 02/18/2008 1200 UTC
Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(B) 18 October 2007 – N15 score 2.5334
Thu 10/18/2007 1200 – Fri 10/19/2007 1200 UTC
Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(C) 25 May 2008 – N15 score 2.8147
Sun 05/25/2008 1200 – Mon 05/26/2008 1200 UTC
Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(D) 7 January 2008 – N15 score 2.9916
Mon 01/07/2008 1200 – Tue 01/08/2008 1200 UTC

RPT TYPE	NUMBER	DEATHS	INJURIES	DAMAGE
TORNADOES:	23	0	77	15.4 M
WIND:	56	0	0	1.3 M
HAIL:	23	0	0	0.0 M
TOTAL:	102	0	77	16.7 M

RPT TYPE	NUMBER	DEATHS	INJURIES	DAMAGE
TORNADOES:	45	3	26	52.9 M
WIND:	278	0	5	6.9 M
HAIL:	114	0	0	0.1 M
TOTAL:	437	3	31	59.9 M

RPT TYPE	NUMBER	DEATHS	INJURIES	DAMAGE
TORNADOES:	32	10	92	128.4 M
WIND:	238	0	1	2.7 M
HAIL:	446	0	0	0.7 M
TOTAL:	716	10	93	131.8 M

RPT TYPE	NUMBER	DEATHS	INJURIES	DAMAGE
TORNADOES:	47	3	33	60.1 M
WIND:	120	0	4	0.6 M
HAIL:	171	0	0	0.9 M
TOTAL:	338	3	37	61.6 M

Fig. 4.26. As in Fig. 4.25, for (a) 17 February 2008, (b) 18 October 2007, (c) 25 May 2008, and (d) 7 January 2008.

ranked above 2.75 are almost always major tornado outbreaks, some cases below this threshold likely would be classified the same way. As the SPC high-risk outlooks have a relatively small POD for a given SR compared to moderate-risk days (see Fig. 4.15a), no obvious reason is apparent for finding this particular ranking index threshold other than the threshold is within the cases ranked as major tornado outbreaks. Moreover, the small sample of high-risk days available from 2003-2008 likely inhibits drawing meaningful conclusions in this manner to a large degree.

The similar ability to discriminate major severe weather outbreaks from less significant events between SPC categorical outlooks of moderate- and high-risk days and the areal coverage diagnoses suggests that areal coverage is a primary means of identifying the most significant severe weather events. However, the identification of major tornado outbreaks within the highest-ranked outbreak cases appears to be more limited, because of a substantial false alarm problem. Large areal coverage can be observed in environments favoring linear convective modes – generally unfavorable for widespread significant tornadoes. Additionally, uncertainties regarding convective inhibition, storm-scale environments that can be vastly different from larger scales, and storm interactions are not accounted for (adequately) using areal coverage. Consequently, observed large areal coverage may be most appropriately regarded as a "necessary condition" for tornado outbreaks, and subsequent thorough analysis of the subsynoptic-scale environment would be necessary for proper identification of major tornado outbreaks.

The somewhat reduced discrimination using the intersect method versus the KDE method indicates that *a priori* knowledge of the location of the outbreak provides beneficial information on the severity of the outbreak. The intersect method requires selection of an initial threshold value for a severe weather parameter, which can be altered, whereas the KDE method does not require a starting threshold. Generally, it was found that increasing the thresholds for the intersect method did not substantially improve outbreak discrimination and typically led to deterioration of the statistics as the thresholds became exceedingly high (not shown). On the other hand, lowering the thresholds commonly led to very large regions of areal coverage, sometimes connecting distinct events, which were generally unrepresentative of the outbreaks (not shown). Importantly, *the intersect method is not guaranteed to represent the best environment for the outbreak* *itself*, as correlations between favorable regions and the affected area of the outbreak are not perfect.

Of course, the KDE method as introduced here is not likely to be implemented in an operational setting, as the location of the outbreak is not known beforehand.[18] Therefore, a technique similar to the intersect method would be necessary to implement in an operational setting. Perhaps considering every contiguous region in which a severe weather parameter exceeds a specified threshold should be incorporated, but this ultimately leads to an excessive number of false alarms, as shown by Hamill et al. (2005).

An important caveat to this work is that areal coverage has only been tested diagnostically. Because of the demonstrated utility of this method, future work should investigate using model simulated fields for computation of areal coverage. Comparisons with SPC categorical outlooks, particularly those issued at times shortly after the meteorological model is initialized, could also be investigated. Future work should continue to incorporate more recent cases (e.g., 2009 data, which have recently become available) in efforts to increase sample size.

As previously mentioned, only cases qualifying as severe weather events using KDE (as in Section 3) were considered when evaluating the areal coverage technique. Inclusion of all days, particularly to investigate how areal coverage may help diagnose slight-risk days, would make for a more complete comparison with operational forecasts. Furthermore, such work could provide valuable information regarding the false alarm problem, as it is likely that a subset of days with little or no severe weather feature

[18] However, a model-derived version of the KDE method, based on characteristics of the simulated fields, could be developed. See Schwartz et al. (2010) and Sobash et al. (2010) for more details.

favorable regions of severe weather parameters. Of course, computational tractability would become a concern in such research.

5. Overall discussion and future work

Discrimination of major tornado outbreaks from all other types of severe weather outbreaks has challenged operational forecasters for decades. Although the synoptic- and subsynoptic-scale environments of tornado outbreaks generally have been identified *a priori*, this comes at a cost of a large number of false alarms. Although a subset of these cases can be recognized as false alarms, based on recognition of environments favoring a particular type of convective mode or of inhibiting factors, the relatively frequent nature of false alarm forecasts suggests this problem continues to plague forecasters.

Limitations to consistent, accurate, and skillful discrimination of major tornado outbreaks from less significant events include the following.

(1) The sample of tornado outbreaks with which to train and test statistical models may not be large enough to identify coherent signals that are consistent with these events and are not observed with null events. It is quite possible many more years of data, perhaps several more decades of data, will be necessary to counteract sample size concerns.

(2) Observations of past events are inundated with nonmeteorological artifacts and are grossly inadequately representative of severe weather outbreaks. From biases associated with population density and political boundaries to secular trends in observing severe weather to modifications in verification emphases in the past five decades, no known method to account for these artifacts can remove them completely. Severe weather will never be observed perfectly, so that creating a ranking scheme to determine the relative severity of these events cannot be compared to “truth”. Furthermore, observations of severe weather can be and often are

erroneous (wind gust estimations, hail size approximations, tornadoes that are rated without hitting substantial structures). *Therefore, any research conducted using the ranking schemes described in SD10a and Section 3 herein are subject to the uncertainties associated with the observations of these events.*

(3) The atmosphere does not provide distinct categories of events. As such, any study that focuses on discriminating types of cases is subject to the uncertainties associated with event classification. In the present work, class types were varied to determine what threshold values of outbreak classification were associated with the most accurate and/or skillful discrimination. However, these thresholds were not necessarily associated with distinct classes of events. Moreover, the ranking schemes developed by D06, SD10a, and Shafer and Doswell 2010b were designed to identify prototypical cases of the relevant types of outbreaks.

