Shafer, C. M., C. A. Doswell III, L. M. Leslie, and M. B. Richman, 2010: On the use of areal coverage of parameters favorable for severe weather to discriminate major outbreaks. *Electronic J. Severe Storms Meteor.,* **5** (7), 1–44.

Electronic Journal of
SEVERE STORMS METEOROLOGY

On the Use of Areal Coverage of Parameters Favorable for Severe Weather to Discriminate Major Outbreaks

CHAD M. SHAFER[1,2], CHARLES A. DOSWELL III[2], LANCE M. LESLIE[1,2], AND MICHAEL B. RICHMAN[1,2]

[1]*School of Meteorology, University of Oklahoma, Norman, Oklahoma*
[2]*Cooperative Institute for Mesoscale Meteorological Studies, Norman, Oklahoma*

(Submitted 1 July 2010; in final form 31 December 2010)

ABSTRACT

Forecaster perceptions of major convective outbreaks include the notion that these events occur within relatively large regions of meteorological conditions favorable for the development of significant severe weather, particularly tornadoes. With recent studies developing a rigorous and scientifically repeatable method of identifying these events and distinguishing them from intermediate or marginal severe weather outbreaks, the investigation of a large sample of these events now is possible. This diagnostic study aims to determine the extent to which the use of areal coverage is successful as a means of discriminating major severe weather outbreaks (primarily but not exclusively major tornado outbreaks) from the less significant outbreaks. Preliminary findings suggest that the areal coverage of severe weather parameters favorable for severe weather indeed is associated with the severity of outbreaks. However, the method produces a substantial number of less significant outbreaks that are misclassified as major severe weather outbreaks. Many of these false alarms can be identified by the presence of synoptic environments that are less favorable for the development of a large number of tornadoes. However, a substantial number of intermediate and marginal severe weather outbreaks feature synoptic patterns and mesoscale environments that are difficult to differentiate from the major events. Limitations of using areal coverage as a means of outbreak discrimination are discussed, and refinements to account for these limitations are proposed.

1. Introduction

Predicting the relative severity of convective outbreaks remains one of the primary challenges for operational severe weather forecasters. As operational models are not capable of resolving tornadoes explicitly, and are not expected to do so in the near future, forecasters must rely on the forecast fields of meteorological parameters associated with various types of severe weather (known as covariates; see Brown and Murphy 1996), simulated convection and forecast convective mode from high-resolution model simulations, short-term forecasting based on current observations, and experience with subjective perceptions of similar past events. Many studies have investigated the utility of a variety of severe weather parameters to distinguish storm modes, significance of severe weather, or types of severe weather (e.g., Davies-Jones et al. 1990; Davies and Johns 1993; Johns et al. 1993; Brooks et al. 1994; Stensrud et al. 1997; Rasmussen and Blanchard 1998; Brooks et al. 2003b; Doswell and Evans 2003; Markowski et al. 2003; Thompson et al. 2003, 2007; Potvin et al. 2010). These studies have focused primarily on storm environments, using observed or model-derived proximity soundings. Notable exceptions include Stensrud et al. (1997; model forecast fields) and Brooks et al. (2003b; reanalysis data), but the focus of these studies has remained primarily on storm environments.

Corresponding author address: Chad Shafer, School of Meteorology, University of Oklahoma
120 David L. Boren Blvd., Suite 5900,
Norman, OK 73072-7307
E-mail: cmshafer@ou.edu

Research on severe convective *outbreaks* mostly has been based on case studies (e.g., Fujita et al. 1970; Fujita 1974; Johns and Hart 1993; Thompson and Edwards 2000; Corfidi et al. 2010). Although much has been learned about the environments in which these outbreaks occur and the challenges of forecasting these events (Johns and Doswell 1992; Doswell et al. 1993; Doswell and Bosart 2001; Moller 2001), the number of studies is surprisingly limited regarding the investigation of using various severe weather parameters in the identification of particular types of outbreaks.

One of the challenges of outbreak discrimination is the identification of prototypical cases. Doswell et al. (2006) proposed an objective scheme to rank tornado outbreak and primarily nontornadic outbreak days using a linear-weighted multivariate index. The definition of a *major outbreak* in this study is based on a variation of this ranking scheme developed by Shafer and Doswell (2010—hereafter, SD10) that has been shown to be consistent with subjective assessment of numerous outbreak events. For a complete description of the outbreak ranking scheme, the limitations of the dataset used for the rankings, and the characteristic cases considered to be major outbreaks and for those considered not to be, the reader is referred to SD10.[1]

Shafer et al. (2009; 2010—hereafter, S09; S10) and Mercer et al. (2009—hereafter, M09) investigated the ability of mesoscale models to discriminate tornadic and primarily nontornadic outbreaks initialized with synoptic-scale data by analyzing forecast meteorological fields. These studies considered a relatively large number of cases, thereby reducing the impact of small sample size limitations of previous outbreak studies (Doswell and Schultz 2006; Doswell 2007).

As S09 and M09 concluded, consistent, skillful discrimination of tornadic and primarily nontornadic outbreaks appears to be possible at least three days in advance of the outbreak. Given the success of these initial investigations, it is now reasonable to consider cases that are not easily classified as either type. Most convective outbreaks feature a mix of severe reports (SD10). Consequently, operational forecasters commonly face the task of determining which outbreak days will be of this "intermediate" type. Major severe weather outbreaks primarily consist of major tornado outbreaks, though a few cases feature a notable lack of tornadoes (high-impact derechos or widespread significant hail events). Intermediate cases feature a small to moderate number of tornadoes and a large number of nontornadic reports, or lower-impact primarily nontornadic outbreaks. Marginal events were also included in SD10, which are events with substantial geographic scatter or multiple regionally-separate clusters of severe reports. The reader is referred to SD10 (their Section 3b) for a more thorough description of these types of events. In this paper, intermediate and marginal outbreaks will be considered as one category (null events). Section 2 discusses how these events are classified in more detail.

To select intermediate outbreak days in a classification study, the work of Doswell et al. (2006) was modified by including any type of outbreak (specifically, the top 30 days of each year from 1960-2006, based on the total number of severe reports) by SD10. The new indices developed by SD10 revealed that a small proportion (~200) of the 1410 cases included in their study could be classified as major severe weather outbreaks (see their Fig. 6), whereas the remaining cases were intermediate or marginal.

There are many possible methods for objectively evaluating the ability of meteorological covariate fields to discriminate the major outbreak days from the intermediate and marginal outbreak days (e.g., see M09). A relatively simple approach is to consider only the areal coverage of several diagnostic variables at the valid times of the outbreaks. Forecaster perceptions of convective outbreaks suggest that the subsynoptic environments on these days are favorable for severe weather over relatively large regions, whereas localized severe weather is associated with a favorable environment only in a relatively small region. S09 illustrated the utility of examining areal coverage in the subjective discrimination of tornadic and nontornadic outbreaks (their Section 3), in which the environments favorable for tornadoes systematically covered a much larger area with tornado outbreaks than for primarily nontornadic outbreaks.

Previous studies focusing on using covariates to discriminate storm modes or severe weather types show a pronounced "false alarm

[1] This study is based on the work of SD10. As a result, it is recommended that readers consult SD10 before this study.

problem" (e.g., Rasmussen and Blanchard 1998; Thompson et al. 2003). Specifically, the discrimination of tornadic supercells from nontornadic supercells resulted in a large number of nontornadic supercells being classified as tornadic supercells. Rasmussen and Blanchard (1998) discussed this problem at length, proposing at least three possible reasons. (1) Large-scale factors that exhibit characteristics favorable for tornadoes may be represented adequately, but factors unfavorable for their development may not be. (2) The convective mode often has profound implications on the type of severe weather that is observed and likely does not correspond well with most severe weather parameters. (3) Large-scale conditions almost never represent the storm-scale environment (see Markowski et al. 1998a,b for a discussion). Although these problems have arisen in considering individual storm environments, rather than outbreaks, it is reasonable to expect these possible inhibiting factors to affect discrimination of outbreak types as well. Thus, before investigation of a mesoscale model forecast's ability to discriminate major convective outbreaks from the intermediate and marginal outbreak days, it is worthwhile to determine the ability of analysis data valid at the time the outbreaks were occurring to discriminate outbreak type. This is the focus of the present work.

Section 2 describes the data and methods incorporated in our study, and Section 3 presents the results. Section 4 provides some subjective interpretations of the results, including the limitations inherent in an "areal coverage" approach. Section 5 provides a discussion of the current work's implications and offers potential topics for future investigation.

2. Data and methods

To analyze the mesoscale fields of meteorological parameters at the valid times of a large number of outbreak days, the North American Regional Reanalysis dataset (NARR; Mesinger et al. 2006) was used in this study. The NARR dataset is available from 1 January 1979 to the present. Horizontal grid spacing is 32 km, with 45 vertical layers. These regional reanalysis data showed significant improvement in temperature and wind fields compared to global reanalysis datasets (For more details please see the PowerPoint™ presentation at http://www.ejssm.org/ojs/public/vol5-7/narr.ppt.) and are available for relatively long periods of time compared to other datasets, such as the RUC, making selection of the NARR preferable. The long period of time is critical, as even with a relatively large sample used in this study, sample size issues remain (see Section 3).

The reanalysis data were converted via bilinear interpolation to a 300×200 18-km horizontal grid, with 31 vertical levels, using the Weather Research and Forecasting model's Preprocessing System (WPS) Version 3.1 (Skamarock et al. 2008). This conversion was made for direct comparison with previous and future model simulations of the 18-km domain used for objective analysis of outbreak classification (see S09; M09; S10). The domain covers the contiguous United States and is not shifted for any outbreak day. Each of the 30 outbreak days for every year from 1979 to 2006 considered in SD10 was analyzed for this study. The outbreak days were split randomly into a training set of 630 cases and a testing set of 210 cases.[2] Analysis of the meteorological fields included the same variables as used in previous work (S09; M09; S10).

The total number of grid points in the 300x200 domain that exceeded a threshold value for a particular parameter was used to measure the areal coverage of any parameter considered to be favorable for a major severe weather outbreak. The grid point sums then were compared for each outbreak day (Section 3), using individual and combined fields (that is, univariate and multivariate sums). Additionally, pseudo-trajectories were computed as another means of describing the severe weather environment. These pseudo-trajectories are intended to approximate how long it would take a storm to traverse the area wherein the environment is deemed to be favorable. This method determines backward and forward trajectories at each grid point, using the wind speed and direction at 500 hPa, within the enclosed area in which the parameter analyzed exceeded the predetermined threshold. The 500-hPa wind is acknowledged to be only a crude estimate of storm motion, but it suffices to compare each outbreak to the others.

