Shafer, C. M., and C. A. Doswell III, 2011: Using kernel density estimation to identify, rank, and classify severe weather outbreak events. *Electronic J. Severe Storms Meteor.,* **6** (2)**,** 1–28.

Electronic Journal of
SEVERE STORMS METEOROLOGY

Using Kernel Density Estimation to Identify, Rank, and Classify Severe Weather Outbreak Events

CHAD M. SHAFER* AND CHARLES A. DOSWELL III
Cooperative Institute for Mesoscale Meteorological Studies, Norman, Oklahoma

(Submitted 10 August 2010; in final form 27 March 2011)

ABSTRACT

A method for ranking severe weather outbreaks of any type using a linear-weighted multivariate scheme has been introduced recently. The results of using this ranking method indicated that the scheme was capable of identifying the most significant severe weather outbreaks. However, the inclusion of days in which numerous reports were widely dispersed across a large region, or in which multiple clusters of reports that were geographically widely separated, was problematic. Though the studies included a variable (the so-called middle-50% parameter) that was effective in identifying these cases, a new way was needed to account for these days in a manner that agrees with subjective perceptions of these events. A candidate scheme introduced here uses nonparametric kernel density estimation to identify clusters of severe weather reports associated with a single severe weather event. Clusters with relatively few reports or sparse coverage within the region associated with the event then can be excluded quite easily. This technique also allows for multiple, regionally-separated clusters of severe reports to be considered in one day. After identifying clusters of severe weather events from 1960-2008, the cases are ranked and classified in a way similar to past research, using multivariate linear-weighting and cluster analysis, respectively. Results suggest that the most significant severe weather outbreaks again are identified appropriately, and the cases could be classified as major tornado, hail-dominant, wind-dominant, and minor mixed-mode events.

1. 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 (Doswell et al. 2006—hereafter D06) and for determining the relative severity of outbreaks of any type (Shafer and Doswell 2010—hereafter SD10). 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 relatively robust to modifications of the weights for the multiple variables used to rank the cases, and could be reproduced using the same technique.

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 SD10). 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 large

*Current affiliation: Department of Earth Sciences, University of South Alabama, Mobile, Alabama.

Corresponding author address: Chad Shafer, University of South Alabama Department of Earth Sciences, LSCB 136, Mobile, AL 36688
E-mail: cmshafer@usouthal.edu

number of widely dispersed reports separated from a 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 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 determined as a range of latitude and longitude (see Fig. 1). As shown, 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 demonstrated in Fig. 1). 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 (SD10).

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. 2). 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 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 2 describes the data and methods used to identify, rank, and classify these severe weather events. Section 3 demonstrates the characteristics of the techniques on various types of severe weather report clusters. Section 4 details the results of the rankings of the severe weather events. Section 5 presents the findings when classifying the severe weather events. Section 6 summarizes the study and discusses some remaining issues associated with the current work.

2. Data and methods

As in D06 and SD10, 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 SD10 (see their Table 1), except for the middle-50% parameter.

Each 24-h period from 1 January 1960 to 31 December 2008 was considered independently. 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 not considered, 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:

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

where n is the number of severe reports on a given day, K_h is a kernel function, and h is a

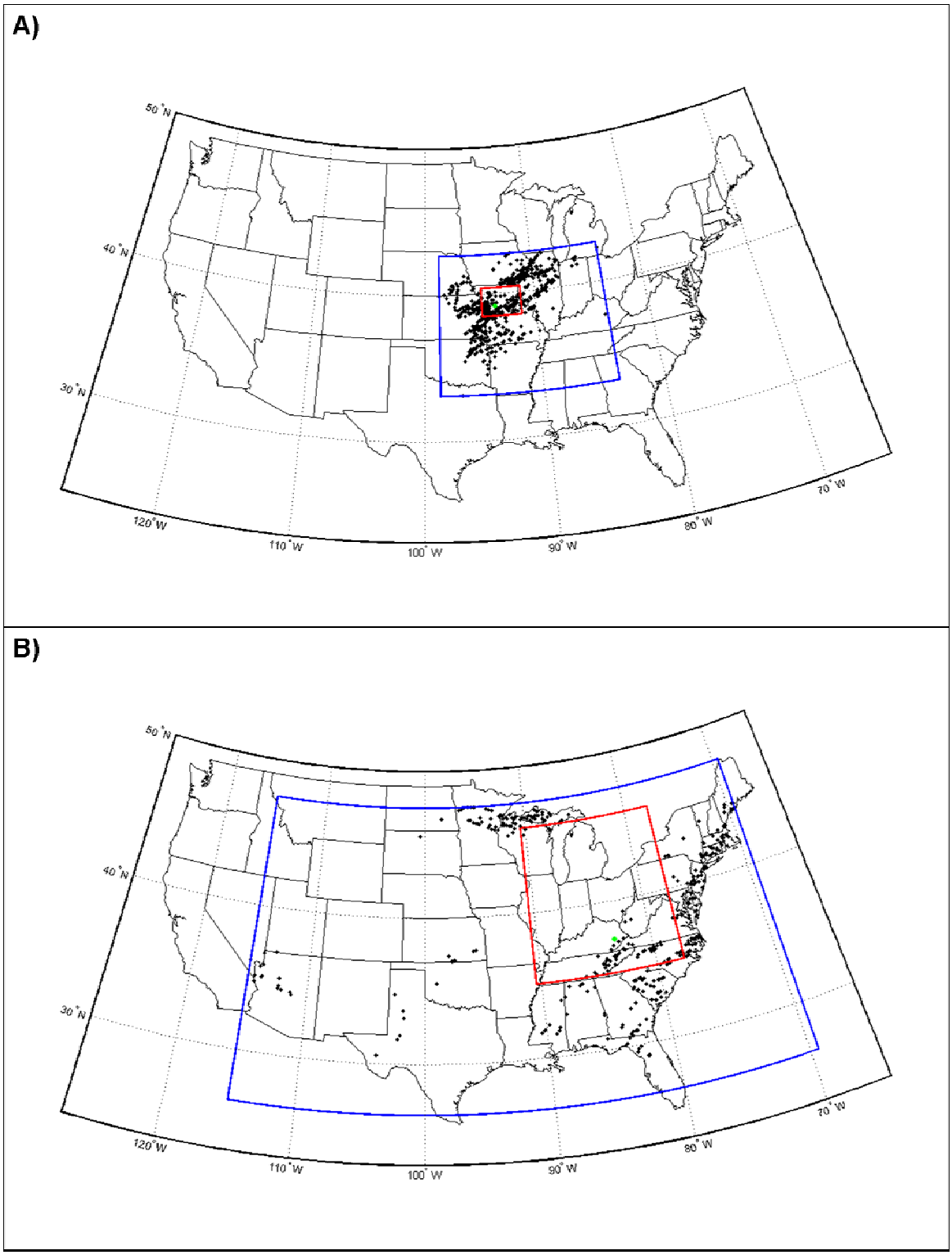

Figure 1: Examples of the middle-50% parameter for: a) 12 March 2006 and b) 28 July 2006. The blue box indicates the maximum and minimum latitudes and longitudes of severe weather reports (black dots). The red box indicates the 25th and 75th percentiles of the reports' latitudes and longitudes. The green dot indicates the median latitude and median longitude of severe reports. From Shafer and Doswell (2010). *Click image to enlarge.*

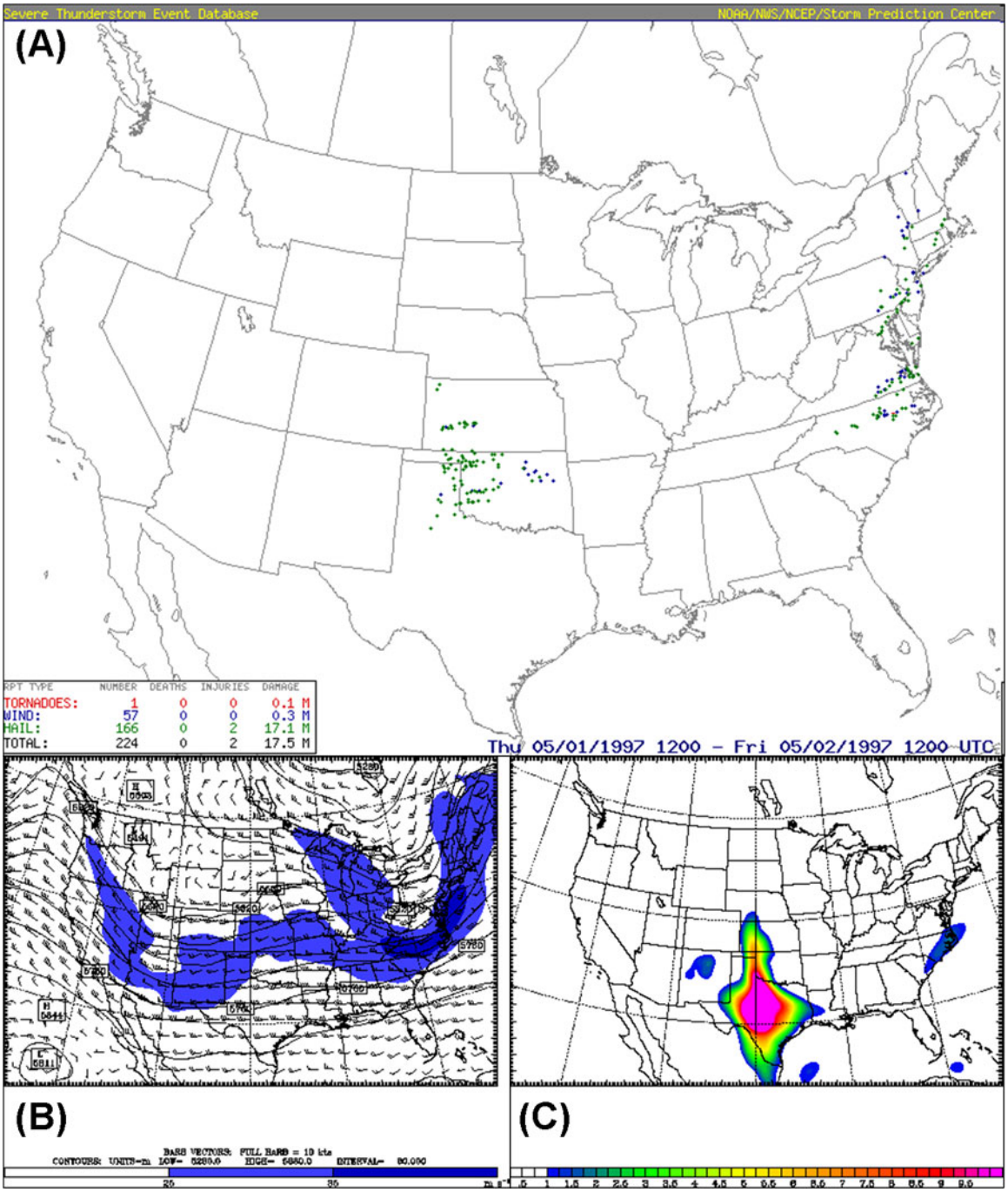

Figure 2: a) Severe reports on 1 May 1997, with severe wind gusts or wind damage in blue, severe hail in green, and tornadoes in red. b) North American Regional Reanalysis (NARR, after Mesinger et al. 2006) 500-hPa wind speeds (filled contours in m s^{-1}), winds (barbs in kt), and geopotential heights (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. *Click image to enlarge.*

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:

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

It can be shown that for multivariate KDE, (1) can be represented as:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n}\left[\prod_{k=1}^{d} K_{h_k}\left(x^{(k)} - x_i^{(k)}\right)\right] \qquad (3),$$

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=2, Eqn. (2) is modified for two dimensions by taking the square of itself, such that the quantity $(x - x_i)$ becomes a two-dimensional distance D_i. The end result is:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2\pi h^2}\exp\left[\frac{-D_i^2}{2h^2}\right]. \qquad (4)$$

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. (2)–(4) 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 shows, differences among the techniques and map projections were minor, as expected, so long as the bandwidth and PDF thresholds were modified accordingly.[1]

If the reports were converted to a grid initially, 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, though relatively coarse grid spacing is preferred because of reduced computational demand. For the latitude-longitude map projection, 1º grid spacing was used.