(4) The process of tornadogenesis is not well understood. Therefore, using meteorological covariates associated with midlevel mesocylones and tornadoes obviously is limited by our lack of understanding of this process. What is clear is that the severe weather parameters presently used to diagnose types of severe weather are not completely explanatory. *As a result, we have yet to identify those predictors needed for accurate tornado forecasting*. Additionally, uncertainties and discrepancies between the local storm environment and the larger-scale environment can be large and generally are not observed consistently and reliably (see Markowski et al. 1998b; Rasmussen and Blanchard 1998).

(5) In conjunction with (4), the links between synoptic-scale processes and tornado outbreaks are not well understood. Based on the research completed so far, it appears

that synoptic-scale processes play *some* role in the occurrence or absence of tornado outbreaks (refer to S09; M09; S10a,b; among others). For example, synoptic-scale processes appear to be associated with the areal coverage of parameters favorable for significant severe weather (S09; S10a). However, *events with large areal coverage include but are not limited to tornado outbreaks* (S10a; Sections 4c-e). Additional synoptic-scale processes also may be important (including degree of boundary-relative midlevel flow and shear, and implications on convective mode). Nevertheless, several important questions remain unanswered. Can synoptic-scale processes dictate when inhibiting factors (such as presence and strength of capping inversions, convective modification of the environment, etc.) prevent a more significant severe weather outbreak from occurring? Do synoptic-scale processes play a primary role in the occurrence of some tornado outbreaks and a limited role in others? If so, how far in advance can these tornado outbreaks be predicted accurately?

(6) The time of year in which outbreaks occur can affect the type of outbreak that occurs (S10b). Unfortunately, eliminating this influence when investigating synoptic-scale links to tornado outbreaks is challenging. Typically, such efforts will reduce the sample size (necessitating inclusion of less significant events, as in S10b). Time of year has not been accounted for in more recent studies (as in S10a; Sections 2 and 4 herein) and may prove effective in reducing the false alarm problem to some degree, particularly given the lack of major tornado outbreaks observed in the summer months (S10b); however, it is also posited that the environmental parameters used to discriminate outbreak events should not require a time-of-year constraint (i.e., the

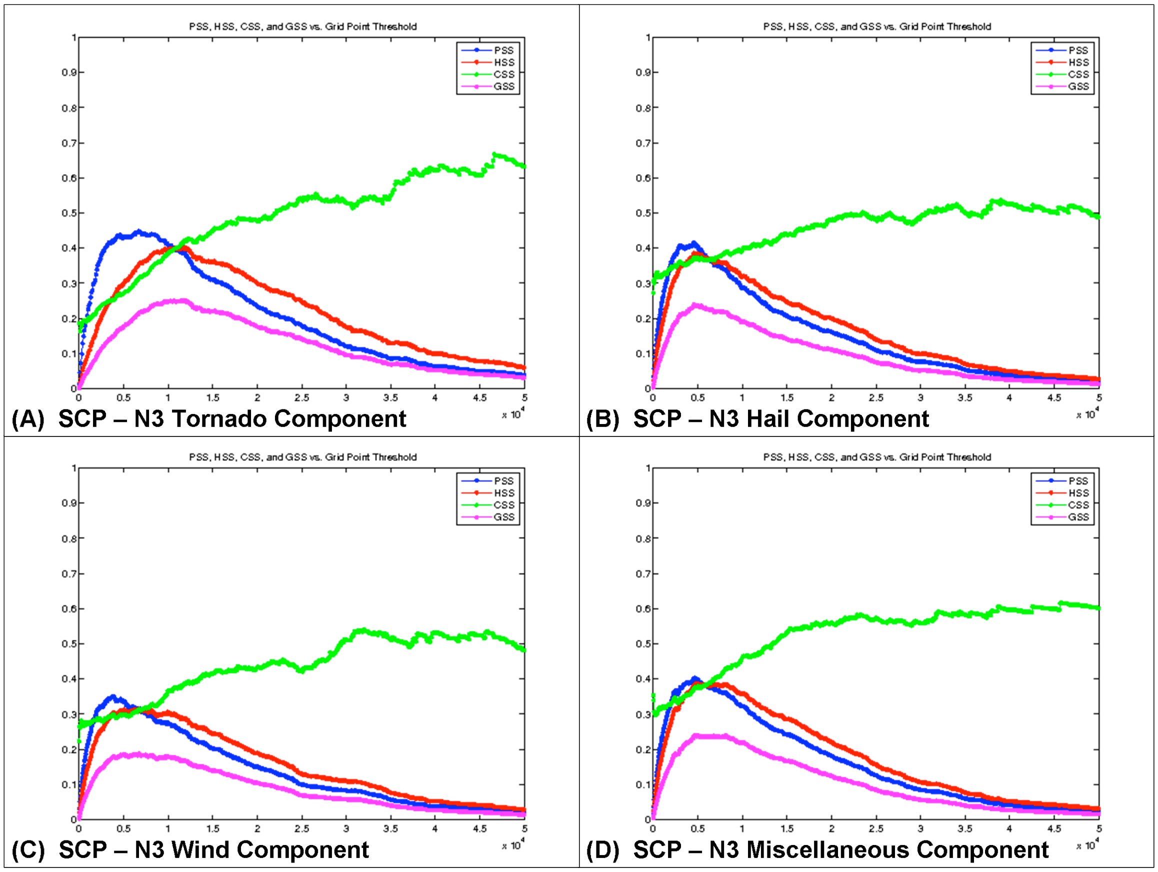
identical environment that produces a major severe weather outbreak in the spring would produce an outbreak in the summer, fall, or winter).

In addition to these limitations, additional research investigating discrimination of severe weather outbreaks is needed, primarily in three areas. First, evaluation of the areal coverage technique in a *forecast* setting is appropriate. Determination of the ability of mesoscale model fields to diagnose major severe weather outbreaks using areal coverage of severe weather parameters can be compared to longer-term operational forecasts (e.g., Day 2+ convective outlooks from the Storm Prediction Center). How far in advance is areal coverage a useful means of predicting the relative severity of outbreaks?