Several classification algorithms were considered, including linear and quadratic

[2] See Section 3b for details on the selection of the number of training and testing cases.

discriminant analysis (Seber 1984; Krzanowski 1988), decision trees (Breiman et al. 1993), and support vector machines (SVMs; Cristianini and Shawe-Taylor 2000), to determine if any of the algorithms used had a distinct advantage or disadvantage in discriminating cases or if the results were consistent among the algorithms. Three methods of interpreting the statistical results were conducted: (1) The entire training set was used to develop a statistical model, and this model was run independently on the test data. Bootstrap confidence intervals (Efron and Tibshirani 1993) of contingency statistics (Wilks 1995) were computed for the results of the testing data. (2) The training set was converted randomly to 25 subsets, in which 63 (10%) of the 630 training cases were removed in each subset (i.e., 567 cases per subset). Removing a larger proportion of cases led to reduced accuracy and skill in discriminating the remaining cases, whereas removing a smaller proportion of cases led to substantially larger uncertainty of the statistics (not shown). These 25 statistical models were assessed using the testing data, and the contingency statistics were analyzed. Techniques (1) and (2) were conducted to determine the uncertainty in the contingency statistics of the testing data and the variability of the training models for the same test data, respectively. (3) Additional analysis was conducted by adjusting incrementally the grid point sum or the total or mean time, distance, or speed of a storm's pseudo-trajectory for those points exceeding a particular threshold value of a severe weather parameter. This method provides analysis on what grid point thresholds may be most accurate and skillful for outbreak discrimination and permits simple interpretation of the results.

SD10 developed 26 multivariate linear-weighted indices (identified by the labels N0-N25) to rank outbreaks of any type. The indices used 14 different types of severe report variables (directly or derived from the Storm Prediction Center severe weather database; see Schaefer and Edwards 1999; see also SD10, their Table 1). Each severe report variable was converted to standard normal (Eqs. [1]–[3], Doswell et al. 2006), to prevent undue weight being given to the parameters with large magnitudes. Additionally, nonmeteorological artifacts are known to be present in the severe weather reports data (Brooks et al. 2003a; Doswell et al. 2005; Verbout et al. 2006). Although we have accounted for some of this by detrending the data in time (SD10), there is no way to account for all of these artifacts. Among other things, this adds an element of uncertainty to our results, the details of which cannot be known, but there is no perfect database to use for our purposes.

The scores of the 26 indices then were computed by taking the weighted sums of the severe report variables and dividing by the sum of the weights (as in Eq. [4], Doswell et al. 2006). The 26 indices differed by modifying the weights of the severe report variables, in accordance with subjective notions of the relative significance of the various types of reports (see Section 3a of SD10 for more discussion). Indices N0–N16 and N20 featured nonzero weights for all of the variables, whereas indices N17–N19 and N21–N25 featured zero weights for six of the eight tornado report variables. This latter point will be discussed further later in this section.

The characteristic curve of the scores of the indices developed by SD10 is similar to the curve from an uncorrelated series of random numbers with standard normal distributions (Fig. 1; see also section 3a and Fig. 6a of SD10 for more discussion). This is because the indices included 14 variables that were standardized. Based on its definition, the variance of an uncorrelated series of 14 standard normal variables is 14 (whereas the mean remains zero). However, each of these variables was given a weight w_i, and the sum of the weighted variables was divided by the sum of the weights. Thus, each variable x_i had a coefficient a_i, where:

$$a_i = \frac{w_i}{\sum_{i=1}^{14} w_i} \tag{2.1}$$

If we assume equal weights, the coefficient for each variable is (1/14) = 0.0714. Based on the definition of the variance of a series of variables:

$$Var\left(\sum_{i=1}^{n} a_i x_i\right) = \sum_{i=1}^{n} a_i^2 Var(x_i) + 2 \sum_{i,j;i<j} a_i a_j Cov(x_i, x_j) \tag{2.2}$$

For uncorrelated variables, the covariance term in Eq. (2.2) is zero, and from Eq. (2.1), the variance of the linear equally-weighted average of 14 uncorrelated variables is (1/14) = 0.0714.

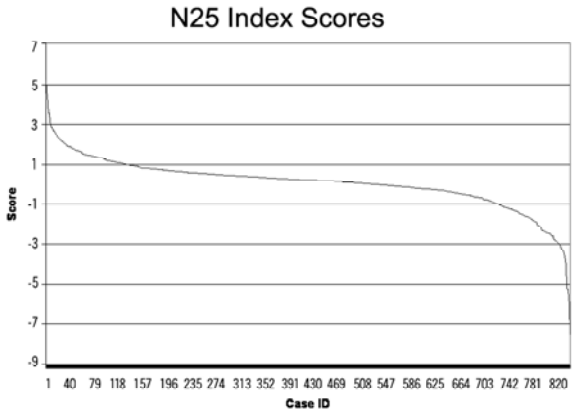


Figure 1: Plot of the values of the N25 index scores for the 840 cases considered in the current study. *Click image to enlarge*.

However, the variables used to determine the outbreak rankings in SD10 are correlated, and generally strongly positively correlated (see their Table 2). As the correlation is the covariance of two variables divided by the product of their individual standard deviations (which is 1, as the variables are standard normal), the covariance term in Eq. (2.2) is positive. Thus, the variances of the index scores are higher than the variance of the equally-weighted uncorrelated series.[3] In general, these variances were near or slightly larger than 1. In SD10, cases above scores of 1 (i.e., the mean plus the standard deviation, approximately) were primarily tornado outbreaks, whereas the cases below this threshold generally were not. This was by design, as the indices weighted the tornado variables highest.

The threshold value of 1 also was supported by a three-class *k*-means cluster analysis (see Gong and Richman 1995) of the multivariate indices (not shown). For example, using the N25 index, the three classes were separated by the scores of 1.10 and -1.15. For the other indices, these values similarly were found to be near 1 and -1. Because of the aforementioned findings and because the cases were not completely rank-invariant (leading to some cases being classified as major outbreaks for some indices and intermediate for others; see SD10), the value of 1 initially was selected to separate *major* outbreaks from *intermediate* or *marginal* outbreak days, where the latter included cases with scores below the value of -1. We by no means are stating that these values are the most appropriate, however. Indeed, selecting various thresholds to examine differences in diagnosing outbreak severity is appropriate. Discussion regarding shifting the index threshold, and the subsequent effects on diagnosing outbreak classification using areal coverage, will be addressed in Section 3b.

The N15 and N25 indices were used for outbreak classification because of the differences in the number of tornado report variables included as weights for the indices.[4] Specifically, N15 incorporates eight variables (the total number of tornadoes, the total number of F2 or greater tornadoes, the total number of F4 or greater tornadoes, the destruction potential index [DPI; Thompson and Vescio 1998], the total path length, the number of killer tornadoes, the total number of fatalities, and the number of long-track tornadoes), whereas N25 only includes the total number of tornadoes and the DPI (SD10, their Fig. 4). Both indices include six additional severe report variables (the total number of all reports, the total number of wind reports, the total number of hail reports, the total number of significant wind reports, the total number of significant hail reports, and the middle-50% parameter—a parameter designed to account for geographic scatter with the reports). The scores for each of the 840 case days (e.g., N25 in Fig. 1) follow the characteristic curves shown in SD10 (their Fig. 6a).

3. Results

a. The need for additional constraints

Initially, the areal coverage of a particular combination of severe weather parameters was computed with no preexisting criteria for a grid point's inclusion in the computation. A case-by- case determination of the areal coverage values for a variety of meteorological parameters (Fig. 2) indicated several important characteristics:

[3] There are other reasons for this increase as well, including unequal magnitudes of the weights for the 14 variables and the treatment of one of the 14 variables separately from the remaining 13 (the middle-50% parameter; see SD10 for more details).

[4] The selection of N15 and N25, besides the differences in the number of tornado report variables used as weights for outbreak ranking, was essentially arbitrary. There is no obvious preference to any of the indices developed by SD10, as their Section 5 discusses.