After all of the clusters for the 49-yr 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. 2003; Doswell et al. 2005; D06; Verbout et al. 2006; SD10), 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 SD10; 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

[1] Note that the PDF threshold must be modified if the bandwidth is modified, as *f(x)* is a function of the bandwidth [see Eqs. (1)–(4)].

the small values and exponential increase of the detrended means over the 49-yr period (e.g., Fig. 3), 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 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 study period. Thus, although we have attempted relatively simple accounting for secular changes in the dataset, some impact from those changes is inevitable.

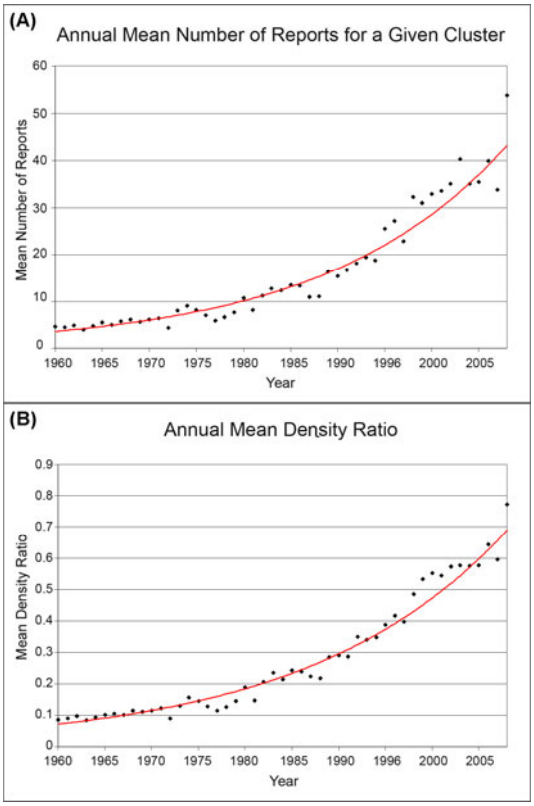

Figure 3: Examples of detrending for: a) the annual mean number of reports for a given cluster, and b) the annual mean density ratio, when considering all of the clusters for a particular year. The results are for a latitude-longitude map projection with 1° grid spacing spanning the conterminous United States, using a bandwidth of unity for each dimension, and a threshold probability of 0.001 for a grid point to be associated with each severe weather event. *Click image to enlarge.*

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 (SD10, their Table 1) were then 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 SD10 (examples in Fig. 4). 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) as follows:

$$\tilde{x}_i^{(j)} = \frac{x_i^{(j)} - \bar{x}_i}{s_i} \qquad (5).$$

In (5), i represents a particular member of the n variables used in the ranking index, and j is one of the m clusters considered for ranking. The mean is symbolized by $\bar{x}$, and the standard deviation is represented as s, with their well-known formulas:

$$\bar{x}_i = \frac{1}{m}\sum_{j=1}^{m} x_i^{(j)} \qquad (6),$$

$$s_i = \frac{1}{m-1}\sqrt{\sum_{j=1}^{m}[x_i^{(j)} - \mu_i]^2} \qquad (7).$$

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

$$I^{(j)} = \frac{\sum_{i=1}^{n} w_i \tilde{x}_i^{(j)}}{\sum_{i=1}^{n} w_i} \qquad (8).$$

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.

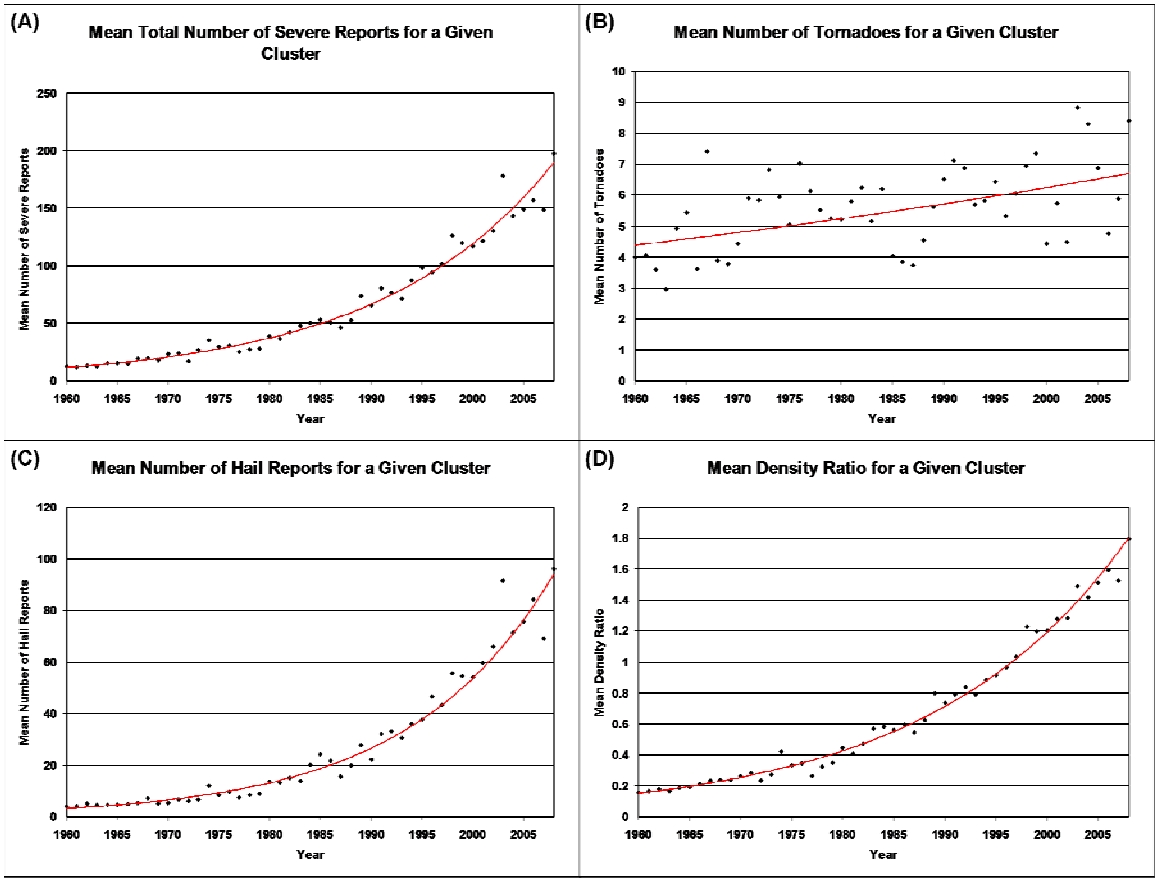

Figure 4: Examples of detrending for the annual means of the clusters with total number of reports and density ratios above the detrended annual means. Variables labeled in each chart. *Click image to enlarge.*

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 SD10 (see their Fig. 4 and 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 SD10).

Given that our objectives are similar to those of D06 and SD10, the techniques used in this study and that of SD10 are intentionally similar, as the former's technique was relatively simple to implement and easy to reproduce. Optimality of our 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. The reader is referred to D06 and SD10 for the reasons involved in the selection of the variables and the various methods used in the ranking of these events.

3. KDE analysis

To identify severe weather events by clusters of severe reports for a given day, modifying Eqn. (4) by changing the bandwidth (h) for a given map projection and analyzing various threshold values of $f(x)$ were required. 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). Using the grid point method and a latitude-longitude map projection with 1º grid spacing, analysis of the severe reports (Fig. 5a) and the resultant two-dimensional PDF contour charts for various modifications of the bandwidth and contour thresholds (Fig. 6) illustrate the process. This