Second, inclusion of cases not considered as severe weather events using the modified ranking scheme described in Section 3 would be a more appropriate comparison to operational forecasters, as *a priori* determination of whether a severe weather outbreak will occur is subject to uncertainty. Such studies face computational tractability concerns, particularly if *every* day is considered. Many days without severe weather are obvious to operational forecasters (e.g., lack of suitable moisture in a large region of the United States, owing to a large surface anticyclone) and may not need to be considered in such studies. Perhaps a useful alternative is to identify minimum-threshold environments for consideration (e.g., nonzero SBCAPE or most-unstable CAPE; dew point temperatures above a certain threshold; nonzero EHI, SCP, or STP; etc.). The objective is to consider true “null events” to determine the degree to which the false alarm problem worsens if any day or event is considered versus those for which at least some severe weather occurs in a distinct geographic cluster.

Third, identification of physical processes associated with the occurrence or absence of major tornado outbreaks, as well as increased areal coverage of parameters favorable for significant severe weather, is essential. To this end, development and idealized simulations of outbreak composites can be used to increase our physical understanding of these events. Possible topics for consideration include the addition of targeted perturbations of initial conditions to determine model simulation sensitivities toward particular environmental parameters at various times before the event, examination of synoptic-scale variables (e.g., Q-vector convergence, quasigeostrophic vorticity or height tendency, etc.) in the hopes of identifying consistent signals in major tornado outbreaks that are not observed with other types of events, and evaluation of model simulations for each composite compared to individual events designated as a particular type of outbreak.

Of course, major tornado outbreaks are not the only type of outbreak to consider. Forecasts of major hail and/or major wind events are also important. The ranking scheme developed by Shafer and Doswell 2010b allows for such investigations. If the ranking index scores are decomposed into the four severe components (tornado, wind, hail, and miscellaneous – see Section 3e), objective techniques can be evaluated to determine if a particular event could be associated with an above-average "component" to the index. For example, areal coverage of SCP was tested to determine the skill the severe weather parameter had in diagnosing days in which the tornado-, hail-, wind-, and miscellaneous-components of the ranking indices exceeds the mean value (zero) of the 6072 cases (for the 4057 cases occurring between 1979 and 2008; Fig. 5.1). For SCP, skill was exhibited for all four of the individual components, though the wind component featured



(A) SCP – N3 Tornado Component

(B) SCP – N3 Hail Component

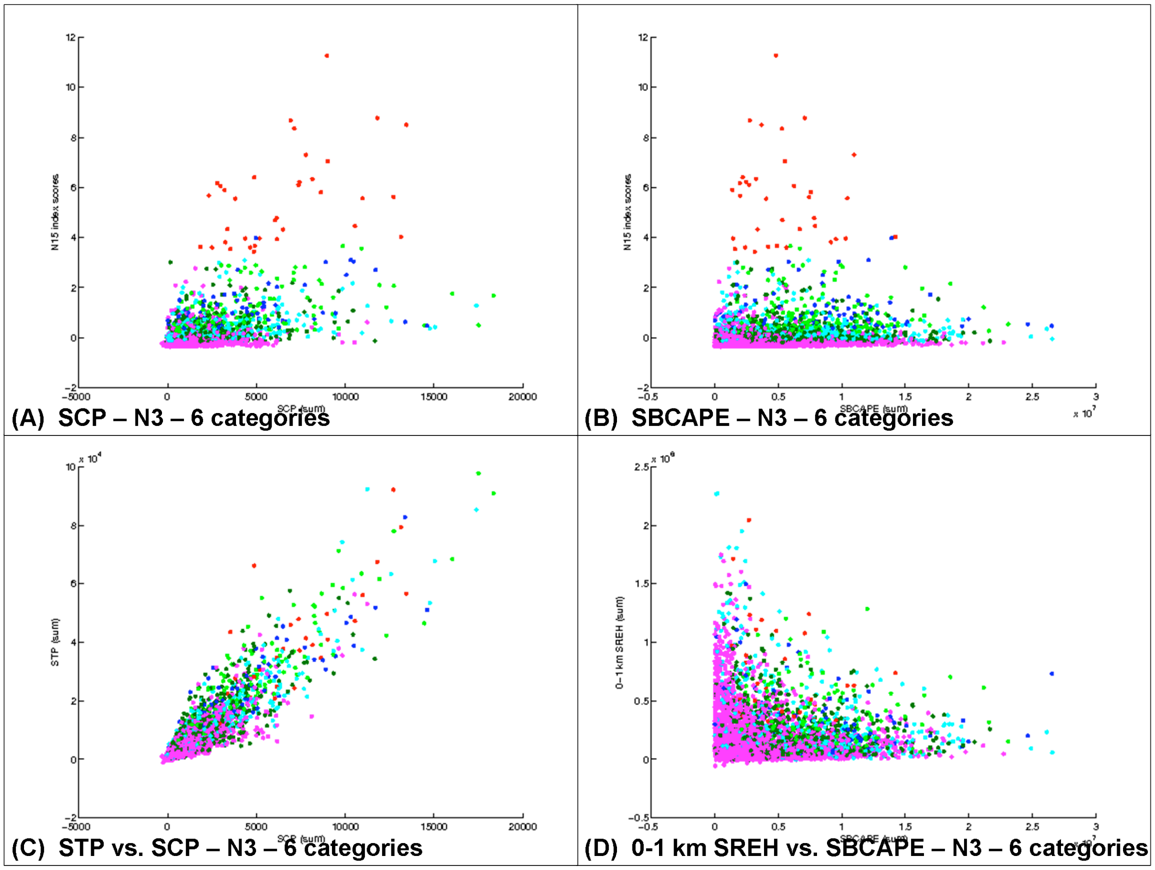
(C) SCP – N3 Wind Component

(D) SCP – N3 Miscellaneous Component

Fig. 5.1. Skill scores (*y*-axis) as a function of the areal coverage of SCP (KDE method; *x*-axis) for discriminating events that are above the mean value of the (a) tornado-, (b) hail-, (c) wind-, and (d) miscellaneous-component of the four-dimensional N3 outbreak ranking index to those below the mean value.

considerably lower skill than the other three components. Note that the miscellaneous component of the index (N3 in Fig. 5.1; note that this is consistent with using N15 – see SD10a, their Sections 3c,d) is comprised of the total number of reports for a given event cluster and the density ratio of the reports.