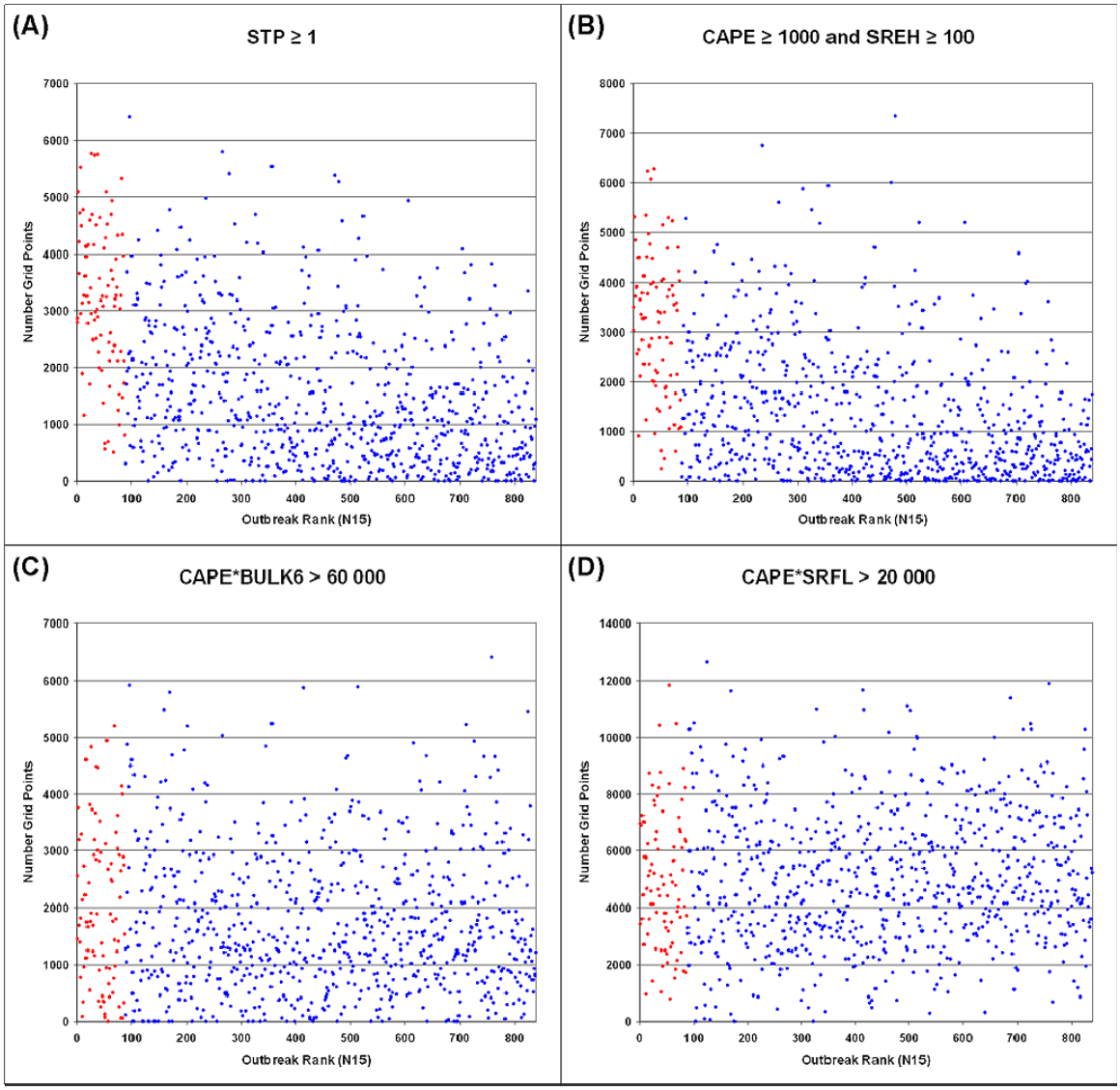


Figure 2: Number of grid points exceeding a) a threshold value of 1 for STP, b) a threshold value of 1000 J kg^{-1} SBCAPE *and* a threshold value of 100 m^2 s^{-2} for 0–1 km SREH, c) a threshold value of 60 000 for the product of SBCAPE (J kg^{-1}) and 0-6 km bulk shear (kts), and d) a threshold value of 20 000 for the product of SBCAPE (J kg^{-1}) and SRFL (m s^{-1}), for each of the 840 cases. Cases ranked in order from 1 to 840 using the N15 index. Major (intermediate or marginal) outbreaks are indicated by red (blue) dots. *Click image to enlarge*.

- The highest-ranked outbreak days tended to have a higher number of grid points exceeding a threshold value for a *certain subset* of parameters. Subjective analysis suggests that the "best" discrimination of major outbreaks from intermediate and marginal outbreak days used either (1) the significant tornado parameter (STP, after Thompson et al. 2003) values ≥1 (Fig. 2a), (2) a combination of surface-based (SB) CAPE ≥1000 J kg^{-1} and 0–1 km storm-relative environmental helicity (SREH) ≥100 m^2 s^{-2} for each grid point (Fig. 2b), or (3) 0-1 km AGL energy helicity index (EHI, after the Rasmussen 2003 version) values ≥1 (not shown). That is, the performance of these covariates was very similar. On the other hand, some areal coverage parameters showed negligible capability distinguishing outbreak days (e.g., the product of SBCAPE and 0-6 km AGL bulk shear (BULK6) exceeding 60 000 (J kg^{-1})(kts), Fig. 2c; or

the product of SBCAPE and ~2 km above ground level storm-relative flow (SRFL) exceeding 20 000 $m^3 s^{-3}$, Fig. 2d). Based on this analysis, the remainder of this paper uses STP for evaluation of the areal coverage technique.[5]

- There is a high level of scatter in the data, suggesting that this method is subject to substantial limitation in determining the relative severity of outbreak days. In particular, a statistical model's ability to predict the index score is likely to be limited.
- Several intermediate and marginal outbreak days are seen to have comparable values of total grid points exceeding a threshold for various severe weather parameters to those of major outbreak days.
- Very few cases exhibited noncontiguous regions of parameters exceeding thresholds. Thus, no effort was made to account for this tendency in subsequent analyses.

Examination of several "false alarm" days showed that the unconstrained initial calculations of favorable areas were susceptible to a number of undesirable characteristics: (1) A large number of water points (e.g., the adjacent Atlantic Ocean, Gulf of Mexico, Great Lakes, etc.) exceeded the specified thresholds. As severe reports are archived only over land, points over water could not be verified to have experienced severe weather, so water points are not considered hereafter. (2) Several severe weather indices [e.g., EHI, the supercell composite parameter (SCP, after Thompson et al. 2003), STP] appeared to be exceptionally high on days with very large values of CAPE. These values were high despite relatively unfavorable values of shear or helicity (e.g., Fig. 3). As a result, only grid points in which SBCAPE $\geq$1000 J kg^{-1} and 0–1 km SREH $\geq$100 $m^2 s^{-2}$ were considered in some subsequent calculations of areal coverage thresholds.[6] (3) On some days with multiple clusters of severe reports or with large geographic scatter in the reports, there were large regions of favorable severe parameters. The additional CAPE and SREH thresholds incorporated as a result of (2) seemed to diminish this problem to a degree, but days with multiple clusters of reports were sometimes unaffected. Moreover, these days were predominantly responsible for any noncontiguous regions of favorable parameters for significant severe weather. As SD10 noted, more rigorous techniques to account for these days are likely necessary when developing an index to rank outbreak days (see their discussion on the "middle-50% parameter"). In Section 5, one possible method to be incorporated in future research is proposed.

Not surprisingly, the incorporation of the additional constraints also contributed to a lowering of the grid point sums for some of the major severe weather outbreaks. *A larger number of "misses" should be expected with any constraint implemented in the inclusion of points exceeding a particular threshold for specified variables.* Moreover, implementing the constrained STP did not reduce adequately the excessive scatter in the values (not shown), as a large number of intermediate and marginal outbreak days maintain values of areal coverage comparable to the major outbreak days. As a result, both constrained and unconstrained calculations of STP are used in subsequent analyses in this paper.

When considering pseudo-trajectories for each point within the area in which the unconstrained STP $\geq$1, we computed the sum of the distances for each grid point storm, the mean distance of the hypothetical storms, the mean speed of storm motion, and the time required for the mean hypothetical storm to enter and exit the favorable region. The results illustrated similar characteristics to those shown in Fig. 2. The lengths of the trajectories (sum and mean) were, in general, larger for major severe weather outbreaks than for intermediate and marginal outbreaks. Storm motions were somewhat faster for major outbreaks, and the time required for a

[5] Other parameters were analyzed, including EHI, and the combination of CAPE and SREH, but results were similar or worse than those of STP. Thus, STP is used throughout the rest of the paper to illustrate the advantages and disadvantages of using the areal coverage technique. However, CAPE and SREH are used as additional constraints in some areal coverage calculations, as this section will later discuss.

[6] Hereafter, any time the water and CAPE/SREH constraints are added to the STP threshold, this will be referred to as the *constrained STP*. If only the water points are eliminated, it will be referred to as the *unconstrained STP*.

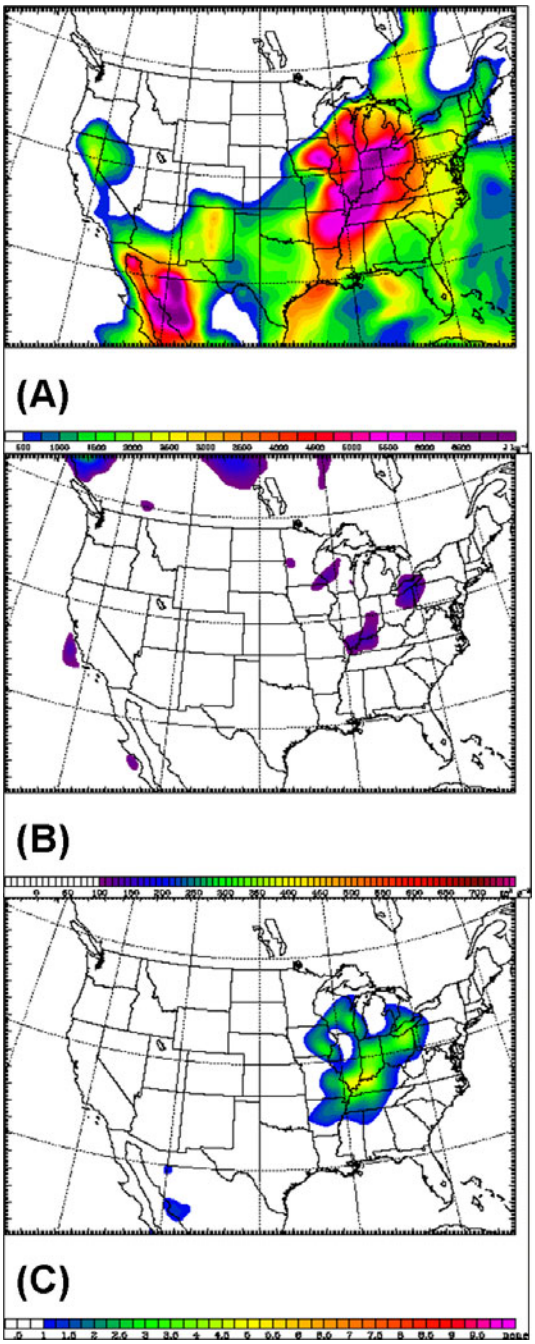


Figure 3: NARR fields of a) SBCAPE (J kg^{-1}), b) 0–1 km AGL SREH (m^2 s^{-2}), and c) 0–1 km AGL EHI, valid at 0000 UTC 30 August 1984. *Click image to enlarge*.

hypothetical storm to traverse the favorable area appeared to be slightly longer for major outbreaks. The storm motion and time of the mean trajectory to traverse the favorable area did not decline substantially, however, as the ranking (severity) of the outbreak decreased. Once again, there was substantial scatter in the results, and there still was a large number of intermediate and marginal cases with similar values to those of major severe weather outbreaks.