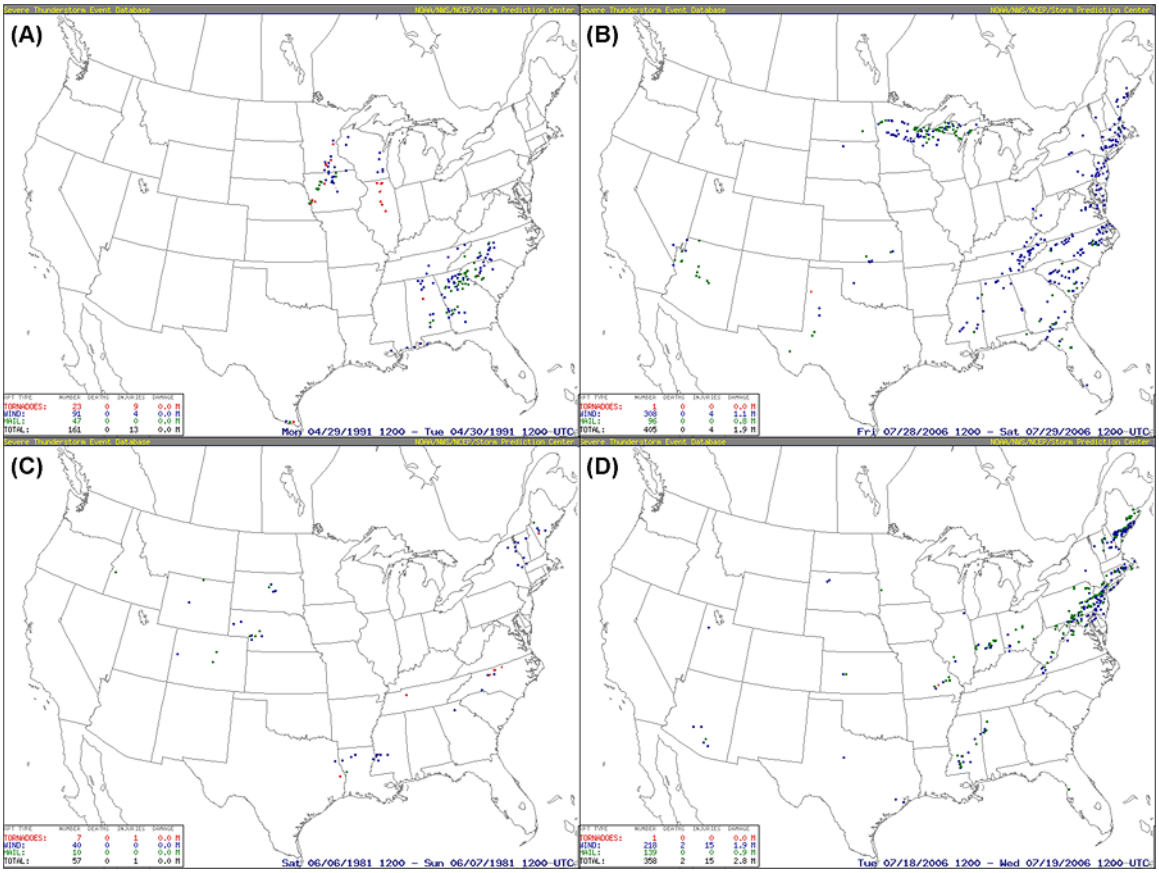

Figure 5: As in Fig. 2a, for a) 29 April 1991, b) 28 July 2006, c) 6 June 1981, and d) 18 July 2006. *Click image to enlarge.*

day featured three distinct regions of severe weather: the Southeast, the Upper 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. 6d), 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. 6a), given their relatively close proximity. Though there is some separation of the reports (Fig. 5a), 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.

The objective, therefore, is to find a range of bandwidths falling in between the two extremes (as in Figs. 6b,c). 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. 6, 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 reports while also not merging regionally separate events. For example, in Fig. 6d, 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, selecting a threshold too high results in separating relatively close regions, such as the 0.005 threshold (second contour) in Fig. 6b.

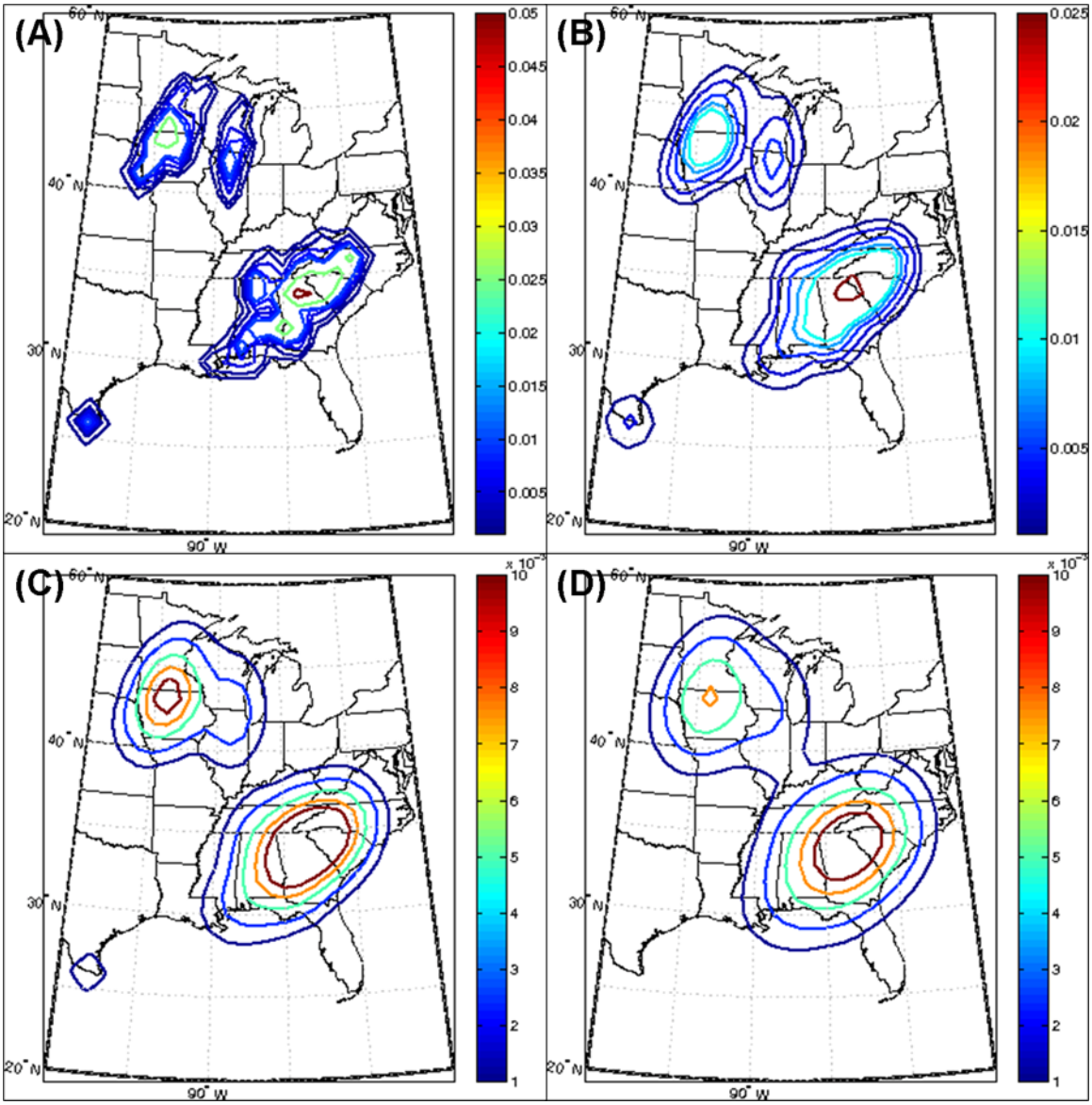

Figure 6: Two-dimensional KDEs of the probability density functions for severe reports from 1200 UTC 29 April 1991 to 1159 UTC 30 April 1991, using bandwidths of a) 0.5, b) 1, c) 1.5, and d) 2 for the latitudinal and longitudinal directions. Plots use severe reports converted to a latitude-longitude map projection with 1° grid spacing. For each plot, outermost contour is 0.001; second outermost contour is 0.005. *Click image to enlarge.*

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. 7a,c,e) are reasonable selections for various types of events. For example, as shown in Fig. 5b on 28 July 2006, distinct 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. 7a,b). On 6 June 1981, relatively few severe reports were scattered through the northern High Plains, Southeast, and Northeast (Fig. 5c). The two-dimensional PDFs were quite different between the two bandwidths, the smaller bandwidth being associated with a larger number of clusters (Figs. 7c,d).

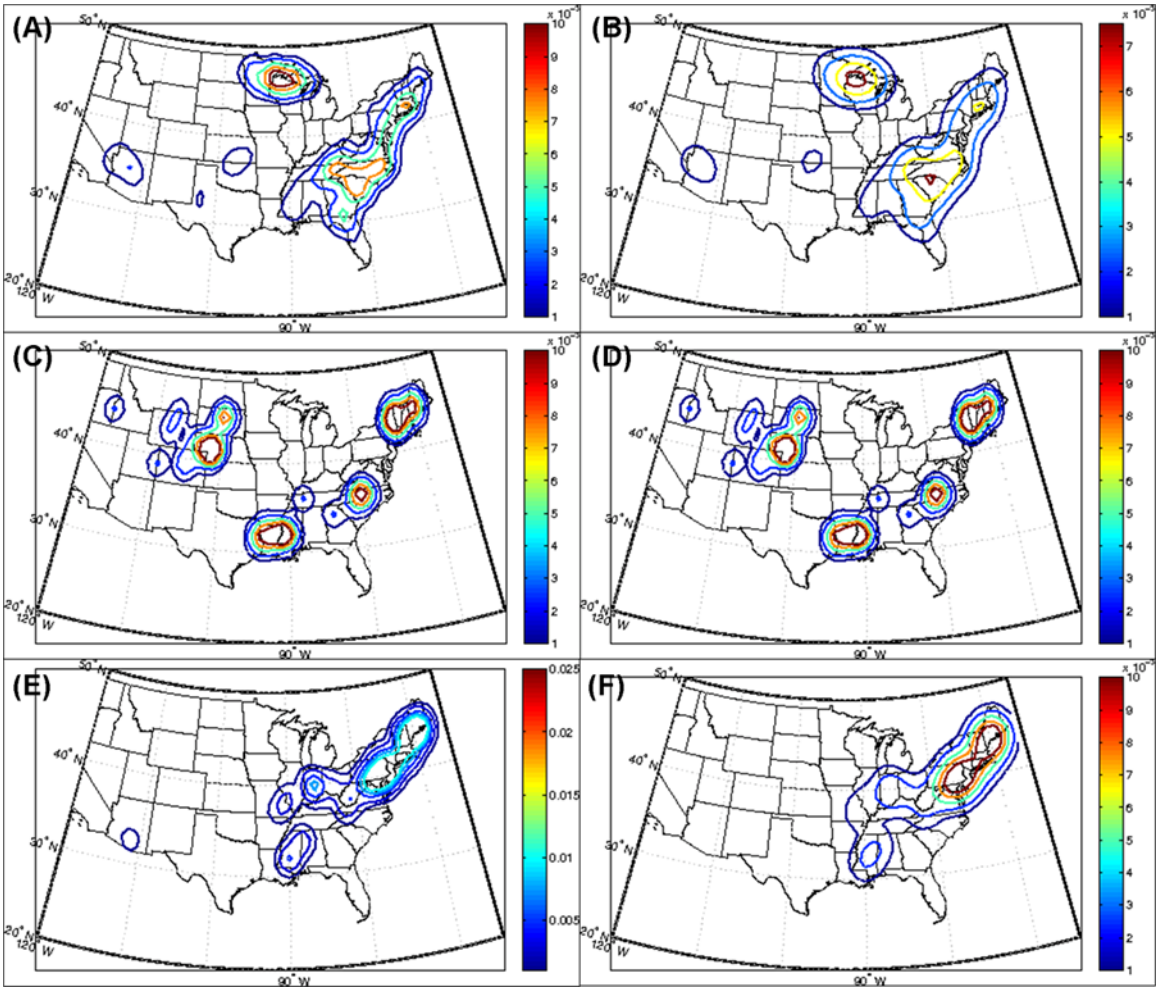

Figure 7: a) As in Fig. 6a, for 28 July 2006. b) As in Fig. 6b, for 28 July 2006. c) As in (a), for 6 June 1981. d) As in (b), for 6 June 1981. e) As in (a), for 18 July 2006. f) As in (b), for 18 July 2006. *Click image to enlarge.*