Additionally, the outbreaks were classified into categories using the four-dimensional decomposition with cluster analysis (Section 3e). Examining how the various categories cluster (or do not cluster) using various severe weather areal coverage parameters (e.g., Fig. 5.2) provides insight into the ability of these variables to diagnose the particular types of outbreaks. As Fig. 5.2 suggests, most characteristic types of



(A) SCP – N3 – 6 categories

(B) SBCAPE – N3 – 6 categories

(C) STP vs. SCP – N3 – 6 categories

(D) 0-1 km SREH vs. SBCAPE – N3 – 6 categories

Fig. 5.2. (a) Areal coverage (KDE method) of SCP (*x*-axis) and N15 scores (*y*-axis). Dots are colored based on 6-category *k*-means clustering of the 4-dimensional decomposition of the N3 index of the 6072 cases ranked in Section 3, with red dots indicating major tornado outbreaks, light green dots indicating hail-dominant outbreaks, blue dots indicating wind-dominant outbreaks, cyan dots indicating mixed-mode events with a slight propensity for more wind events, dark green dots indicating mixed-mode events with a slight propensity for more hail events, and magenta dots indicating mixed-mode, relatively minor events. (b) As in (a), using SBCAPE. (c) As in (a), using areal coverage of STP (*y*-axis). (d) As in (b), using areal coverage of 0-1 km SREH (*y*-axis).

outbreaks do not cluster appreciably. Hail-dominant and wind-dominant events have significant overlap, suggesting their discrimination would be quite difficult using most severe weather parameters. Statistical analyses similar to those shown in Section 4 can be conducted on certain pairs of classes (e.g., hail-dominant versus wind-dominant) or all classes (via statistical techniques such as multi-categorical support vector machines) to

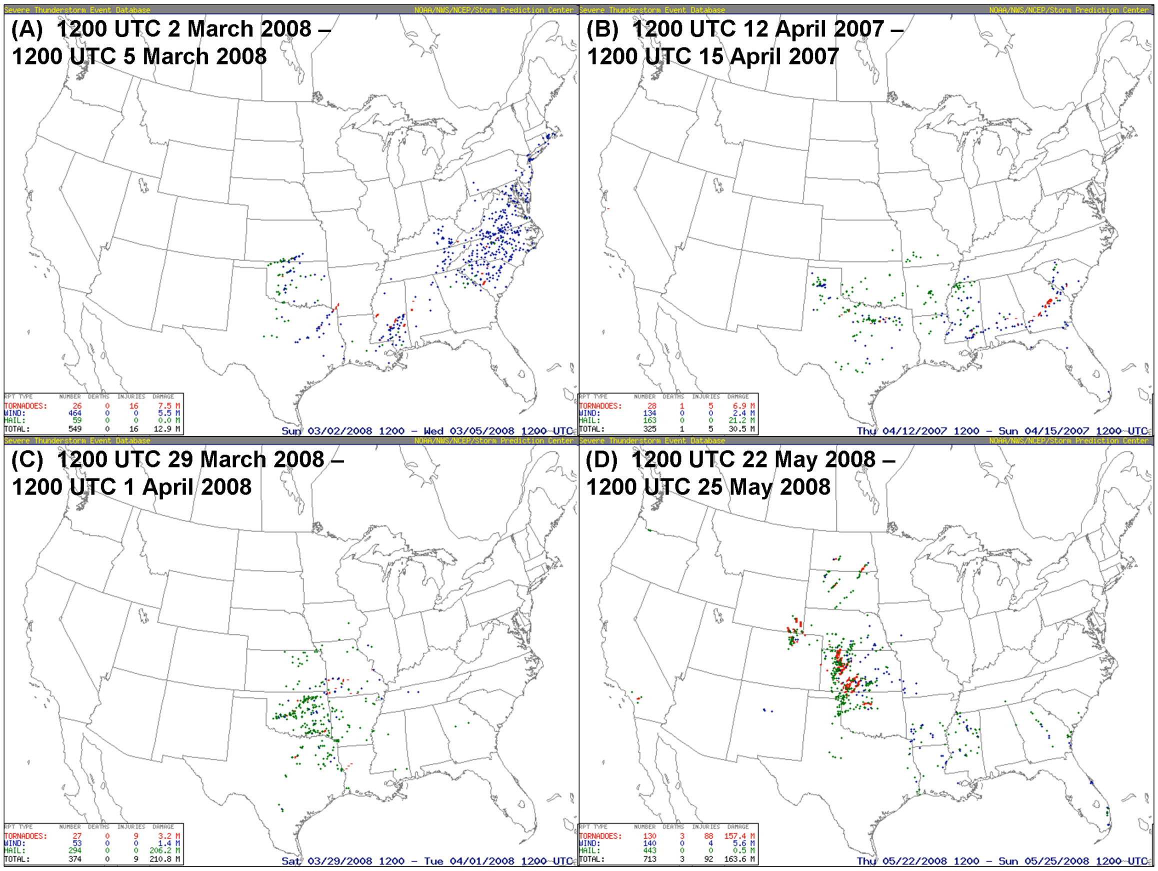

Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(A) 1200 UTC 2 March 2008 –
1200 UTC 5 March 2008
RPT TYPE NUMBER DEATHS INJURIES DAMAGE
TORNADOES: 26 0 16 7.5 M
WIND: 464 0 0 5.5 M
HAIL: 59 0 0 0.0 M
TOTAL: 549 0 16 12.9 M
Sun 03/02/2008 1200 – Wed 03/05/2008 1200 UTC
Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(B) 1200 UTC 12 April 2007 –
1200 UTC 15 April 2007
RPT TYPE NUMBER DEATHS INJURIES DAMAGE
TORNADOES: 28 1 5 6.9 M
WIND: 134 0 0 2.4 M
HAIL: 163 0 0 21.2 M
TOTAL: 325 1 5 30.5 M
Thu 04/12/2007 1200 – Sun 04/15/2007 1200 UTC
Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(C) 1200 UTC 29 March 2008 –
1200 UTC 1 April 2008
RPT TYPE NUMBER DEATHS INJURIES DAMAGE
TORNADOES: 27 0 9 3.2 M
WIND: 53 0 0 1.4 M
HAIL: 294 0 0 206.2 M
TOTAL: 374 0 9 210.8 M
Sat 03/29/2008 1200 – Tue 04/01/2008 1200 UTC
Severe Thunderstorm Event Database
NOAA/NWS/NCEP/Storm Prediction Center
(D) 1200 UTC 22 May 2008 –
1200 UTC 25 May 2008
RPT TYPE NUMBER DEATHS INJURIES DAMAGE
TORNADOES: 130 3 88 157.4 M
WIND: 140 0 4 5.6 M
HAIL: 443 0 0 0.5 M
TOTAL: 713 3 92 163.6 M
Thu 05/22/2008 1200 – Sun 05/25/2008 1200 UTC

Fig. 5.3. As in Fig. 4.25, for the time periods labeled.

determine the utility of a range of severe weather parameters to discriminate these types of events from each other.

Another focus for future research should include multi-day outbreaks. These events, currently accounted for as separate daily outbreaks for the work presented in Section 3, occur over multiple diurnal periods. They generally can be categorized as two types. The first features a progressive, long-track cluster of severe weather reports associated with a single synoptic-scale system (e.g., Figs. 5.3a,b). These types of events may have diurnal maxima and minima, associated with daytime heating and nocturnal boundary layer stabilization, but clearly evolve over a long stretch of the nation. The second type of multi-day event occurs in the same general vicinity over consecutive days (e.g., Figs. 5.3c,d). These events typically feature a nearly stationary longwave trough with multiple progressive smaller-scale vorticity maxima associated with individual clusters of reports at distinct times.