b. Results using incremental thresholds

A simple technique to measure the accuracy and skill of using areal coverage as a criterion for outbreak discrimination is to change the thresholds incrementally, from small to large values, and to compute contingency statistics for these thresholds. The following analysis includes all 840 cases. However, later in this section, training and testing sets are created, and more advanced statistical algorithms are incorporated into the analysis. An overview of contingency statistics is provided in Wilks (1995), and the standard binary contingency variables *a* (correct hits), *b* (false alarms), *c* (misses), and *d* (correct nulls) will be used in the following discussion. Equations for the binary statistics employed in the following discussion are:

$$HR = \frac{a+d}{N} \tag{3.1}$$

$$POD = \frac{a}{a+c} \tag{3.2}$$

$$FAR = \frac{b}{a+b} \tag{3.3}$$

$$POFD = \frac{b}{b+d} \tag{3.4}$$

$$SR = 1 - FAR \tag{3.5}$$

$$CSI = \frac{a}{a+b+c} \tag{3.6}$$

$$PSS = POD - POFD \tag{3.7}$$

$$HSS = \frac{(a+d) - E_c}{N - E_c} \tag{3.8}$$

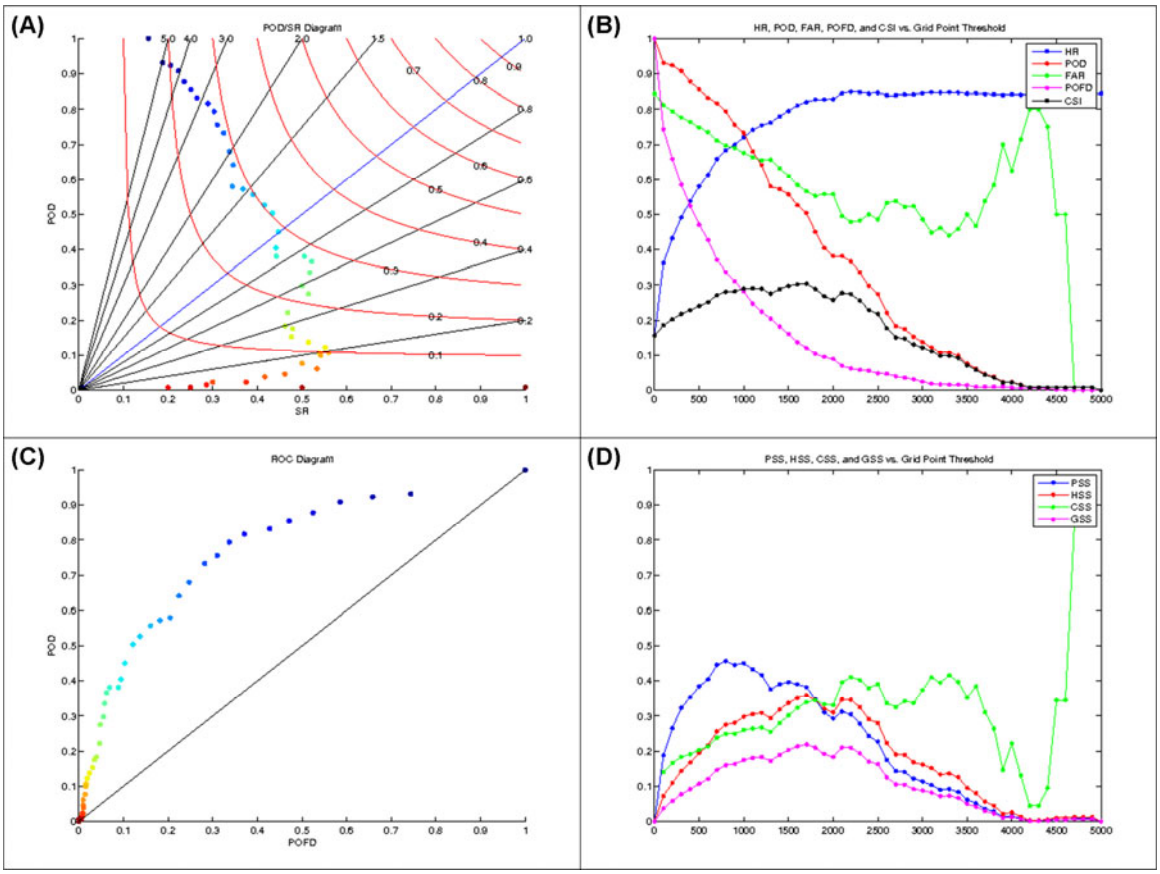


Figure 4: a) Roebber (2009) diagram, using the total number of land grid points for the constrained STP ≥1 for diagnosis of outbreak type and the N25 index for outbreak classification. Bias is shown using diagonal black lines (with the blue line showing a bias of unity). Red curves indicate CSI. Individual points indicate grid point thresholds in increments of 100, starting from 0 (blue) to 5000 (maroon). b) Hit rate (blue), probability of detection (red), false alarm ratio (green), probability of false detection (magenta), and critical success index (black) of the grid point thresholds specified on the *x*-axis. c) Relative operating characteristics (ROC) diagram for each of the grid point thresholds, as in (a). d) Pierce (blue), Heidke (red), Clayton (green), and Gilbert (magenta) skill scores for each of the thresholds in (b). *Click image to enlarge*.

$$CSS = \frac{a}{a+b} - \frac{c}{c+d} \tag{3.9}$$

$$GSS = \frac{a - \dfrac{(a+c)(a+b)}{N}}{a+b+c-\dfrac{(a+c)(a+b)}{N}}, \tag{3.10}$$

where N is the sum of each of the components of the contingency table and

$$E_c = \frac{1}{N}((a+c)(a+b)+(c+d)(b+d)). \tag{3.11}$$

The preceding equations are the hit rate (HR), probability of detection (POD), false alarm ratio (FAR), success ratio (SR), probability of false detection (POFD), critical success index (CSI), Pierce skill score (PSS), Heidke skill score (HSS), Clayton skill score (CSS), and Gilbert skill score (GSS). These four skill statistics were chosen because of their different properties in rare-events datasets (e.g., Doswell et al. 1990), the inclusion or exclusion of the "correct null" category (i.e., GSS—see Murphy 1996), the property observed by Richardson (2000) that the maximum value of a 2×2 decision problem is given by the PSS for a complete range of users, and the property observed by Wandishin and Brooks (2002) that the CSS can pinpoint when forecasts no longer have value.

As an example of the proposed technique, the total number of grid points in which the

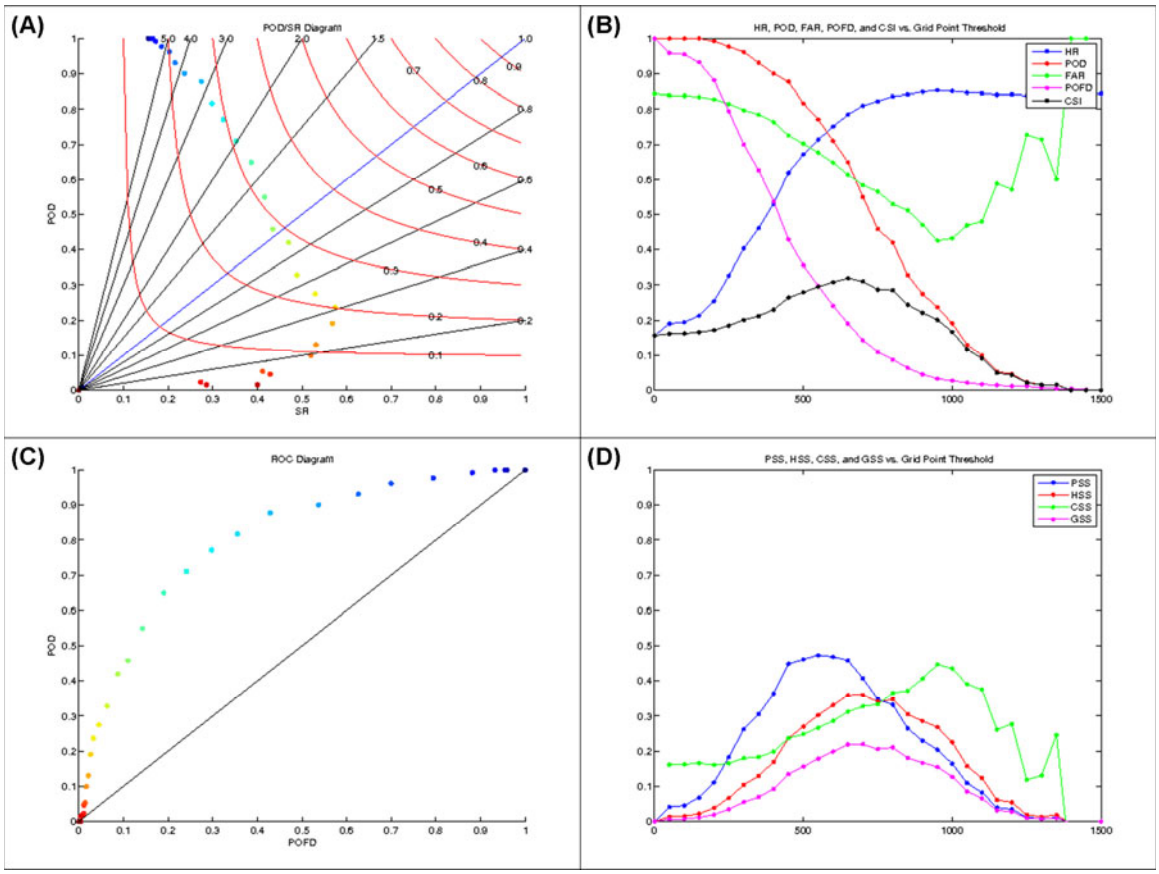


Figure 5: As in Fig. 4, using the mean hypothetical storm distance in the area through which the unconstrained STP ≥1. *Click image to enlarge*.

discriminating parameter, with the N25 index used for the classification of outbreak type (Fig. 4). The major severe weather outbreaks generally have a higher number of favorable grid points than the marginal outbreak days, with a considerable overlap observed with the intermediate outbreak days. As the threshold number of grid points is increased in increments of 100 from 0 to 5000, the HR increases, the POD decreases, the FAR decreases, and the POFD decreases. The FAR is quite large (>0.6 up to a threshold of 1500 grid points) and exceeds the POD for grid point thresholds of 1200 or more, in agreement with the subjective analysis discussed in Section 3a.

The POFD becomes quite small with increased threshold, as a consequence of $(a + c) << (b + d)$, $(a + b)$ decreasing, and $(c + d)$ increasing. The inequality is true because the dataset is highly imbalanced (i.e., the number of intermediate/marginal outbreak days $(b + d)$ far exceeds the number of major outbreak days $(a + c)$). Specifically, for the N25 index, 130 out of 840 days are considered major outbreaks. As the threshold number of grid points is incrementally increased, both a and b (c and d) decrease (increase) by design. However, because b << d with increased threshold, the POFD decreases rapidly in the small-to-moderate portion of the thresholds, as a large number of correct nulls are associated with a small number of grid points exceeding the favorable threshold. The POD, on the other hand, remains comparatively large within the small thresholds, as $c << a$. As a result, the PSS can become quite large compared to other skill statistics when using a low threshold to discriminate the outbreak types.