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 2 and 4) 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. 5d). 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. 7e). This was not the case for the bandwidth of 1.5, unless the PDF threshold was increased for the former and decreased for the latter (Fig. 7f). 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 1º 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. 8). If the bandwidth and PDF threshold were identical (in magnitude) to that of

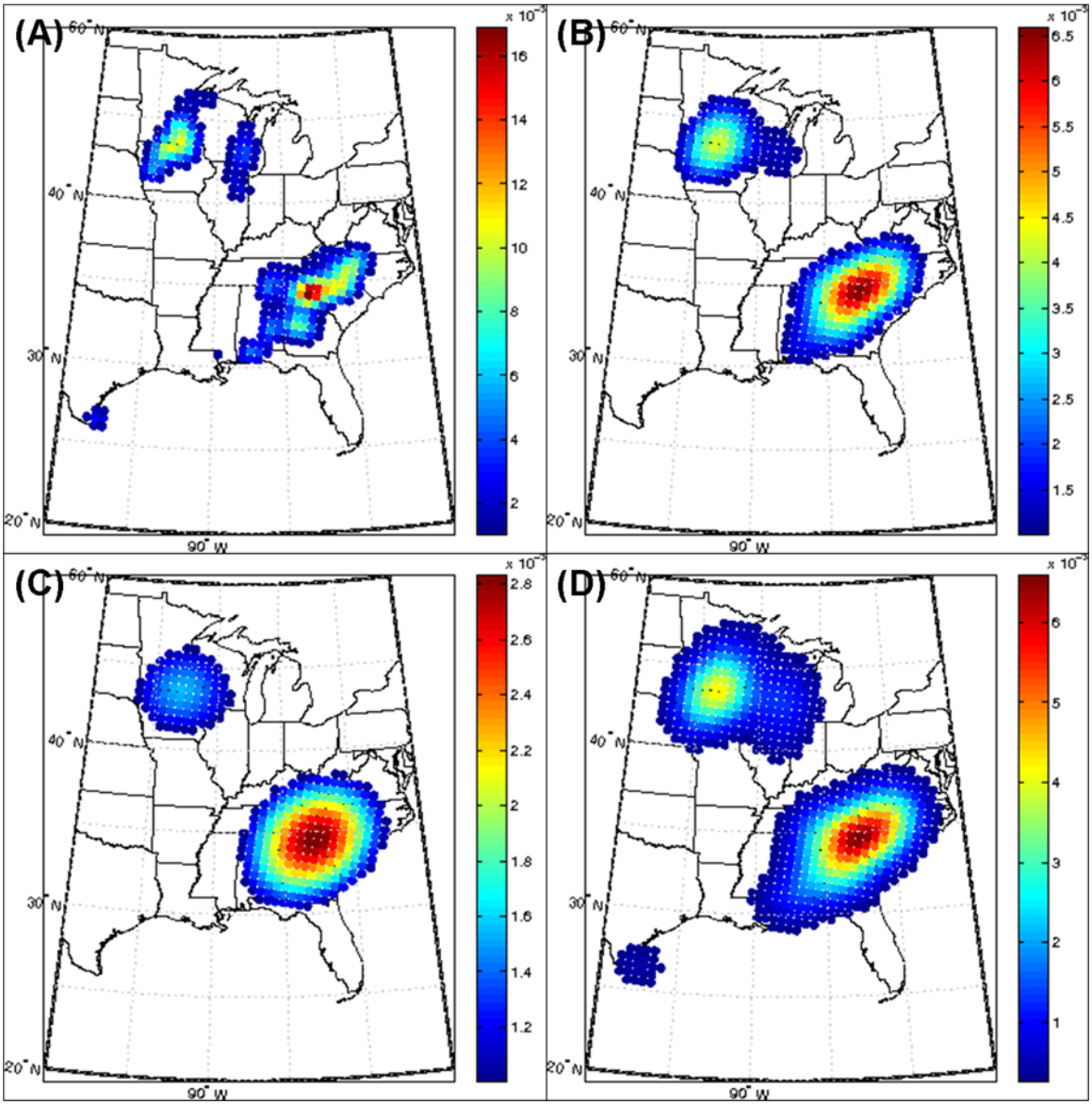

Figure 8: As in Fig. 6, using the grid point method and a Lambert conformal projection with 54-km grid spacing, with a bandwidth of a) 1, b) 2.5, and c) 5. Shading begins with a threshold of 0.001. d) As in (b), shading beginning with a threshold of 0.00025. *Click image to enlarge.*

the latitude-longitude projection, the results were markedly different (cf. Figs. 6b and 8a). These differences were 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. 8d).

Using relatively high bandwidths for the Lambert conformal projection resulted in characteristic shapes that did not match the reports well (cf. Figs. 5a and 8c), indicating 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. 6b and 8b), but the PDF threshold of 0.001 did not capture the reports in Deep South Texas. Lowering the threshold to 0.00025 (Fig. 8d) 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

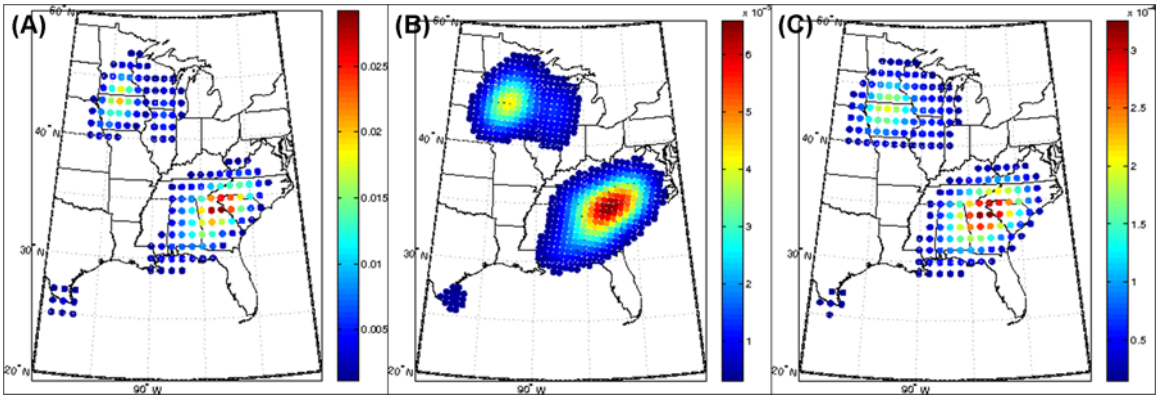

Figure 9: Scatter plots showing the grid points that exceed a specified probability density function threshold, using KDE of the severe reports from 1200 UTC 29 April 1991 to 1159 UTC 30 April 1991. a) The grid point method (see relevant text) is used, with a latitude-longitude projection with 1º grid spacing, a bandwidth of 1, and a PDF threshold of 0.001. 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. c) The distance method (see relevant text) is used, with a latitude-longitude projection, 1º grid spacing, a bandwidth of 150 km, and a PDF threshold of 0.000014. *Click image to enlarge.*

are similar to the changes in grid spacing between the two map projections. A 1º 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 when using the grid point method versus the distance method are also quite minor (Fig. 9). *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 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. 7a,b). The results generally are reasonably robust to modifications of the bandwidth and PDF threshold selections, as long as these selections avoid the tendencies shown in Figs. 6a,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., see Figs. 7c,d).

Any cluster in which the number of reports within the region was less than the detrended annual mean for a cluster was removed from consideration (refer to Fig. 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

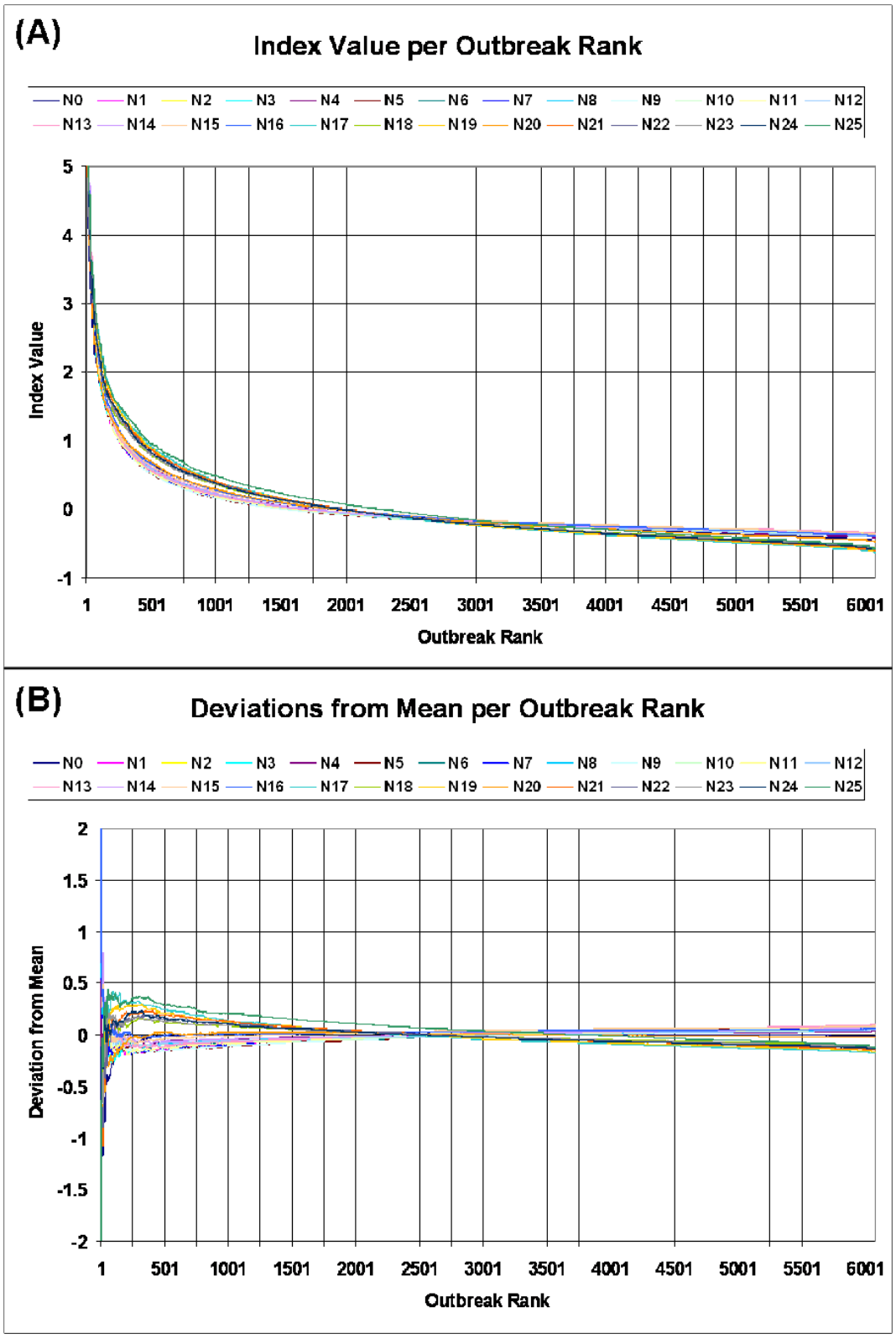

Figure 10: a) The index scores (*y*-axis) and rankings (*x*-axis) of each of the 6072 cases for each of the indices in the study (labeled). b) The deviations (*y*-axis) of each of the indices (labeled) from the mean score of all the indices for a particular rank (*x*-axis). *Click image to enlarge.*