Multi-day outbreaks provide several foci for additional research. For example, future work can address the association between areal coverage of severe weather parameters and peak frequency and/or intensity of the outbreaks, particularly with significant tornadoes. Additionally, identification of any consistent signals with the synoptic-scale systems associated with the most significant multi-day outbreaks would be appropriate. For example, are anomalously strong synoptic-scale systems (which could be measured in various ways) more likely to produce multi-day outbreaks? Evaluation of mesoscale model simulations of multi-day outbreaks may provide insight into the predictability of such events, and the sensitivities of the evolution of these outbreaks to the synoptic and subsynoptic environments, scale interactions, and convective modification. This research could provide a means for forecasting the evolution of outbreaks in a systematic way, rather than providing a single 24-h forecast.

Finally, can antecedent environmental conditions be used to predict the relative severity of convective outbreaks? In Sections 2 and 4 herein, severe weather parameters were used as diagnostic variables, where the values of these parameters were valid roughly when the event was ongoing. Evaluation of these variables as forecast parameters, valid at times preceding the event of interest, may be valuable (see Doswell and Schultz 2006 for a discussion). It is quite common in short-term forecasting of severe weather to use the current values of severe weather parameters to predict the characteristics of the severe weather a short time in the future. This necessitates evaluation of these parameters that are valid before the event occurs. For outbreak events, a dataset such as NARR can be obtained before the event (e.g., rather than using the fields valid just before the median time of the event, one can use the fields three hours in advance of the fields valid before the median time of the event). Datasets with higher temporal resolution (e.g., RUC analyses) may be even more helpful in such investigations, though this typically results in a smaller sample of cases that can be tested.

With a large number of studies in the literature that are at least somewhat associated with tornado outbreaks (over 800 in the AMS journals alone), there are very few that look at a large sample of these events to examine general characteristics and how these differ from less significant outbreaks. The results presented herein and discussed in M09, S09, Mercer et al. 2010, S10a, S10b, Shafer et al. (2010c), SD10a, and Shafer and Doswell (2010b) suggest that our understanding of these events, although certainly improved from initial investigations in the 1950s and 1960s, remains quite limited. It is hoped that the work presented in this paper provides substantial support for a renewed focus on these high-impact events, particularly in efforts to increase our *physical understanding* of the underlying processes responsible for the widespread development of (significant) tornadoes and to improve our observations of these events. Without this increased physical understanding, the false alarm problem that pervades outbreak discrimination currently will continue for the foreseeable future. Without systematic improvement in the observation of severe weather, uncertainties regarding an outbreak's meteorological significance and relative severity will plague evaluation and verification studies hereafter. Based on the significant threat these events pose to life and property, these limitations should be unacceptable to severe weather forecasters and researchers.

References

Barnes, L. R., E. C. Gruntfest, M. H. Hayden, D. M. Schultz, and C. Benight, 2007: False alarms and close calls: A conceptual model of warning accuracy. *Wea. Forecasting*, **22**, 1140–1147.

Barnes, S. L., 1964: A technique for maximizing details in numerical weather map analysis. *J. Appl. Meteor.*, **3**, 396–409.

Beebe, R. G., 1955: Types of air masses in which tornadoes occur. *Bull. Amer. Meteor. Soc.*, **36**, 349–350.

——, 1956: Tornado composite charts. *Mon. Wea. Rev.*, **84**, 127–142.

Blanchard, D. O., 1998: Assessing the vertical distribution of convective available potential energy. *Wea. Forecasting*, **13**, 870–877.

Bowman, A. W., and A. Azzalini, 1997: *Applied Smoothing Techniques for Data Analysis: the Kernel Approach Using S-Plus Illustrations*. Oxford University Press, 208 pp.

Breiman, L., J. Friedman, R. Olshen, and C. Stone, 1993: *Classification and Regression Trees*. Wadsworth, 358 pp.

Breznitz, S. 1984: *Cry Wolf: The Psychology of False Alarms*. Lawrence Earlbaum Associates, 265 pp.

Briggs, W., 2005: A general method of incorporating forecast cost and loss in value scores. *Mon. Wea. Rev.*, **133**, 3393–3397.

Brooks, H. E., 2004a: On the relationship of tornado path length and width to intensity. *Wea. Forecasting*, **19**, 310–319.

——, 2004b: Tornado-warning performance in the past and future: A perspective from signal detection theory. *Bull. Amer. Meteor. Soc.*, **85**, 837–843.

——, C. A. Doswell III, and J. Cooper, 1994a: On the environments of tornadic and nontornadic mesocyclones. *Wea. Forecasting*, **9**, 606–618.

——, ——, and R. B. Wilhelmson, 1994b: On the role of midtropospheric winds in the evolution and maintenance of low-level mesocyclones. *Mon. Wea. Rev.*, **122**, 126–136.

——, ——, and M. P. Kay, 2003a: Climatological estimates of local daily tornado probability for the United States. Wea. Forecasting, 18, 626–640.

——, J. W. Lee, and J. P. Craven, 2003b: The spatial distributions of severe thunderstorm and tornado environments from global reanalysis data. *Atmos. Res.*, **67–68**, 73–94.

——, M. Kay, and J. A. Hart, 1998: Objective limits on forecasting skill for rare events. *Preprints*, Nineteenth Conf. on Severe Local Storms, Minneapolis, MN, Amer. Meteor. Soc., 552–555.

Brown, B. G., and A. H. Murphy, 1996: Verification of aircraft icing forecasts: The use of standard measures and meteorological covariates. Preprints, *13th Conf. on Probability and Statistics in the Atmospheric Sciences*, San Francisco, CA, Amer. Meteor. Soc., 251–252.

Bunkers, M. J., 2002: Vertical wind shear associated with left-moving supercells. *Wea. Forecasting*, 17, 845–855.

Colby, F. P., Jr., 1984: Convective inhibition as a predictor of convection during AVE-SESAME II. *Mon. Wea. Rev.,* **112**, 2239–2252.

Collins, W. G., and L. S. Gandin, 1990: Comprehensive hydrostatic quality control at the National Meteorological Center. *Mon. Wea. Rev.*, **112**, 2239–2252.

——, and ——, 1992: Complex quality control of rawinsonde heights and temperatures (CQCHT) at the National Meteorological Center. NMC Office Note 390, 30 pp. [Available from NOAA/NCEP, 5200 Auth Rd., Washington, DC, 20233.]

Corfidi, S. F., S. J. Weiss, J. S. Cain, S. J. Corfidi, R. M. Rabin, and J. J. Levit, 2010: Revisiting the 3-4 1974 super outbreak of tornadoes. *Wea. Forecasting*, **25**, 465–510.