The discussion above illustrates a potential drawback of using the PSS for this dataset. The PSS can be quite large for classifications with a large number of false alarms (high FAR) and consequently a large positive bias (Fig. 4a). As $b << (b + d)$ for increasingly large thresholds, b can be quite large despite a relatively high PSS. Some skill statistics (such as the HSS) do not

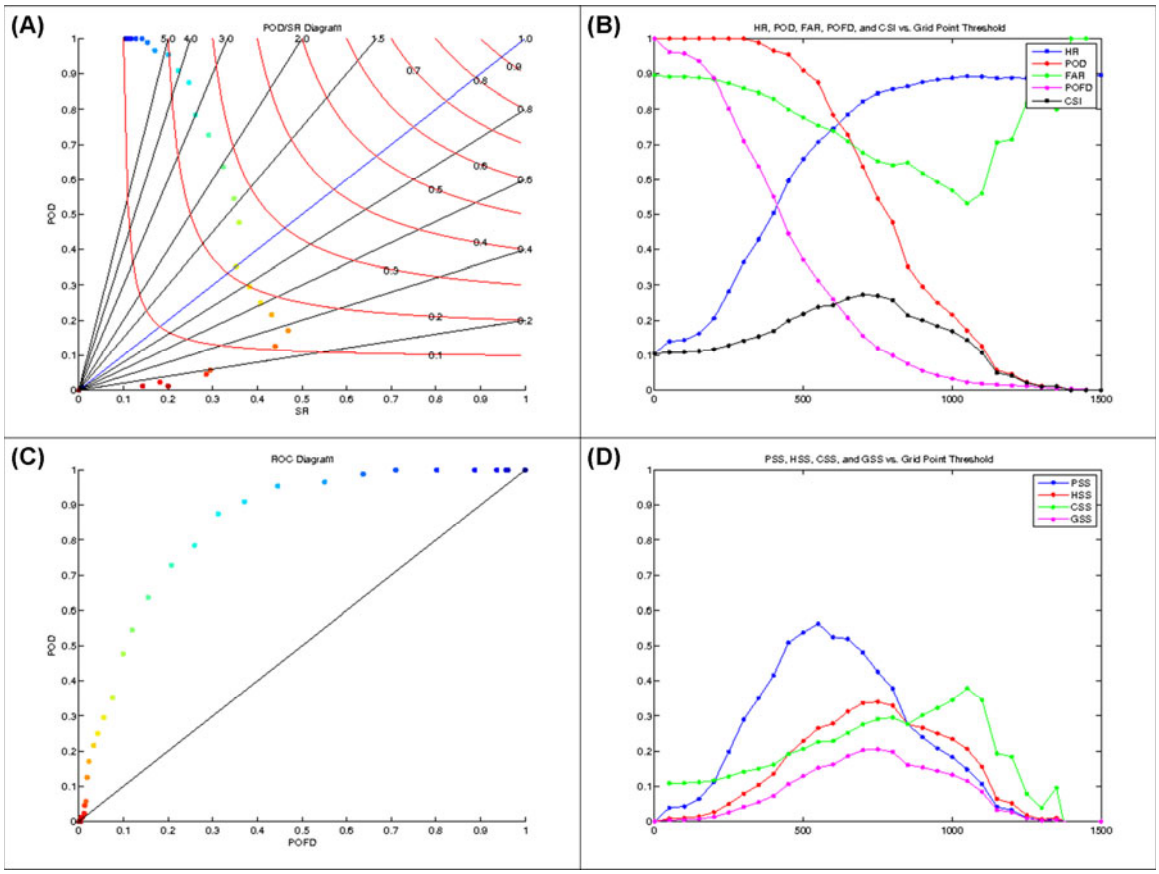


Figure 6: As in Fig. 5, with outbreak classification modified by using the N15 index rather than the N25 index. *Click image to enlarge*.

exhibit this tendency (see Doswell et al. 1990 for a discussion).

The above discussion assumes that the cost of a miss is equal to the cost of a false alarm, which is not necessarily true. The choice of skill statistics requires careful consideration when the costs of false alarms and misses are not equivalent. For example, generally the cost of false negatives is assumed to be higher than that of false positives (see, e.g., Doswell 2004). Although the cost of false positives is certainly nonzero, as preparatory efforts by emergency management and weather forecast agencies could be costly and a false-alarm effect (Breznitz 1984) may exist, false negatives may lead to increased casualty rates as a result of inadequate preparation and public awareness. Thus, even though the CSI, HSS, and GSS tend to be highest at grid point thresholds in which the bias is near unity (see Figs. 4a,b,d), diagnoses may be preferred to have a large positive bias, as the POD is higher at these thresholds with relatively little decrease to the CSI, HSS, and GSS but a large increase in the FAR. See Murphy (1977), Katz and Murphy (1997), Roebber and Bosart (1998), Briggs (2005), and references within for more discussion of forecast value.

The contingency statistics computed in Fig. 4 are remarkably similar to previous work discriminating storm types. Thompson et al. (2003) described the utility of STP in discriminating tornadic and nontornadic supercells, using POD, FAR, CSI, and PSS (their Fig. 19). They found that the FAR was very high (>0.6 for all values of STP ≤4), with the POD becoming less than the FAR, as STP was increased to values greater than unity. The PSS was relatively high (>0.4) for an STP threshold of 1, with a CSI much lower (<0.3) that peaked at higher thresholds. Similarities of the results in this study with those of studies focusing on storm discrimination suggest the accuracy and skill when using severe weather parameters in diagnosing outbreak types may be limited in a comparable way to studies attempting to discriminate storm types.

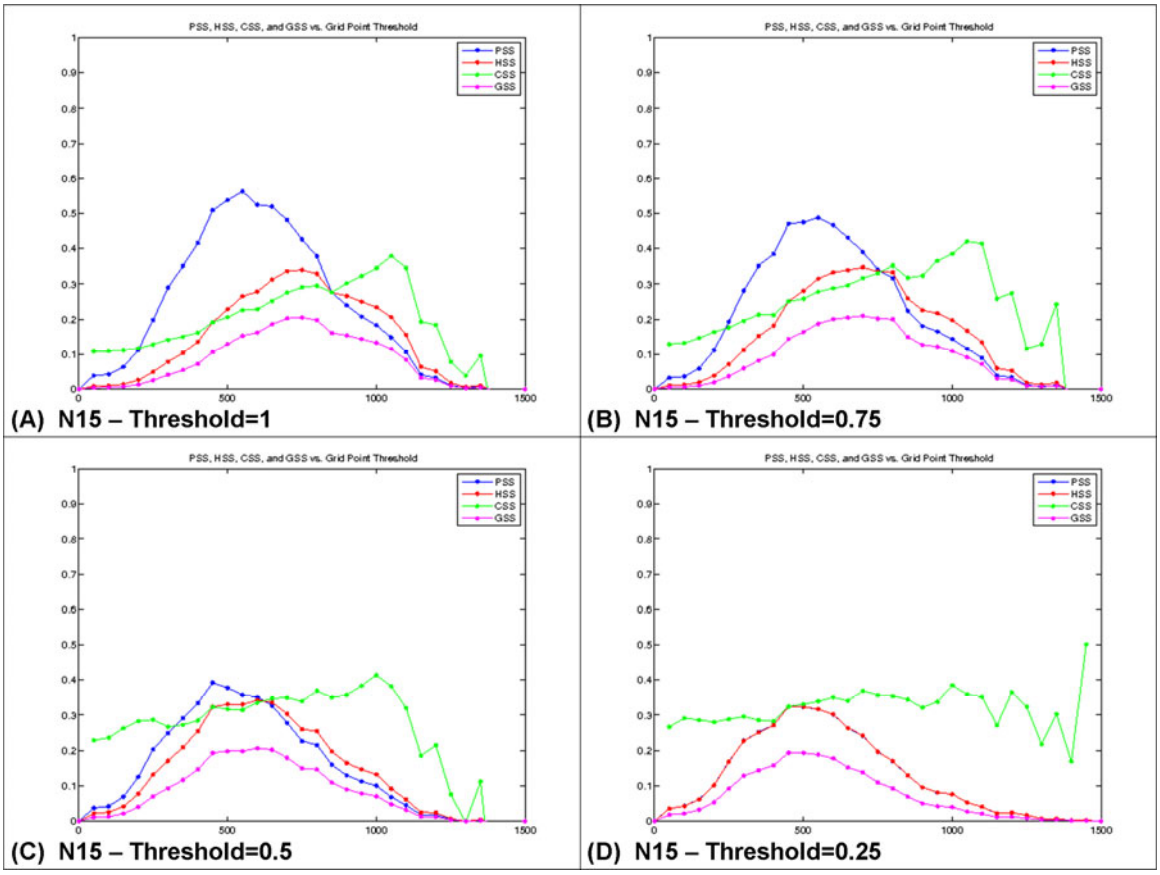


Figure 7: a) Same as Fig. 6d. b), c), d) As in (a), with the N15 index threshold, used to discriminate major from intermediate and marginal outbreak days, changed to 0.75, 0.5 and 0.25 respectively. *Click image to enlarge*.

The statistics exhibit only limited changes if the threshold criteria for diagnosing outbreak type are altered, or if the actual outbreak type is defined using different criteria (see below). For example, changing the threshold criteria by using the mean hypothetical storm distance of the trajectory within the area in which STP $\geq$1 resulted in contingency statistics that are quite similar to those of using the constrained STP areal coverage (cf. Figs. 4 and 5). When the criteria for actual outbreak classification were altered by using the N15 index (Fig. 6), some subtle changes were observed. Approximately 20% of major outbreaks using the N25 index were no longer classified as major outbreaks using the tornado-dominant indices, and nearly a third of these cases were no longer classified as misses, as a result. These cases typically involved only a small number of tornadoes. This led to an increase in POD (PSS) from 0.89 (0.47) to 0.91 (0.55) using mean hypothetical storm distances of 500 km within the region of STP values greater than unity (cf. Figs. 5 and 6). As STP was formulated for the discrimination of tornadic and nontornadic supercells (Thompson et al. 2003), the improvement is predictable. Note, however, this also led to an increase in FAR (from 0.72 to 0.78), as the remaining cases classified as intermediate by the N15 index were diagnosed to be false alarms. These differences in outbreak classification by the various indices are unavoidable, given the modifications to the weights of the variables used to develop the ranking schemes (SD10).