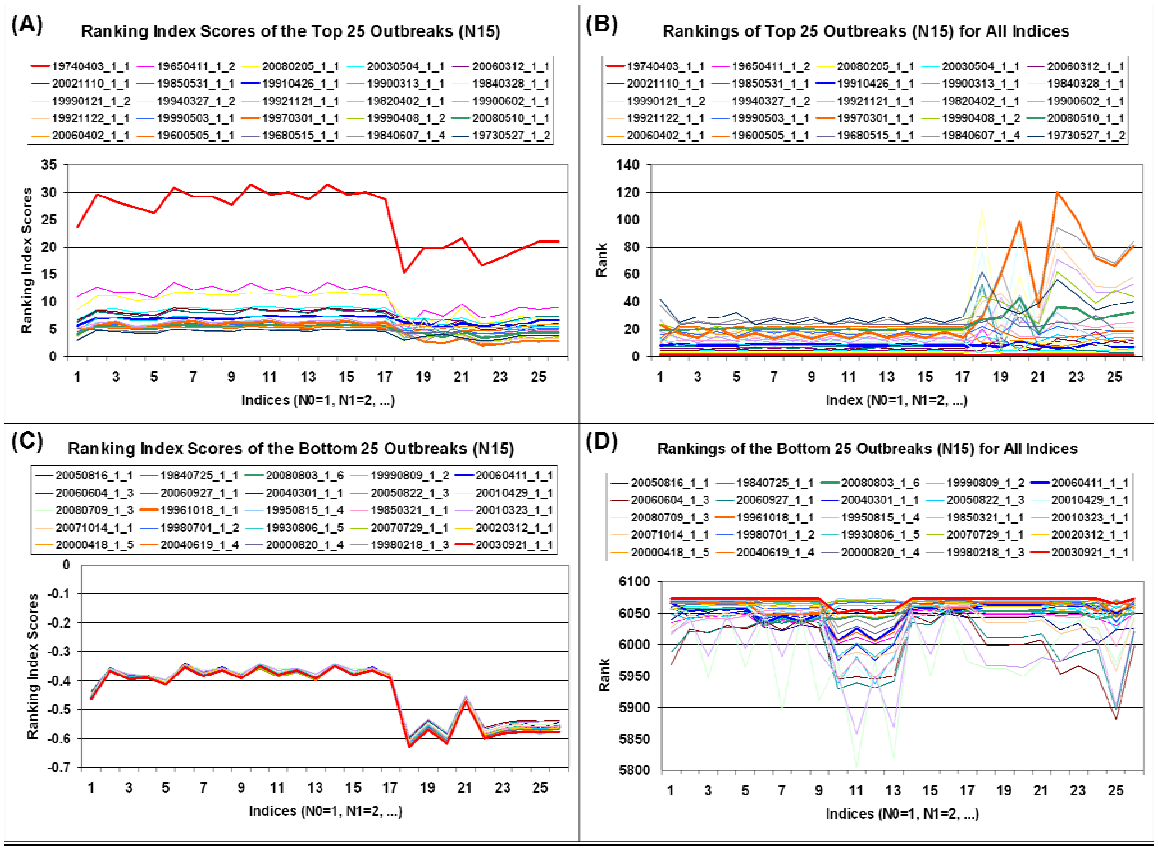

Figure 11: a) The ranking index scores (*y*-axis) for each of the 26 indices (*x*-axis; N0=1, N1=2, etc.) for the top 25 outbreaks based on the rankings of the N15 index. b) The rankings for each of the 26 indices [as indicated in (a)] of the top 25 outbreaks based on the ranking of the N15 index. c) 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. *Click image to enlarge.*

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.

4. 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. 4). Each variable (detrended or otherwise) was standardized as in Eqs. (5)–(7), and the scores for each cluster were computed as in Eq. (8). The relative weights of the variables were altered to develop 26 indices, the weights being equivalent to those of SD10 (their Fig. 4), with the same notation. As Section 2 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 SD10, 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.[2] These are 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 are 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

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

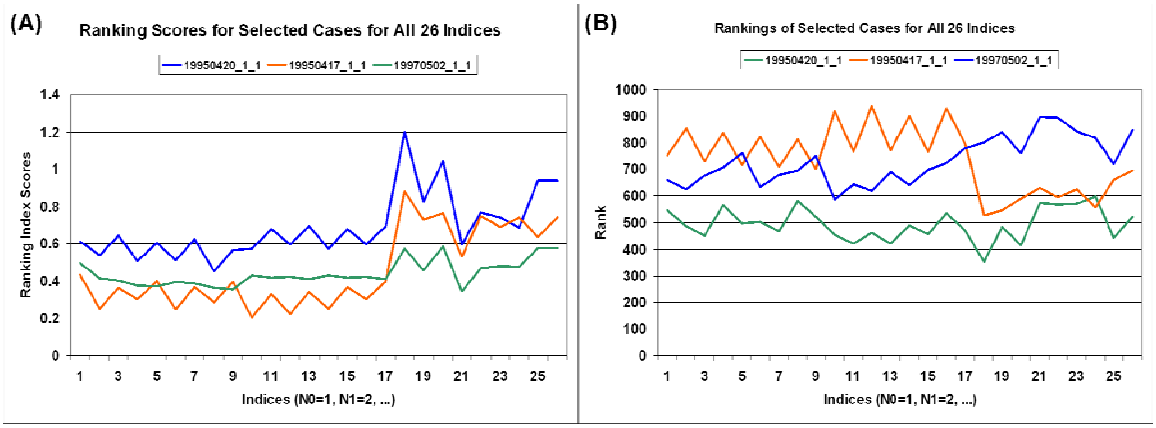

Figure 12: a) As in Fig. 11a, for three specific events (17 April 1995, 20 April 1995, 2 May 1997) as referenced in the text. b) As in Fig. 11b, for the cases in (a). *Click image to enlarge.*

counteract a negative bias for severe report clusters with a large number of significant nontornadic events, and a determination of 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 SD10 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 (plotted as a function of rank in Fig. 10). There were three main findings: (1) The highest-scored (approximately 250) cases have a very steep negative slope (when plotted as a function of rank), analogous to the first ~200 cases in SD10 (cf. their Fig. 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 SD10). The deviations of the individual index scores from the mean score of all the indices for each rank (Fig. 10b) 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 5000 cases. These tendencies indicated that the rankings of the top cases were reasonably consistent no matter what index was used (e.g., Figs. 11a,b), whereas the rankings were relatively volatile for the lower cases (e.g., Fig. 12—see also SD10 for more details). (2) None of the curves exhibit a second steep negatively sloped section analogous to the final ~200 cases in SD10. 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 comprises the two-tornado indices. These differences are more noticeable than in SD10, and are likely a result of the strong correlations among the tornado variables (SD10; 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 subset of the tornado variables from the indices can affect the rankings of the cases noticeably. This is confirmed from inspection of Figs. 11b and 12.

As an example of this last point, consider the exclusion of the 1 March 1997 tornado outbreak (the orange bold curve in Figs. 11a,b; reports in Fig. 13a) from the top 25 outbreaks in the two-tornado indices (N17–N19 and N21–N25) and its presence in the remaining indices. This is a result of the relative lack of nontornadic reports on that day. Some other cases exhibit

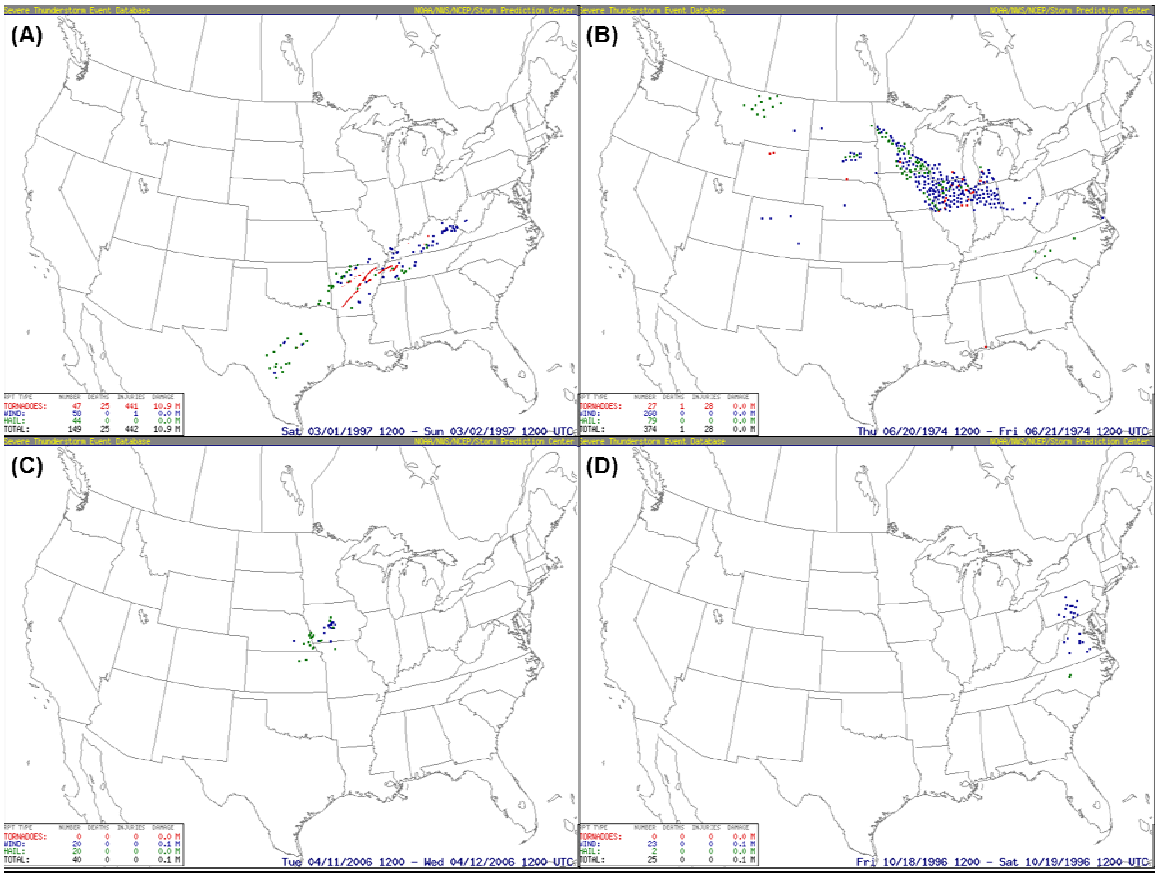

Figure 13: As in Fig. 2a, for a) 1 March 1997, b) 20 June 1974, c) 11 April 2006, and d) 18 October 1996. *Click image to enlarge.*