Cortinas, J. V., Jr., and D. J. Stensrud, 1995: The importance of understanding mesoscale model parameterization schemes for weather forecasting. *Wea. Forecasting*, **10**, 716–740.

Craven, J. P., H. E. Brooks, and J. A. Hart, 2002a: Baseline climatology of sounding derived parameters associated with deep, moist convection. Preprints, *21st Conf. on Severe Local Storms*, San Antonio, TX, Amer. Meteor. Soc., 643–646.

——, R. E. Jewell, and H. E. Brooks, 2002b: Comparison between observed convective cloud-base heights and lifting condensation level for two different lifted parcels. *Wea. Forecasting*, **17**, 885–890.

Cristianini, N., and J. Shawe-Taylor, 2000: *Support Vector Machines and other kernel-based learning methods.* Cambridge University Press, Cambridge, England, 189 pp.

Davies, J., and R. Johns, 1993: Some wind and instability parameters associated with strong and violent tornadoes. Part I: Wind shear and helicity. *The Tornado: Its*

Structure, Dynamics, Prediction and Hazards, Geophys. Monogr., No. 79, Amer. Geophys. Union, 573–582.

Davies-Jones, R., D. Burgess, and M. Foster, 1990: Test of helicity as a tornado forecast parameter. Preprints, *16th Conf. on Severe Local Storms*, Kananaskis Park, AB, Canada, Amer. Meteor. Soc., 588–592.

Deng, A., and D. R. Stauffer, 2006: On improving 4-km mesoscale model simulations. *J. Appl. Meteor.*, **45**, 361–381.

Dial, G. L., J. P. Racy, and R. L. Thompson, 2010: Short-term convective mode evolution along synoptic boundaries. *Wea. Forecasting*, **in press**.

Doswell, C. A. III, 2004: Weather forecasting by humans – Heuristics and decision making. *Wea. Forecasting*, **19**, 1115–1126.

——, 2007a: Historical overview of severe convective storms research. *Electronic J. Severe Storms Meteor.*, **2 (1)**, 1–25.

——, 2007b: Small sample size and data quality issues illustrated using tornado occurrence data. Electronic J. Severe Storms Meteor., **2** (5), 1–16.

——, and L. F. Bosart, 2001: Extratropical synoptic-scale processes and severe convection. *Severe Convective Storms, Meteor. Monogr.*, No. 27, Amer. Meteor. Soc., 1–27.

——, and J. S. Evans, 2003: Proximity sounding analysis for derechos and supercells: An assessment of similarities and differences. *Atmos. Res.*, **67–68**, 117–133.

——, and D. M. Schultz, 2006: On the use of indices and parameters in forecasting severe storms. *Electronic J. Severe Storm Meteor.*, **1** (3), 1–14.

——, R. Davies-Jones, and D. L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting*, **5**, 576–585.

——, R. H. Johns, and S. J. Weiss, 1993: Tornado forecasting: A review. *The Tornado: Its Structure, Dynamics, Prediction and Hazards, Geophys. Monogr.*, Vol. 79, Amer. Geophys. Union, 557–571.

——, D. V. Baker, and C. A. Liles, 2002: Recognition of negative mesoscale factors for severe-weather potential: A case study. *Wea. Forecasting*, **17**, 937–954.

——, H. E. Brooks, and M. P. Kay, 2005: Climatological estimates of daily local nontornadic severe thunderstorm probability for the United States. *Wea. Forecasting*, **20**, 577–595.

——, R. Edwards, R. L. Thompson, J. A. Hart, and K. C. Crosbie, 2006: A simple and flexible method for ranking severe weather events. *Wea. Forecasting*, **20**, 577–595.

Droegemeier, K. K., S. M. Lazarius, and R. Davies-Jones, 1993: The influence of helicity on numerically simulated convective storms. *Mon. Wea. Rev.*, **121**, 2005–2029.

Dudhia, J., 1989: Numerical study of convection observed during the winter monsoon experiment using a mesoscale two-dimensional model. *J. Atmos. Sci.*, **46**, 3077 – 3107.

Efron, B., and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap*. Chapman and Hall/CRC, Boca Raton, Florida. 436 pp.

Fawbush, W. J., and R. C. Miller, 1952: A mean sounding representative of the tornadic air mass environment. *Bull. Amer. Meteor. Soc.*, **33**, 303–307.

——, and ——, 1954: A basis for forecasting peak wind gusts in nonfrontal thunderstorms. *Bull. Amer. Meteor. Soc.*, **46**, 3077–3107.

Ferguson, E. W., F. P. Ostby, and P. W. Leftwich Jr., 1987: The tornado season of 1985. *Mon. Wea. Rev.*, **115**, 1437–1445.

Fujita, T. T., 1974: Jumbo outbreak of 3 April 1974. *Weatherwise*, **27** (3), 116–126.

——, 1981: Tornadoes and downbursts in the context of generalized planetary scales. *J. Atmos. Sci.*, **38**, 1511–1534.

——, D. L. Bradbury, and C. F. van Thullenar, 1970: Palm Sunday tornadoes of April 11, 1965. *Mon. Wea. Rev.*, **98**, 29–69.

Glickman, T. S., Ed., 2000: *Glossary of Meteorology*, 2d ed. Amer. Meteor. Soc., 782 pp.

Gong, X., and M. B. Richman, 1995: On the application of cluster analysis to growing season precipitation data in North America east of the Rockies. *J. Climate*, **8**, 897–931.

Grell, G. A., and D. Devenyi, 2002: A generalized approach to parameterizing convection combining ensemble and data assimilation techniques. *Geophys. Res. Lett.*, **29**, 1693, doi:10.1029/2002GL015311.

Hamill, T. M., R. S. Schneider, H. E. Brooks, G. S. Forbes, H. B. Bluestein, M. Steinberg, D. Melendez, and R. M. Dole, 2005: The May 2003 extended tornado outbreak. *Bull. Amer. Meteor. Soc.*, **86**, 531–542.

Hart, J. A., and W. Korotky, 1991: The SHARP workstation v1.50 users guide. NOAA/National Weather Service, 30 pp. [Available from NWS Eastern Region Headquarters, 630 Johnson Ave., Bohemia, NY 11716.]

Haykin, S., 1999: *Neural Networks: A Comprehensive Foundation*. Pearson Education, 842 pp.

Hodges, K. I., 2008: Confidence intervals and significance tests for spherical data derived from feature tracking. *Mon. Wea. Rev.*, **136**, 1758–1777.

Hong, S.-Y., and H.-L. Pan, 1996: Nonlocal boundary layer vertical diffusion in a medium-range forecast model. *Mon. Wea. Rev.*, **124**, 2322–2329.