The selection of the threshold index score of 1 to separate major from intermediate outbreaks can be modified, depending on the desires of the decision-maker regarding skill and value of the diagnosed classifications. Lowering the threshold from 1 to 0.25 in increments of 0.25 (Fig. 7) suggests the following: (1) The maximum PSS decreases as the outbreak classification threshold is lowered, whereas the other skill statistics exhibit little change in their maximum values. (2) As the outbreak classification threshold is lowered, the maximum skill scores (aside from PSS) occur at lower areal

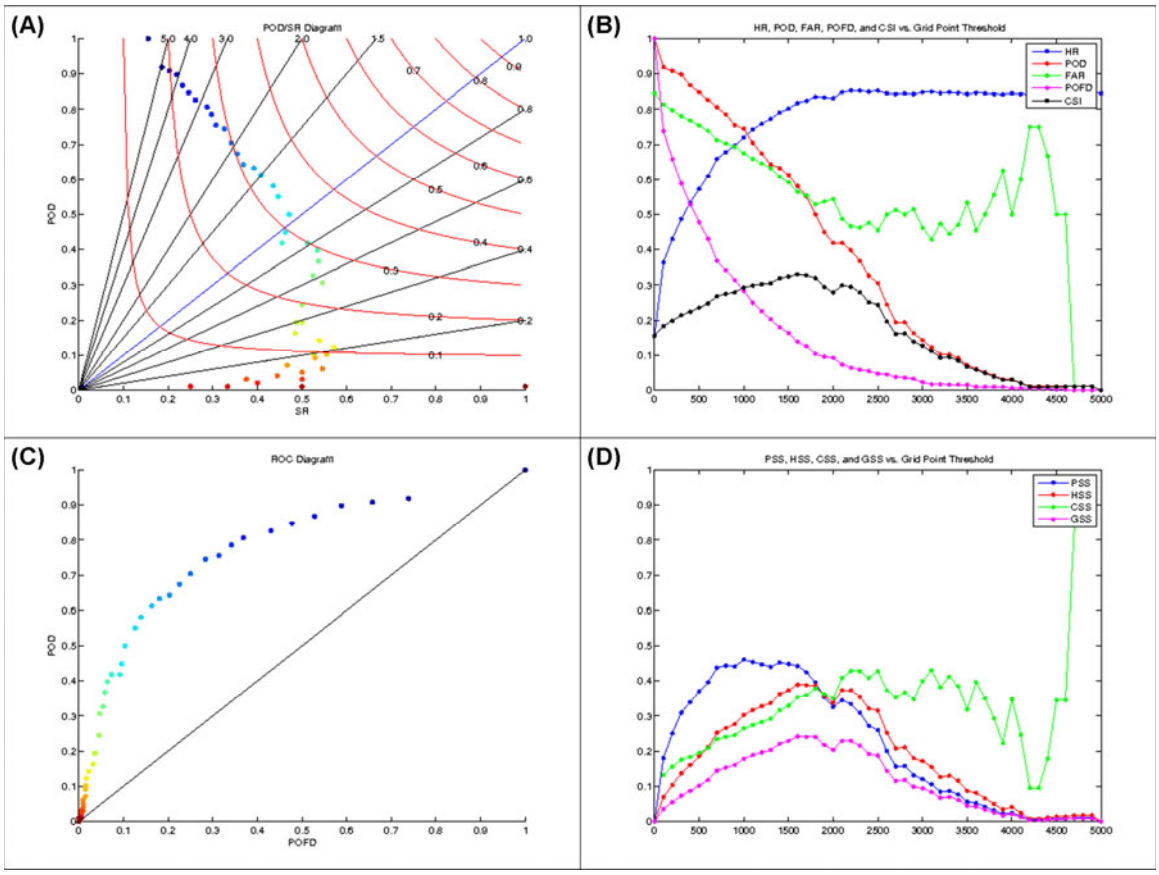


Figure 8: As in Fig. 4, for the training cases only. *Click image to enlarge*.

coverage thresholds, whereas the maximum PSS occurs at similar areal coverage thresholds. (3) As a result of (1) and (2), the PSS approaches the HSS at lower outbreak classification thresholds.

If the index score thresholds are increased above 1 (not shown), the following results are observed: (1) The POD increases up to index thresholds of 1.75 (specifically, from 0.90 to 0.97 using the same configuration as in Fig. 7). (2) The FAR and POFD also increase. The difference between POD and POFD increases up to index thresholds of 1.75. The POD/FAR tradeoff is similar to what is observed in Brooks (2004), regarding tornado warning evaluation (see also the discussion of the duality of error in Doswell 2004). (3) As a result of (1) and (2), the PSS increases slightly up to an index threshold of 1.75. The tendency of PSS to increase in this manner resembles trends discussed in Doswell et al. (1990). (4) Accuracy (HR) decreases as index threshold increases, leading to lower HSS and GSS with increased index threshold.

The preceding discussion implies that a determination of value when discriminating outbreaks is appropriate when selecting threshold values of areal coverage, the index used to classify outbreak types, and the threshold score of the index chosen. If the objective of predicting the occurrence of every major severe weather outbreak is more important than the over-prediction of these events (i.e., if the cost of "misses" is greater than the cost of "false alarms"), the areal coverage threshold criteria for discriminating outbreak type generally should be lower (near the peak of PSS) and the index threshold should be higher (above 1). On the other hand, if the objective is to minimize the number of false alarms, higher areal coverage thresholds (near the peaks of CSI and GSS) and relatively low values of the index threshold should be used (at or below 1). Determination of these thresholds in an operational setting is beyond the scope of this study, however.

The preceding analysis included all 840 cases. However, splitting the cases into training and testing sets is more instructive, as it provides

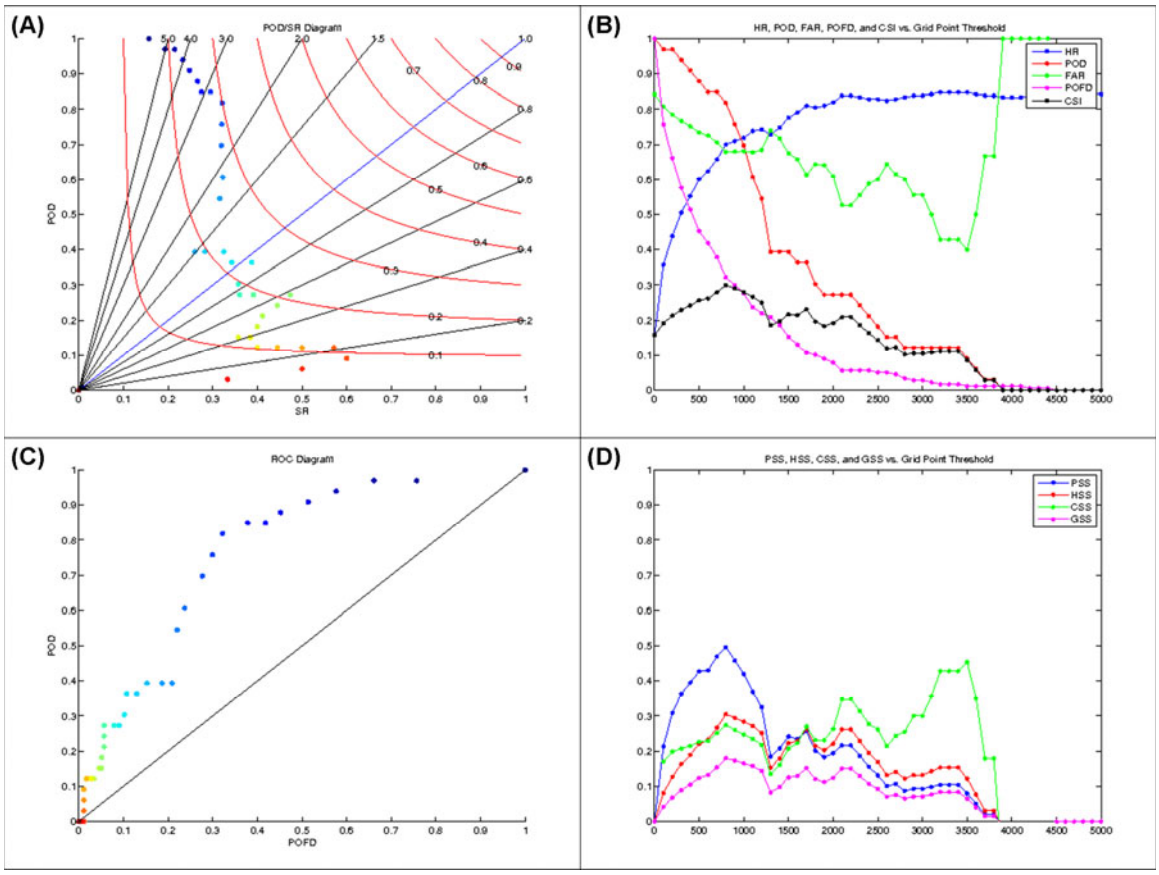


Figure 9: As in Fig. 4, for the testing cases only. *Click image to enlarge*.

Table 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

information on the generalization of outbreak discrimination criteria. *It is emphasized that the preceding analysis with all 840 cases was not used in any way to conduct the training/testing analysis discussed below.* Using the constrained STP criteria (as in Fig. 4) for the training cases (Fig. 8) and testing cases (Fig. 9), we see that, indeed, the behavior of the contingency statistics

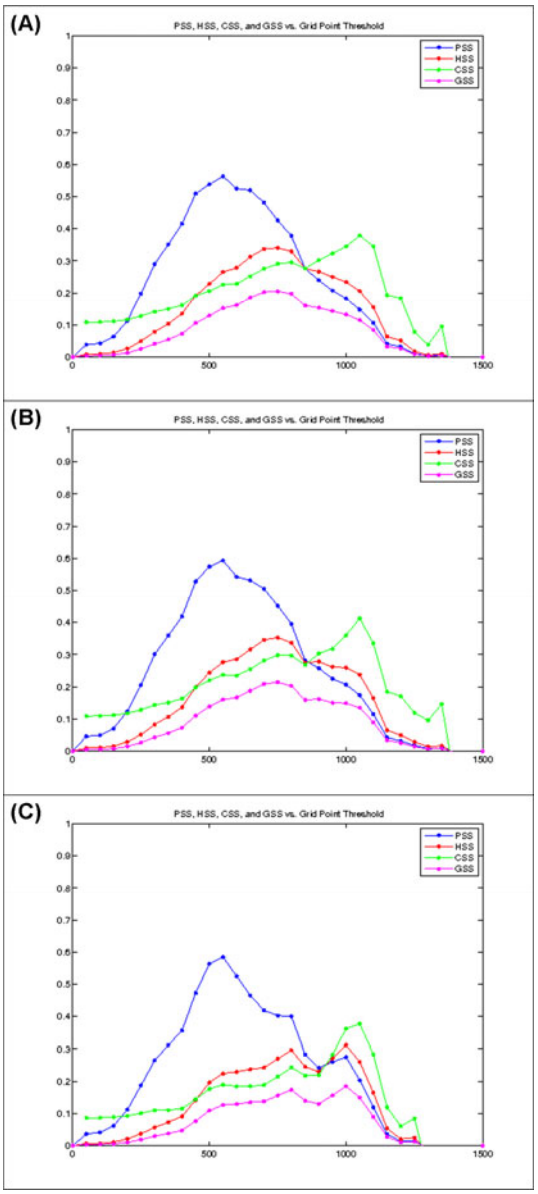


Figure 10: a) Same as Fig. 6d. b) As in (a), for 630 training cases. c) As in (a), for 420 training cases. *Click image to enlarge.*

of the training data is quite similar to the characteristics when using all 840 cases (cf. Figs. 4 and 8). However, the results of the testing data appear to be more volatile. The testing data consist of a small number of cases at relatively high grid point values (not shown). This suggests that the contingency statistics, when using higher thresholds, are subject to substantial uncertainty.[7]

Unfortunately, the availability of a large number of testing cases while maintaining a relatively large training set is limited. Reducing the training size to 50% of the total number of cases (420 out of 840) resulted in a worsening of the various statistics. For example, when using the mean distance of a storm trajectory within the region in which the unconstrained STP ≥ 1 and classifying outbreaks based on the N15 index (as in Fig. 6), the skill scores of the entire data set and the training set using 630 cases appear to be comparable (cf. Figs. 10a,b). However, a reduction of the training set to 420 cases showed a marked change in the skill scores at relatively high thresholds, indicating that the number of cases featuring high values of the mean distance is not adequately sampled (cf. Figs. 10a,c). These results were consistent, no matter which threshold and classification criteria were used (not shown). As a result of the above analysis, a training sample of 75% of the total number of cases (630 out of 840) was deemed appropriate. However, the disadvantages of the relatively small size of the testing data sample are important to keep in mind as the results are described in the following sections.

c. Volatility of the statistical models

To determine the volatility of the training statistical models developed using the areal coverage criteria, a collection of 25 subsets of the training sample, using 90% of the cases (randomly selected) for each subset, was used to develop 25 statistical models. The contingency statistics describing the capability of the statistical models to discriminate major severe weather outbreaks from intermediate and marginal outbreak days were computed. Plots of the 25 scores then could be used to provide guidance on the uncertainty of the statistical models *if they are tested on the same data (the testing data)*. The choice of 25 statistical models was based on computational practicality and explanatory power of the volatility of the models. The use of multiple statistical classifying algorithms was applied (Table 1), and their algorithm IDs are applied in the relevant figures for the rest of this section.