this behavior (e.g., 15 November 2005; not shown), indicating a potential drawback of removing some tornado variables from consideration. These days commonly were not considered in SD10 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 in this study 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. 11; reports in Fig. 13b) in the top 25 outbreaks of the two-tornado variables 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 to be a desirable characteristic of the variables including few tornado variables; however, this comes at the cost of lower ranks for cases with a large number of strong tornadoes and relatively few nontornadic reports (e.g., the N25 index places 1 March 1997 as 81st). The preferred group of indices is dependent on research goals. If the task is to identify tornado outbreaks and discriminate from all other events, the indices with 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, tornado outbreaks are dominant for the highest-ranked cases without regard to which index is used, another result considered desirable in this study.

The volatility of rankings increases for cases below the steep portion of the curves (Fig. 12). For the 20 April 1995, 17 April 1995, and 2 May 1997 severe weather events (SD10; 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 SD10, 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 SD10, 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. 13c,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 of similar events prior to this time. As noted earlier, nonmeteorological artifacts have not been removed completely.

5. 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 SD10, 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 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 SD10, 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 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) for the N3 index (Fig. 14). 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 portion of the cases in this cluster (Fig. 14a), significant severe weather of any type was included (Fig. 14b). The 3 April 1974 tornado outbreak is an outlier in Fig. 14a (reports in Fig. 15a). The 21 April 1996 hail-dominant outbreak (Fig. 15b) and the 1 July 1994 and 30 May 1998 wind-dominant outbreaks (Figs. 15c,d) are distinct outliers in the four-dimensional decomposition as well. The 20 June 1974 outbreak (refer to Fig. 13b) is a noticeable outlier when the miscellaneous component is analyzed (Fig. 14b), 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. 14c). The major tornado outbreaks (red; 57 cases), hail-dominant outbreaks (green; 887 cases), and wind-dominant outbreaks (blue; 340 cases) are similar to the cluster analysis findings in SD10 (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. 14a and 15 are in the classes one would expect in the 4-cluster analysis.

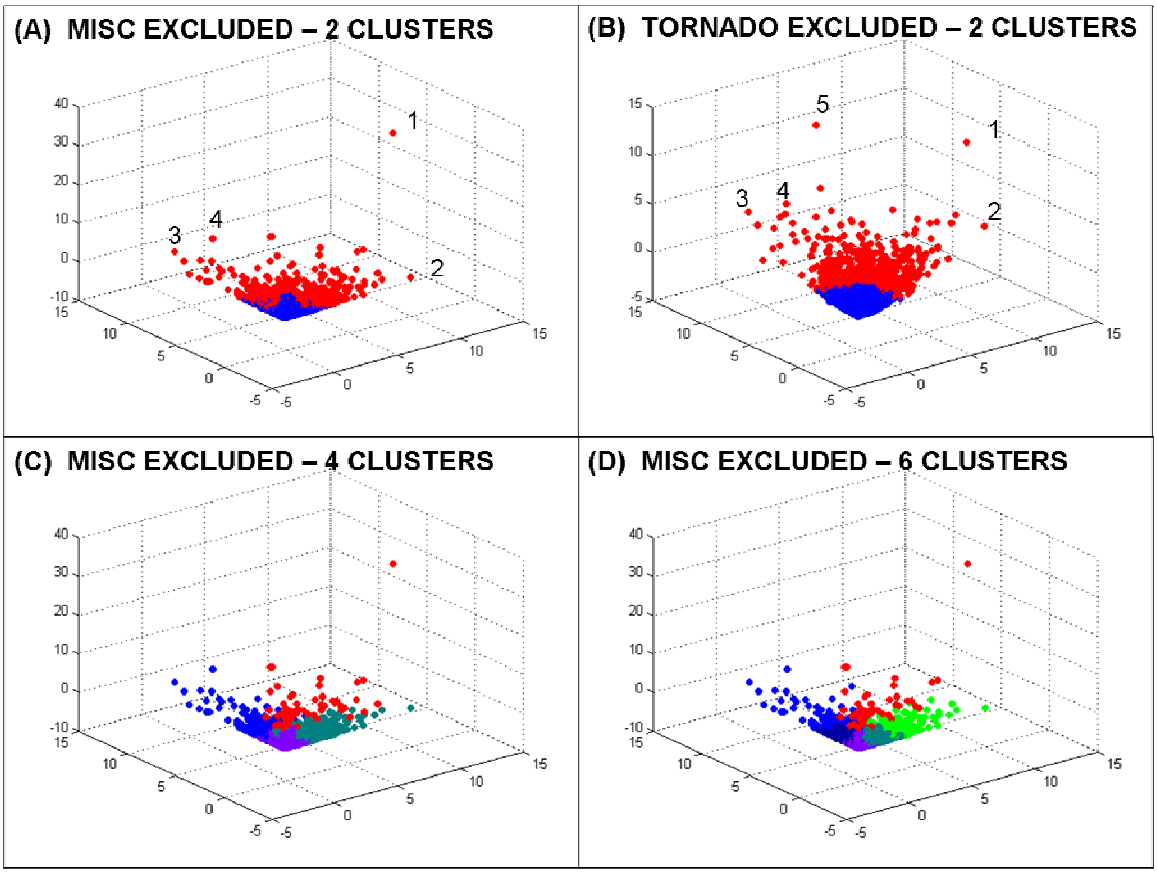

Figure 14: Clusters obtained using the four-dimensional decomposition of the N3 index and k-means cluster analysis. Clusters identified by color, and excluded components of the analysis are labeled. Cases identified include; 1) 3 April 1974, 2) 21 April 1996, 3) 1 July 1994, 4) 30 May 1998, and 5) 20 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. *Click image to enlarge.*

The 6-cluster analysis separates the hail-dominant and wind-dominant groups into two classes each (Fig. 14d). The major hail (wind) events, in which a large number of hail (wind) reports and/or a large number of significant hail (wind) reports were observed, are separated from the relatively minor events. 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. 10a.

The *k*-means 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 (Fig. 16). However, this technique appeared to group major tornado outbreaks and significant wind events (primarily derechos; see Johns and Hirt 1987) together (cf. Figs. 14c and 16c), 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 somewhat similar for the *k*-means and Ward's techniques (cf. Figs. 14d and 16d),

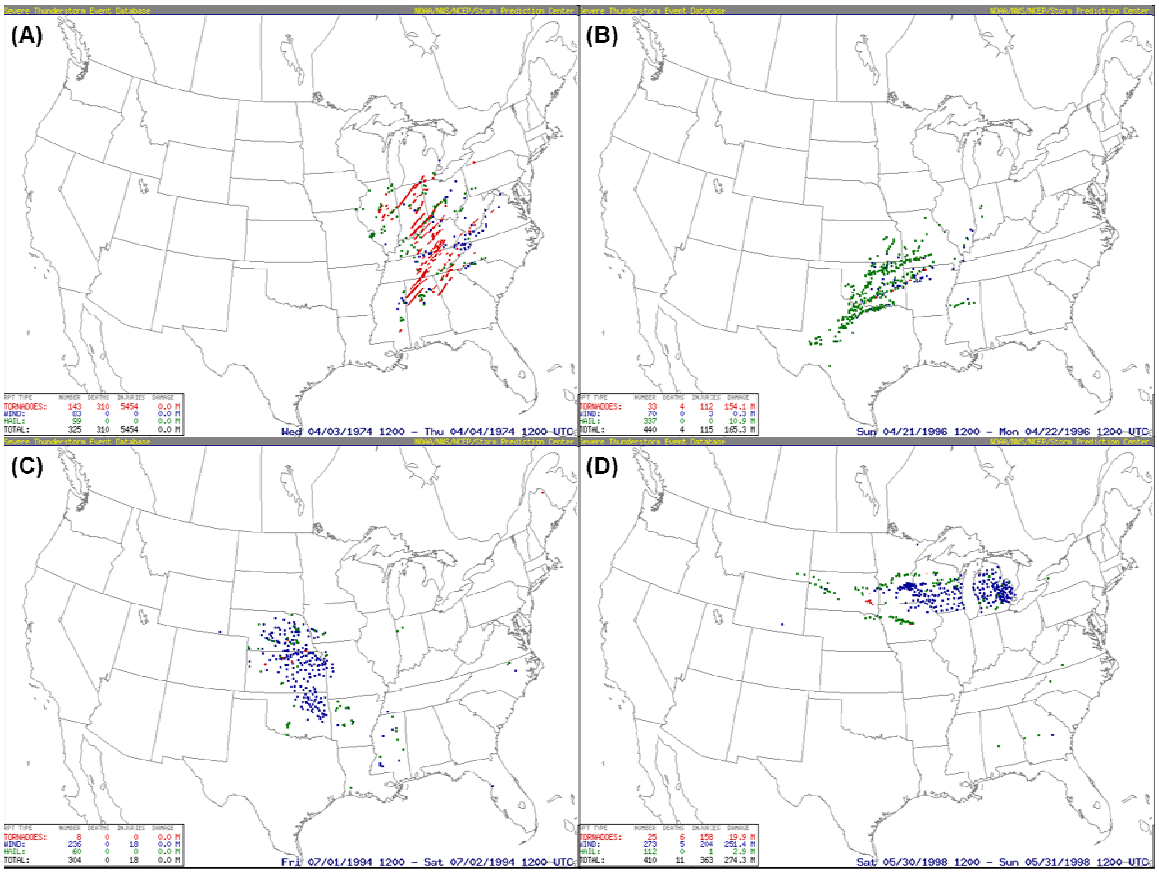

Figure 15: As in Fig. 2a, for a) 3 April 1974, b) 21 April 1996, c) 1 July 1994, and d) 30 May 1998. *Click image to enlarge.*

though the minor hail-dominant and wind-dominant classes for each technique are quite different. 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 our purposes. 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.