——, H.-M. Juang, and Q. Zhao, 1998: Implementation of prognostic cloud scheme for a regional spectral model. *Mon. Wea. Rev.*, **126**, 2621–2639.

Hotelling, H., 1933: Analysis of a complex of statistical variables into principal components. *J. Educ. Phys.*, **24**, 417–441, 498–520.

James, R. P., J. M. Fritsch, and P. M. Markowski, 2005: Environmental distinctions between cellular and slabular convective lines. *Mon. Wea. Rev.*, **133**, 2669–2691.

Ji, M., A. Kumar, and A. Leetmaa, 1994: A multiseason climate forecast system at the National Meteorological Center. *Bull. Amer. Meteor. Soc.*, 75, 557–569.

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

——, and C. Doswell III, 1992: Severe local storms forecasting. *Wea. Forecasting*, **7**, 588–612.

——, and J. A. Hart, 1993: Differentiating between types of severe thunderstorm outbreaks: A preliminary investigation. Preprints, *17th Conf. on Severe Local Storms*, Saint Louis, MO, Amer. Meteor. Soc., 46–50.

——, J. Davies, and P. Leftwich, 1993: Some wind and instability parameters associated with strong and violent tornadoes. Part II: Variations in the combinations of wind and instability parameters. *The Tornado: Its Structure, Dynamics, Prediction and Hazards, Geophys. Monogr.*, No. 79, Amer. Geophys. Union, 583–590.

Kain, J. S., S. J. Weiss, J. J. Levit, M. E. Baldwin, and D. R. Bright, 2006: Examination of convection-allowing configurations of the WRF model for the prediction of severe convective weather: The SPC/NSSL Spring Program 2004. *Wea. Forecasting*, **21**, 167–181.

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

Kanamitsu, M., 1989: Description of the NMC global data assimilation and forecast system. *Wea. Forecasting*, **4**, 334–342.

——, and Coauthors, 1991: Recent changes implemented into the global forecast system at NMC. *Wea. Forecasting*, **6**, 425–435.

Katz, R. W., and A. H. Murphy, 1997: *Economic Value of Weather and Climate Forecasts*. Cambridge University Press, 222 pp.

Kaufman, L., and P. Rousseeuw, 1990: *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, 342 pp.

Kerr, B. W., and G. L. Darkow, 1996: Storm-relative winds and helicity in the tornadic thunderstorm environment. *Wea. Forecasting*, **11**, 489–505.

King, P. S. W., 1997: On the absence of population bias in the tornado climatology of southwestern Ontario. *Wea. Forecasting*, **12**, 939–946.

Krzanowski, W. J., 1988: *Principles of Multivariate Analysis: A User's Perspective*. Oxford University Press, 563 pp.

Lin, Y.-L., R. D. Farley, and H. D. Orville, 1983: Bulk parameterization of the snow field in a cloud model. *J. Climate Appl. Meteor.*, **22**, 1065–1092.

Lee, Y., G. Wahba, and S. A. Ackerman, 2004: Cloud classification of satellite radiance data by multicategory support vector machines. *J. Atmos. Oceanic Technol.*, **21**, 159–169.

Maddox, R. A., 1976: An evaluation of tornado proximity wind and stability data. *Mon. Wea. Rev.*, **104**, 133–142.

Markowski, P. M., E. N. Rasmussen, and J. M. Straka, 1998a: The occurrence of tornadoes in supercells interacting with boundaries during VORTEX-95. *Wea. Forecasting,* **13,** 852–859.

——, J. M. Straka, E. N. Rasmussen, and D. O. Blanchard, 1998b: Variability of storm-relative helicity during VORTEX. *Mon. Wea. Rev.,* **11,** 2959–2971.

——, C. Hannon, J. Frame, E. Lancaster, A. Pietrycha, R. Edwards, and R. L. Thompson, 2003: Characteristics of vertical wind profiles near supercells obtained from the Rapid Update Cycle. *Wea. Forecasting*, **18**, 1262–1272.

Mercer, A. E., M. B. Richman, H. B. Bluestein, and J. M. Brown, 2008: Statistical modeling of downslope windstorms in Boulder, Colorado. *Wea. Forecasting*, **23**, 1176–1194.

——, C. M. Shafer, C. A. Doswell III, L. M. Leslie, and M. B. Richman, 2009: Objective classification of tornadic and non-tornadic outbreaks. *Mon. Wea. Rev.*, **137**, 4355–4368.

——, ——, ——,——, and ——, 2010: Synoptic composites of tornadic and nontornadic outbreaks. *Mon. Wea. Rev.*, **in review**.

Mesinger F., Coauthors, 2004: NCEP North American regional reanalysis. Preprints, *15th Symp. on Global Change and Climate Variations,* Seattle, WA, Amer. Meteor. Soc., CD-ROM, P1.1.

Miller, R., 1972: Notes on analysis and severe-storm forecasting procedures of the Air Force Global Weather Center. Air Weather Service Tech. Rep. 200 (rev.), Air Weather Service, Scott Air Force Base, IL, 184 pp. [Available online at http://stinet.dtic.mil/cgi-bin/GetTRDoc?AD=AD744042&Location=U2&doc=GetTRDoc.pdf.]

Mitchell K., Coauthors, 2004: NCEP completes 25-year North American Reanalysis: Precipitation assimilation and land surface are two hallmarks. *GEWEX Newsletter,* No. 14, International GEWEX Project Office, 9–12.

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res.*, **102** (D14), 16 663–16 682.

Moller, A. R., 2001: Severe local storms forecasting. *Severe Convective Storms, Meteor. Monogr.*, No. 50, Amer. Meteor. Soc., 433–480.

Moncrieff, M. W., and J. S. A. Green, 1972: The propagation and transfer properties of steady convective overturning in shear. *Quart. J. Roy. Meteor. Soc.*, **98**, 336–352.

Monteverdi, J. P., C. A. Doswell III, and G. S. Lipari, 2003: Shear parameter thresholds for forecasting tornadic thunderstorms in northern and central California. *Wea. Forecasting*, **18**, 357-370.

Murphy, A. H., 1977: The value of climatological, categorical, and probabilistic forecasts in the cost-loss ratio situation. *Mon. Wea. Rev.*, **105**, 803–816.

——, 1996: The Finley affair: a signal event in the history of forecast verification. *Wea. Forecasting*, **11**, 3–20.

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

Ostby, F. P., 1992: Operations of the National Severe Storms Forecast Center. *Wea. Forecasting*, **7**, 546–563.

Potvin, C. K., K. L. Elmore, and S. J. Weiss, 2010: Assessing the impact of proximity sounding criteria on the climatology of significant tornado environments. *Wea. Forecasting*, **in press**.

Rasmussen, E. N., and D. O. Blanchard, 1998: A baseline climatology of sounding-derived supercell and tornado forecast parameters. *Wea. Forecasting*, **13**, 1148–1164.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteorol. Soc.*, **126**, 649–667.