[7] The ratio of major severe weather outbreaks to the total in each dataset, using the value of the N25 index as the outbreak classification criterion, was 0.154 for the 630 training cases, 0.157 for the 210 testing cases, and 0.155 for the entire dataset.

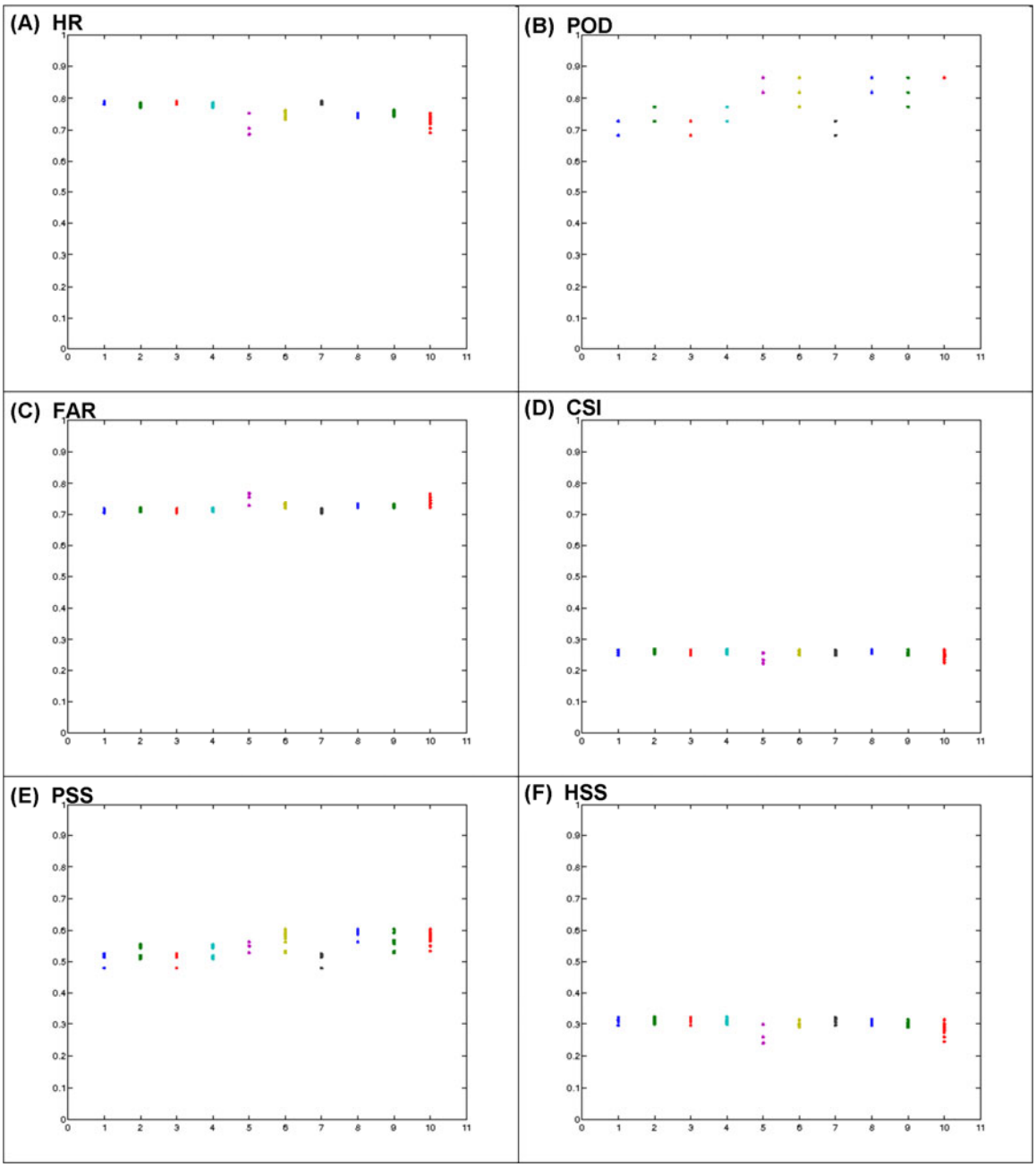


Figure 11: Contingency statistics (labeled) of 25 statistical models developed from the training data, using the mean hypothetical distance a storm travels through the area in which the unconstrained STP ≥1 as the diagnostic criterion, and the N15 index as the verifying outbreak classification criterion. Statistical algorithm IDs are on the *x*-axes and are identified in Table 1. Statistical models with the same contingency scores as others are overlain on preceding dots. *Click image to enlarge*.

The results of the ten statistical algorithms using the N15 index to classify outbreaks as major (if score ≥1) or null events and the mean distance of the hypothetical storm through the area in which the unconstrained STP ≥1 as the diagnostic criterion (as in Fig. 6), can be summarized as follows: (1) The contingency statistics obtained by the statistical algorithms are comparable to the "best" thresholds using the iterative threshold approach (cf. Figs. 4b,d and Fig. 11). (2) No statistical algorithm seemed to perform substantially better or worse than the others (for this particular variable/index combination). This result was

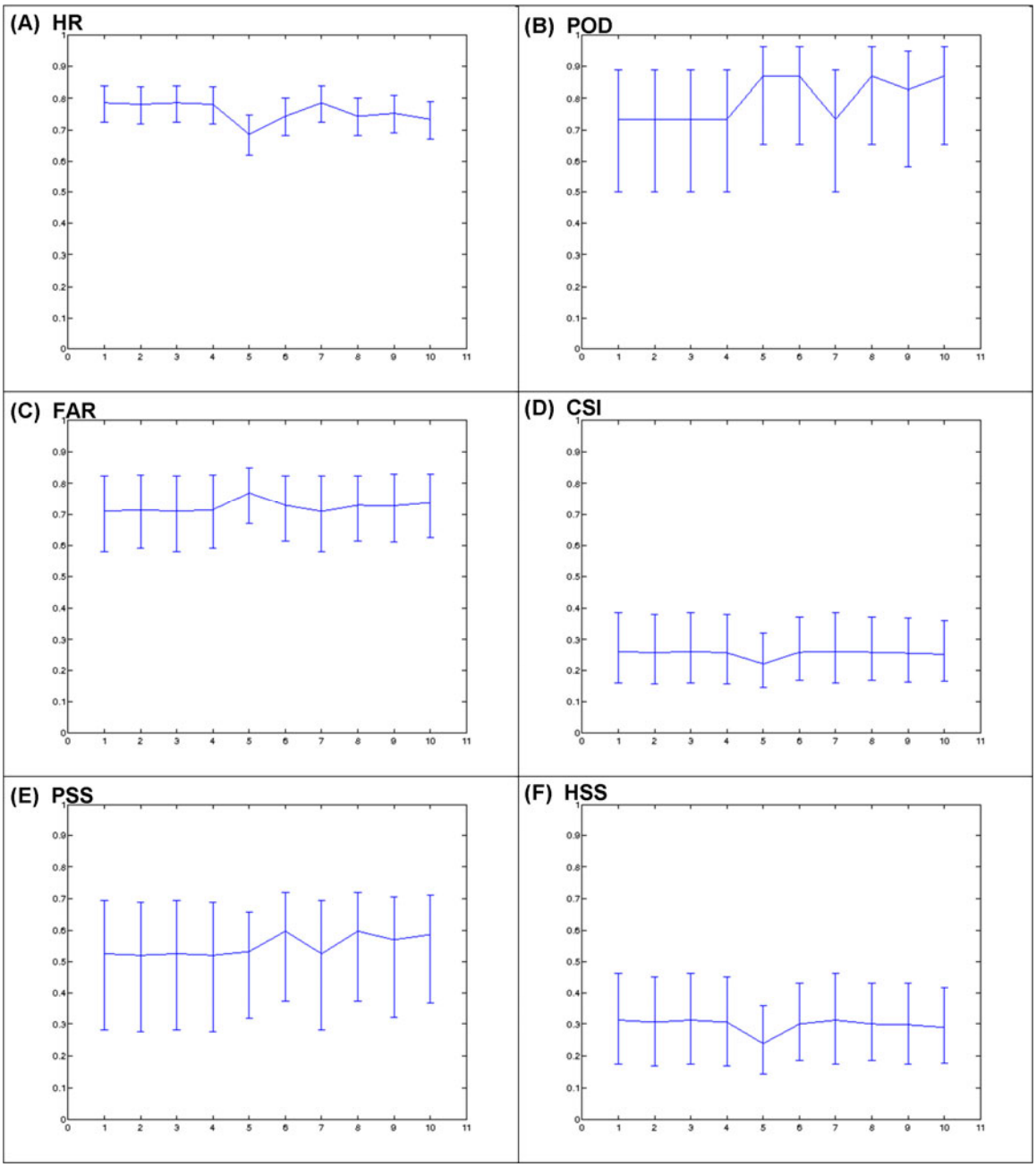


Figure 12: Bootstrap 95% confidence intervals of the contingency statistics (labeled) for the testing set, using all 630 training cases. The medians are shown by the connecting line for each of the 10 statistical models (*x*-axes; algorithms identified in Table 1). Threshold and outbreak classification criteria as in Fig. 11. *Click image to enlarge*.

true in general. (3) The POD is subject to relatively large variability, as the number of major severe weather outbreaks ($a + c$) is small (33) in the testing set. However, the remaining statistics featured very little variability among the 25 statistical models, indicating relatively little uncertainty in the statistical models that were trained. (4) The relatively high PSSs compared to HSSs (cf. Figs. 11e,f) agree with the iterative threshold analyses (e.g., cf. Figs. 6d and Figs. 11e,f), and support the conclusions of Doswell et al. (1990) regarding the limitations of using PSS in a rare-events dataset.