Cluster analysis also was performed on the total one-dimensional scores, for guidance on possible categorization of severe weather events based on relative severity. The average scores were taken of all of the indices, for each rank from 1 to 6072 (the same computation used to create Fig. 10b). Ward's hierarchical and *k*-means cluster analyses then 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. In other words, the clusters of the Ward's technique indicated the characteristics of the indices rather than of the outbreaks.

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

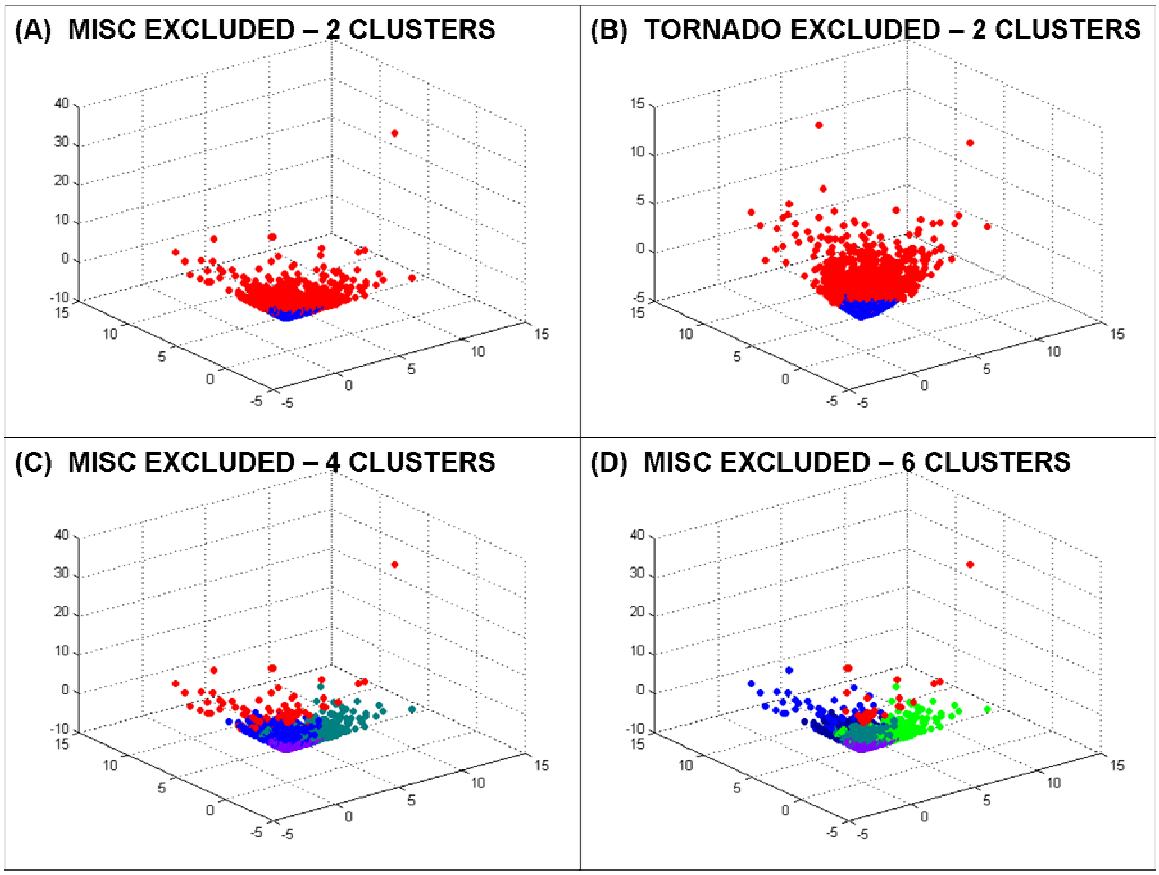

Figure 16: As in Fig. 14, using the Ward's hierarchical clustering technique. *Click image to enlarge.*

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.

6. Summary and conclusions

This study is a follow-up to that of SD10, which presents an innovative technique to account for cases with large geographic scatter or multiple clusters of regionally separated severe weather reports, with the goal of developing a way to rank and classify severe weather events of any type. The technique proposed uses KDE to identify regions associated with a particular cluster of severe reports, rather than 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 SD10.

We have shown that 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

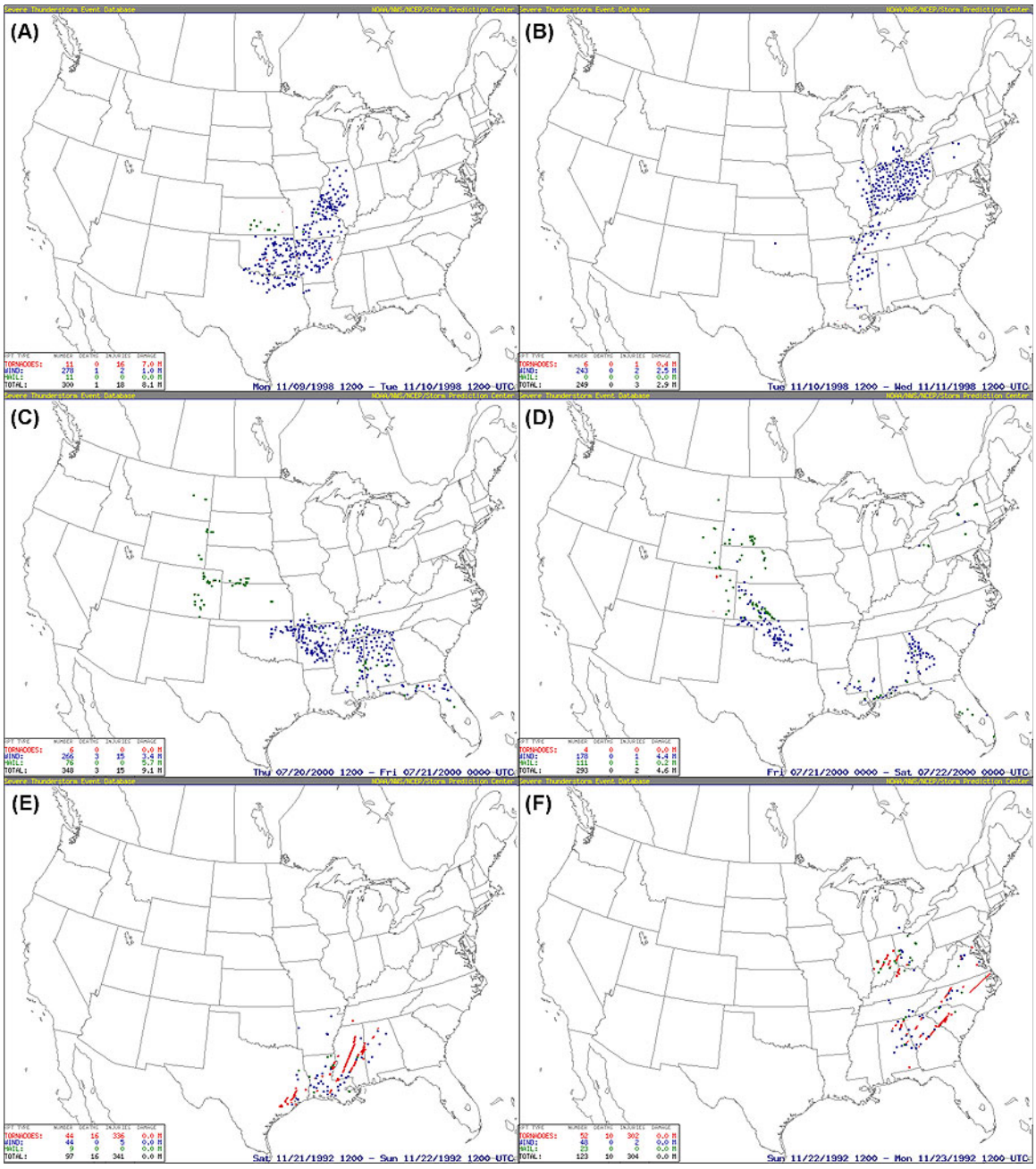

Figure 17: As in Fig. 2a, for a) 9 November 1998, b) 10 November 1998, c) 1200 UTC 20 July 2000 to 0000 UTC 21 July 2000, d) 0000-1200 UTC 21 July 2000, e) 11 November 1992, and f) 12 November 1992. *Click image to enlarge.*

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 SD10. The results were also similar, as major tornado outbreaks were the highest-ranked events. Up to 250 severe weather events appeared to be meteorologically distinct from the remaining cases, as evident in a very steep slope for the scores of the indices for these cases (versus a gradual slope for the remainder). The rankings for the highest-ranked cases were relatively

similar regardless of the index, though important modifications were observed when a subset of the tornado variables was removed from the indices. 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 SD10, 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 regardless of category, the 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 SD10. 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, also would 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 SD10. 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 KDE 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 technique 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. 17a,b). The current method would identify such a circumstance as two separate events and likely would underestimate the severity of the event. Though examples are quite rare in the dataset, future work is planned to try to account for such events.

Furthermore, multiple events can occur in the same region within a 24-hr period (e.g., Figs. 17c,d). The current method would consider this situation a single event, overestimating its severity. These two limitations suggest that a time dimension should be added to the density estimation method; however, its inclusion presents challenges that require substantial investigation before implementation. Objective identification, ranking, and classification of multi-day events associated with a single synoptic-scale system (e.g., Figs. 17e,f) also are prudent, and would provide a valuable resource for investigating these events. Such work is beyond the scope of our current study, however.

We reiterate that the technique presented in this study is not the only way in which severe weather events (outbreaks) could be identified, ranked, and classified. Alternative techniques in the identification of outbreak events using severe weather reports exist (e.g., the contiguity-enhanced hierarchical *k*-means clustering technique; Lakshmanan et al. 2003). The choices made here 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 we do not claim that our scheme is optimal, the results of our 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.).

ACKNOWLEDGMENTS

Funding was provided by NSF Grant AGS-0831359. SVRPLOT V3.0 was used for plots of the severe reports, from a website maintained by Jared Guyer and John Hart. We gratefully appreciate Kim Elmore's suggestion for the formatting of Figs. 11 and 12. Valliappa Lakshmanan and Kim Elmore provided thought-provoking reviews that led to an improved manuscript.

REFERENCES

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

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.