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

Roebber, P.J, 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608.

——, and L. F. Bosart, 1998: The complex relationship between forecast skill and forecast value: A real-world analysis. Wea. Forecasting, 11, 544–558.

Schaefer, J. T., 1986: Severe thunderstorm forecasting: A historical perspective. *Wea. Forecasting*, **1**, 164–189.

——, and R. Edwards, 1999: The SPC tornado/severe thunderstorm database. Preprints, *11th Conf. on Applied Climatology*, Dallas, TX, Amer. Meteor. Soc., 603–606.

Schwartz, C. S., J. S. Kain, S. J. Weiss, M. Xue, D. R. Bright, F. Kong, K. W. Thomas, J. J. Levit, M. C. Coniglio, and M. S. Wandishin, 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263–280.

Seber, G. A. F., 1984: *Multivariate Observations*. Wiley Press, 686 pp.

Shafer, C. M., and C. A. Doswell III, 2010a: A multivariate index for ranking and classifying severe weather outbreaks. *Electronic J. Severe Storms Meteor.*, **5** (1), 1–39.

——, and ——, 2010b: Using kernel density estimation to identify, rank, and classify severe weather outbreak events. *Electronic J. Severe Storms Meteor.*, **in review**.

——, ——, L. M. Leslie, and M. B. Richman, 2010a: On the use of areal coverage of parameters favorable for severe weather to discriminate major outbreaks. *Electronic J. Severe Storms Meteor.*, **accepted**.

——, A. E. Mercer, L. M. Leslie, M. B. Richman, and C. A. Doswell III, 2010b: Evaluation of WRF model simulations of tornadic and nontornadic outbreaks occurring in the spring and fall. *Mon. Wea. Rev.*, **in press**.

——, M. B. Richman, L. M. Leslie, C. A. Doswell III, and A. E. Mercer, 2010c: Diagnosing major severe weather outbreaks: Comparison of NARR and NCEP reanalysis datasets and evaluation of principal component and areal coverage techniques. *Mon. Wea. Rev.*, **in preparation**.

——, ——, C. A. Doswell III, M. B. Richman, and L. M. Leslie, 2009: Evaluation of WRF forecasts of tornadic and nontornadic outbreaks when initialized with synoptic-scale input. *Mon. Wea. Rev.*, **137**, 1250–1271.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers, 2005: A description of the Advanced Research WRF Version 2. *NCAR Tech. Note*, NCAR/TN-468+STR, 88 pp. [Available from UCAR Communications, P.O. Box 3000, Boulder, CO 80307.]

——, ——, ——, ——, ——, M. G. Duda, X.-Y. Huang, W. Wang, and J. G. Powers, 2008: A description of the Advanced Research WRF Version 3. *NCAR Tech. Note*, NCAR/TN-475+STR, 113 pp. [Available from UCAR Communications, P.O. Box 3000, Boulder, CO 80307.]

Sobash, R., J. S. Kain, M. C. Coniglio, A. R. Dean, D. R. Bright, and S. J. Weiss, 2010: Using convection-allowing models to produce forecast guidance for severe thunderstorm hazards vai a “surrogate severe” approach. Preprints, *25th Conf. on Severe Local Storms*, Denver, Colorado.

Speheger, D. A., C. A. Doswell III, and G. J. Stumpf, 2002: The tornadoes of 3 May 1999: Event verification in central Oklahoma and related issues. *Wea. Forecasting*, **17**, 362–381.

Stensrud, D. J., 2001: Using short-range ensemble forecasts for predicting severe weather events. *Atmos. Res.*, **56**, 3–17.

——, and J. M. Fritsch, 1994: Mesoscale convective systems in weakly forced large-scale environments. Part III: Numerical simulations and implications for numerical forecasting. *Mon. Wea. Rev.*, **122**, 2084–2104.

——, J. V. Cortinas, and H. E. Brooks, 1997: Discriminating between tornadic and nontornadic thunderstorms using mesoscale model output. *Wea. Forecasting*, **12**, 613–632.

——, J.-W. Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensembles of mesoscale convective systems. *Mon. Wea. Rev.*, **128**, 2077–2107.

Thompson, R. L., and M. D. Vescio, 1998: The destruction potential index—A method for comparing tornado days. Preprints, *19th Conf. on Severe Local Storms*, Minneapolis, MN, Amer. Meteor. Soc., 280–282.

——, and R. Edwards, 2000: An overview of environmental conditions and forecast implications of the 3 May 1999 tornado outbreak. *Wea. Forecasting*, **15**, 682–699.

——, ——, J. A. Hart, K. L. Elmore, and P. Markowski, 2003: Close proximity soundings with supercell environments obtained from the Rapid Update Cycle. *Wea. Forecasting*, **18**, 1243–1261.

——, C. M. Mead, and R. Edwards, 2007: Effective storm-relative helicity and bulk shear in supercell thunderstorm environments. *Wea. Forecasting*, **22**, 102–115.

Trafalis, T. B., M. B. Richman, I. Adrianto and S. Lakshmivarahan, 2009: Tornado detection with machine learning classifiers. Invited submission to *Journal of Communications, Network and System Science,* **in review**.

——, C. M. Mead, and R. Edwards, 2007: Effective storm-relative helicity and bulk shear in supercell thunderstorm environments. *Wea. Forecasting*, **22**, 102–115.

Verbout, S. M., H. E. Brooks, L. M. Leslie, and D. M. Schultz, 2006: Evolution of the U.S. tornado database: 1954-2003. *Wea. Forecasting*, **21**, 86–93.

Wandishin, M. S., and H. E. Brooks, 2002: On the relationship between Clayton's skill score and expected values for forecasts of binary events. *Meteorol. Appl.*, **9**, 455–459.

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

Weiss, S. J., J. A. Hart, and P. R. Janish, 2002: An examination of severe thunderstorm wind report climatology: 1970–1999. Preprints, *21st Conf. on Severe Local Storms*, San Antonio, TX, Amer. Meteor. Soc., 446–449.

Wicker, L. J., and W. C. Skamarock, 2002: Time splitting methods for elastic models using forward time schemes. *Mon. Wea. Rev.*, **130**, 2088–2097.

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

Woollen, J. S., 1991: New NMC operational OI quality control. Preprints, *Ninth Conf. on Numerical Weather Prediction*, Denver, CO, Amer. Meteor. Soc., 24–27.

——, E. Kalnay, L. Gandin, W. Collins, S. Saba, R. Kistler, M. Kanamitsu, and M. Chelliah, 1994: Quality control in the reanalysis system. Preprints, *10th Conf. on Numerical Weather Prediction*, Portland, OR, Amer. Meteor. Soc., 13–14.