Brooks, H. E., M. P. Kay, and J. A. Hart, 1998: Objective limits on forecasting skill for rare events. Preprints, *19th Conf. on Severe Local Storms*, Minneapolis, MN, Amer. Meteor. Soc., 552–555.

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

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.

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

——, 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*, **21**, 939–951.

Glickman, T. S. Ed., 2000: *Glossary of Meteorology*, 2nd ed. Amer. Meteor. Soc., 855 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.

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.]

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

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

Lakshmanan, V., R. Rabin, and V. DeBrunner, 2003: Multiscale storm identification and forecast. *Atmos. Res.*, **67**, 367-380.

Mesinger F., and Coauthors, 2006: North American regional reanalysis. *Bull. Amer. Meteor. Soc.*, **87**, 343–360.

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

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

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

Thompson, R. L., R. Edwards, 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.

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.

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

REVIEWER COMMENTS

[Authors' responses in *blue italics.*]

REVIEWER A (Kimberly L. Elmore):

Initial Review:

Recommendation: Accept with minor revision.

General Comment: I have carefully reviewed the paper "Using kernel density estimation to identify, rank, and classify severe weather outbreak events." While not a ground-breaking work, it is a useful example of ways to apply kernel density estimation. It should be accepted with minor revisions.

Substantive Comments: Most of the comments presented here are general as I have few specific comments to make. Overall, the paper presents how to perform and use a kernel density estimate (KDE) of a discrete probability density function in two dimensions. The authors have worked out KDE bandwidths and PDF thresholds that depict what they are after and have shown by example that, within limits, there is no difference between performing the KDE on a latitude/longitude grid and on a regular grid of roughly the same spacing as a 1° latitude/longitude grid.

This application of multidimensional KDE is neither new nor particularly innovative, as it is mentioned in Silverman (1986) and also in Kaluzny, et al. (1998) and Venables and Ripley (2002). What the authors characterize in their method is referred to as the "intensity" (λ) of spatial point patterns or processes (SPPs) in some spatial statistics texts and is considered a first-order property. It is usually thought of as the number of points per unit area. There are several ways to analyze SPP intensity, including two-dimensional Gaussian KDE, as done in this work. Two-dimensional smoothed histograms are also used as are other kernel weighting functions. All tend to produce similar results.

There are also ways to analyze second order processes within SPPs, which is roughly the expected number of points within a distance *d* from any point in the pattern. Ripley's K function (Kaluzny et al. 1993), or a scaled version of it, is usually used for this. The second-order measure doesn't define the structure, but does indicate if non-random structure is present. Since it is obvious by inspection that the SPP is *not* random and isotropic, such an analysis serves no purpose here.

In fact, based on work by Barnes (1964), the method introduced here has actually been incorporated (to some degree and in various ways) for many decades. The technique of KDE is definitely not new.

One theme of this work is that using KDE sufficiently works for our intended purposes. We do not claim that KDE is the only method to identify outbreak events via the location and clustering of the reports (and we state so explicitly in the text – see the last paragraph of Section 6, e.g.), but we strongly suspect that alternative methods will not provide substantially improved results. Of course, part of the problem is that it is not entirely clear what "substantially improved results" would look like, based on the uncertainties inundating the severe weather reports archive (see Shafer and Doswell 2010). However, as the cases considered agree with subjective notions regarding which events are severe weather outbreaks, and the regions identified by the KDE technique appear to correspond to the locations of the reports, we believe alternative analyses provide little if any benefit to our presentation.

Overall, the text size needs to be increased on all figures. Since this is generally the case, I won't call out each figure for that separately.

This is an artifact of the plots being reduced to the size of the pages. As EJSSM allows images to be enlarged to their initial size, this should not be a problem in the final version. If the reviewer would like to see these figures in their initial size, we will be happy to provide these.

The authors go on to generate a cluster analysis derived from N-scores. Clusters are intended to represent particular types of events, such as hail-dominant, wind-dominant, etc. The cluster analysis appears to work well, but may be more complex than necessary. This appears to be the case in [then-] Fig. 15, where the clustering is based on the mean of all available N-scores. This seems to mimic a simple rank threshold. The clustering appears almost "too good" because there is no mixing at the margins of the clusters. In effect, the clusters are "perfect" regardless of how many are chosen, a state rarely seen in cluster analysis.

The reviewer makes a good point regarding the N-score vs. rank clustering. We have no problem removing most of this portion of the manuscript, as the results do not appear to tell us much about the nature of the relative severity of the outbreaks (see below).

This is not true, however, about the four-dimensional ranking index decomposition.

I am also concerned about the nature of the function used in the cluster analysis (N-score vs. rank), especially for the very steep part of the N-score vs. rank plots. There seems to be a very high inter-cluster variability in the clusters representing the highest ranks. What's more, the N-score values beyond the first 250 highest ranked cases are essentially constant. How is it that a meaningful distinction can be drawn between event types in the orange and red segments in [then-] Fig. 15d?

These are excellent points, and we did not clarify these enough in our original draft. The basic result is that there is not much guidance provided by the cluster analysis on grouping outbreaks by their relative severity. What the Ward's technique was showing was essentially where the scores (as a function of rank) for the two types of ranking indices (all-tornado vs. two-tornado indices) intersected—meaningful only because of the characteristics of the indices themselves and not the outbreaks. The k-means technique was showing the high inter-cluster variability the reviewer refers to. As neither technique really provides any guidance as to obvious distinctions between major severe weather outbreaks and less significant events (and subgroups within), we are inclined to remove much of this portion of the manuscript, especially given the reviewer's concerns. We have only included the points discussed in our response, and the need for some type of distinction from an operational standpoint.

References

Kaluzny, S. P., S. C. Vega, T. P. Cardoso and A. A. Shelly, 1998: *S+ Spatial Stats User's Manual for Windows and Unix.* Springer, 327 pp.

Silverman, B. W., 1986: *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, 175 pp.

Venables, W. N. and B. D. Ripley, 2002: *Modern Applied Statistics with S.* Springer, 495 pp.

[Minor comments omitted...]

Second review:

Recommendation: Accept.

General Comments: The corrections are entirely satisfactory and I have no additional comments.

REVIEWER B (Valliappa Lakshmanan):

Initial Review:

Reviewer recommendation: Revisions required.

Substantive Comments: The authors introduce a technique to cluster severe weather reports in order to rank outbreak events by cluster instead of (as in their earlier work) by a 24-hour period. To cluster the

reports, they smooth the reports using a 2D Gaussian and then find contiguous regions within the resulting gridded field. The paper is clear, technically correct and concise. Therefore, I have only minor comments:

[Editor's Note: The following review points (and the authors' responses) are rather substantive and are included here.]

Equation 3 is probably better written in terms of x and y since this is a 2D spatial Gaussian. I think that the vector notation just serves to obscure the basic point.

Good suggestion. We have included this as well as the vector term.

I disagree with the footnote on page 5 and the corresponding text on page 8. You should not have to change *both* the variance of the Gaussian and the probability threshold. You can arbitrarily set the threshold and find a sigma that achieves the separation that you need. There are only two degrees of freedom here (one if you set sigma_x = sigma_y).

We think the reviewer's point is valid but more an issue of semantics. In our approach to this problem, we did not arbitrarily select any values (PDF thresholds) beforehand. We did know, however, that only a range of Gaussian thresholds would work for our purposes, because of the well-known bias-vs.-variance tradeoff associated with kernel density estimation. We needed to test multiple variances in order to determine if smoothing was too strong or too weak. The best PDF threshold then is determined once we select the variance of the Gaussian that features the characteristics that agree most with our subjective notions (i.e., smooth regions that do not combine two geographically separate clusters—Figs. 6c,d provide examples of "overfitting" and too much smoothing, respectively).

We note in the text that the same PDF threshold would not work with any value of Gaussian variance because the PDF is a function of the variance. Our example is a comparison of Figs. 6c,d. Note that in Fig. 6c, we have selected a bandwidth of 1.5. The PDF threshold that agrees most with our subjective criteria for usage is the outermost contour (0.001) —as the area within it encompasses almost all of the reports associated with the event without combining geographically separate clusters. However, using a bandwidth of 2, the PDF threshold that agrees with our subjective criteria becomes ~0.005, because lower thresholds had a tendency to join regionally separate clusters of events. Now, in the latter case, we discounted the bandwidth of 2 and the PDF threshold of 0.005 because these regions typically did not encompass all of the reports associated with the outbreak cluster.

This latter point is critical. If we select a bandwidth that contains too much smoothing (as in Fig. 6d), there is no PDF threshold we could select that works sufficiently for our purposes. Indeed, no PDF threshold between 0.001 and 0.005 encompasses the reports in the manner that certain PDF thresholds at lower Gaussian variances do. Similarly, selecting a bandwidth that is too small results in too little smoothing, and no selected PDF threshold works adequately. Because there is a finite range of Gaussian variances that work for our purposes, that means there is also a finite range of PDF thresholds (based on the function relating the PDF threshold and bandwidth). Thus, we cannot (or at least should not) arbitrarily set a PDF threshold beforehand. This is why we state that the PDF threshold has to be modified based on the selection of bandwidth. Because one is a function of the other, modifying one means modifying the other.

Finally, we note that f(x) is not a unitless value. If we use the distance method described in the text, the units of f(x) depend on the dimensionality of the KDE by (1/h)d, where h is the bandwidth and d is the number of dimensions. In two dimensions, d = 2 and h is measured in distance (km) such that f(x) has units of $(km)^{-2}$. Based on the choice of grid point method or distance method, the units of f(x) will be different. This affects the values of PDF threshold that are selected when using alternative map projections.

I've done work before with contiguity-enhanced hierarchical k-means clustering, for the purpose of identifying storms (Lakshmanan et al. 2003). Wouldn't a clustering approach like that also work here? You could explicitly specify the desired intercluster distance in kilometers and let the clustering technique

find the clusters based on the intuitive parameter. [That] would save you a lot of the work in finding sigmas and PDF threshold based on an implicit assumption of intercluster distance.

We see no reason why this would not work, and it may indeed be a more efficient means of doing so. We are by no means suggesting that the method shown in our paper is the only method or even the best method of identifying, ranking, and classifying outbreak events (and we state this explicitly in the text). This is perhaps an alternative approach that could be investigated in future work. In fact, we ***strongly encourage*** *attempting alternative methods. We do believe, however, that the results shown in our paper meet the goals we set out to accomplish.*

We have added some text in the final section regarding this proposed alternative and have included the reference you have provided.

Lakshmanan, V., R. Rabin, and V. DeBrunner, 2003: Multiscale storm identification and forecast. *J. Atm. Res*., **67**, pp. 367-380.

Second review:

Recommendation: Accept.

General Comments: The paper's fine; I'm satisfied with the authors' edits.