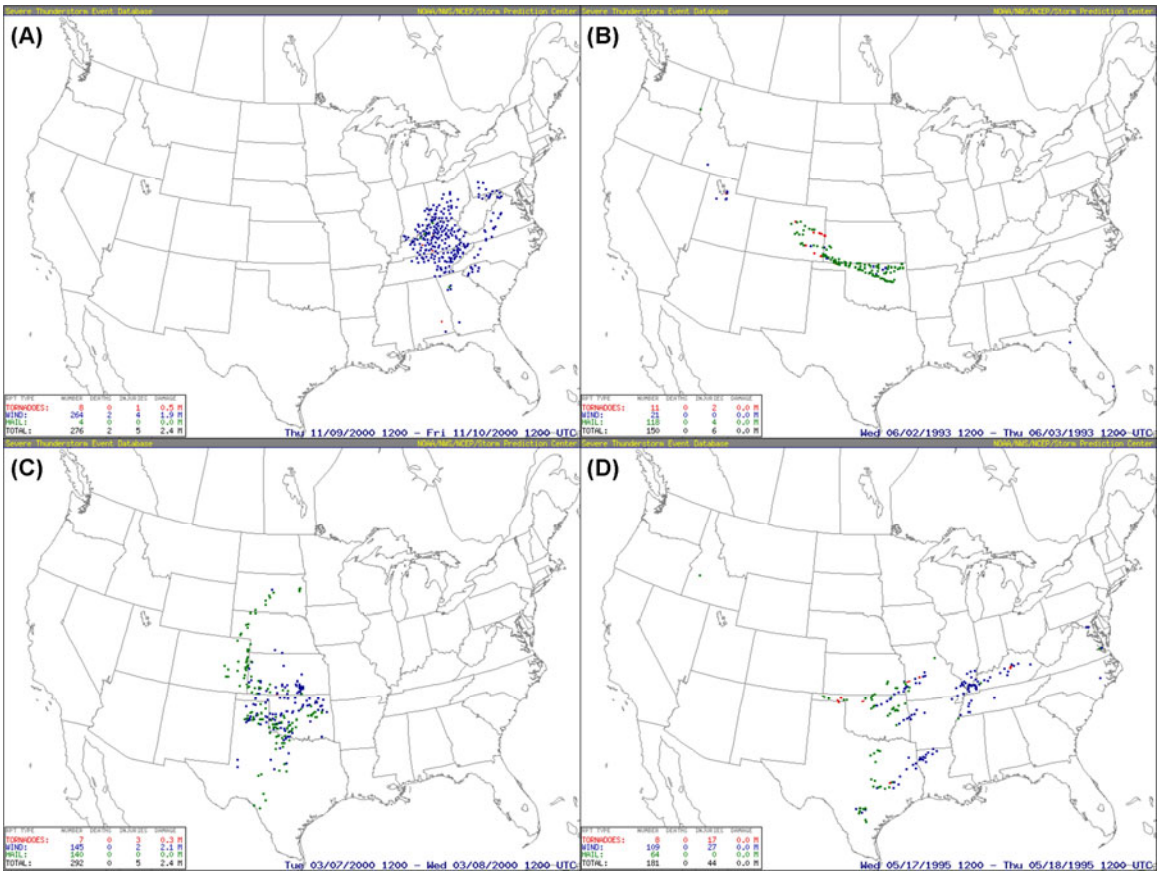


Figure 13: Severe reports from 1200 UTC on the indicated date to 1200 UTC the following day, with red dots denoting tornadoes, green dots denoting large hail, and blue dots denoting wind gusts or wind damage, for the a) 9 November 2000, b) 2 June 1993, c) 7 March 2000, and d) 17 May 1995 outbreak days. *Click image to enlarge*.

Multiple threshold criteria can be used as multidimensional input into the statistical algorithms. If the constrained STP threshold (as in Fig. 7) and the mean distance of a hypothetical storm in the area in which the unconstrained STP ≥1 (as in Fig. 6) are combined, the results suggest subtle improvement in the various contingency statistics with some of the statistical algorithms (not shown). However, numerous false alarms remain, no matter which combination of thresholds is used.

d. Uncertainty in the testing data

An additional means of expressing the uncertainty in the measurements is by obtaining bootstrap confidence intervals of the contingency statistics of the testing data results. For this procedure, all 630 training cases were used, and contingency statistics were computed for the testing data. The 210 (testing cases) × 4 (one column for each possible entry on a binary contingency table) matrix was resampled with replacement, and bias-corrected and accelerated bootstrapping (see Efron and Tibshirani 1993 and Hodges 2008 for a description of the technique) was employed to obtain 95% confidence intervals (CIs) of the statistics for each of the ten statistical algorithms tested. As noted above, the uncertainty was expected to be large, as the number of major severe weather outbreaks in the testing dataset was small, and the number of cases with relatively high values of whatever threshold criteria are employed was small. Using the same threshold and outbreak classification criteria as in Fig. 11, the 95% CIs of the testing data were found to be quite large (Fig. 12). For example, the 95% CIs had a range on the order of 25% for FAR and CSI, 40% for PSS, and 25% for HSS. Although variations in the median skill scores were present using the various statistical algorithms, these differences were not significant.

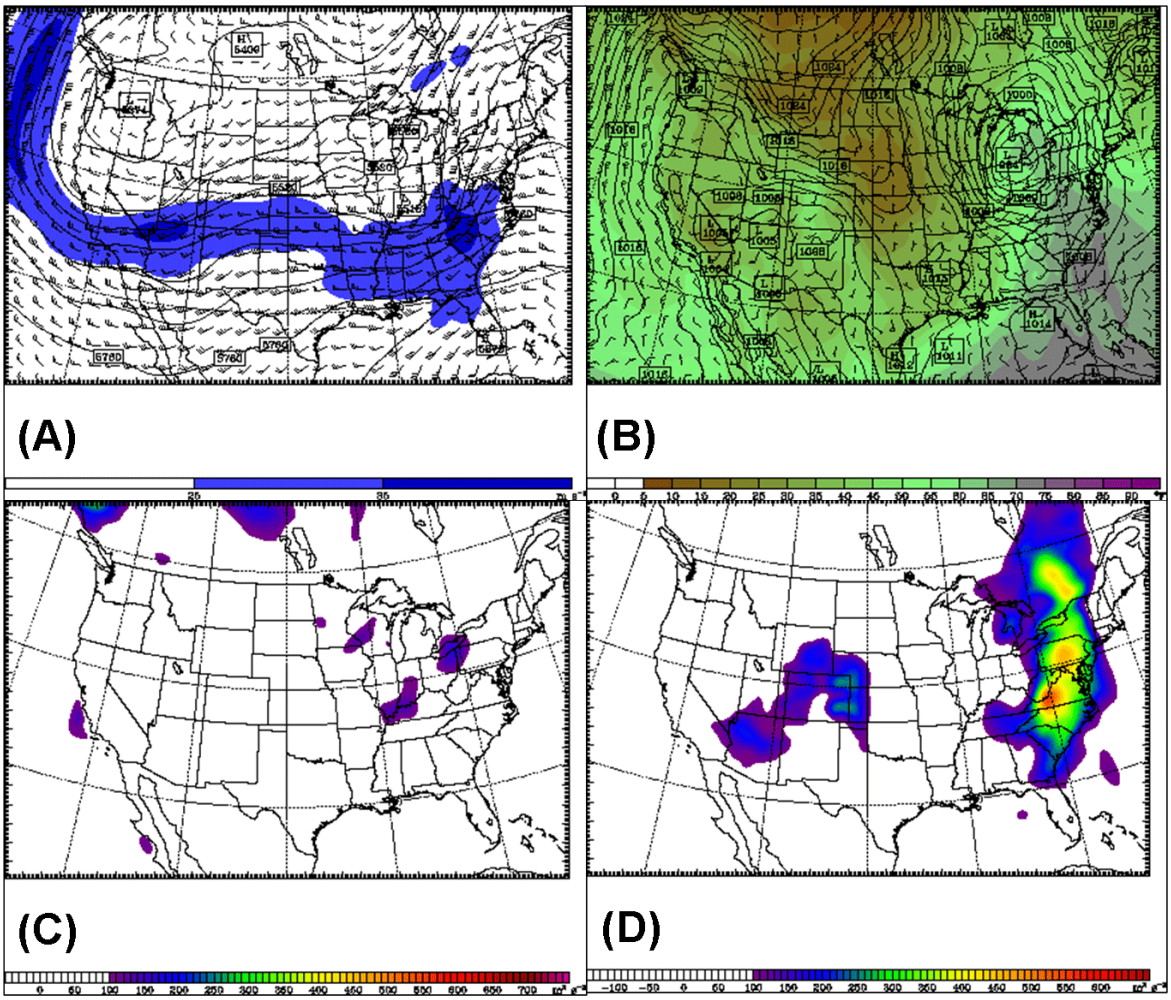


Figure 14: NARR fields valid at 0000 UTC 10 November 2000 of: a) 500-hPa winds (barbs in kt), isotachs (filled contours in m s^{-1}), and geopotential heights (contours in m); b) surface dew point temperature (filled contours in ºF), winds (barbs in kt), and mean sea-level pressure (contours in hPa); c) SBCAPE (J kg^{-1}); and d) 0–1 km SREH (m^2 s^{-2}). *Click image to enlarge.*

As a result of the large uncertainty in the statistics of the testing data (i.e., the large CIs), there was difficulty in identifying particular areal coverage parameters that were statistically significantly better than others in discriminating major from intermediate and marginal outbreak days. However, the findings were relatively consistent with the analysis in sections 3b and 3c, as desired.

4. Subjective interpretation

The interpretation of discrimination of major severe weather outbreaks from intermediate and marginal outbreak days is subject to several challenges and limitations. False alarms are quite common, with many intermediate and sometimes even marginal outbreak days incorrectly classified as major severe weather outbreaks using the areal coverage of severe weather parameters. Furthermore, some major severe weather outbreak days are classified incorrectly as intermediate to marginal outbreak days. Analysis of the cases that commonly are misclassified provides valuable insight into the weaknesses of the methods both in ranking the outbreak cases and in the methods used for outbreak discrimination.

Several major severe weather outbreaks between the index values of 1 and 2, using the indices in which only a subset of the tornado variables are used (N17–N19 and N21–N25; see SD10), are misclassified as intermediate or

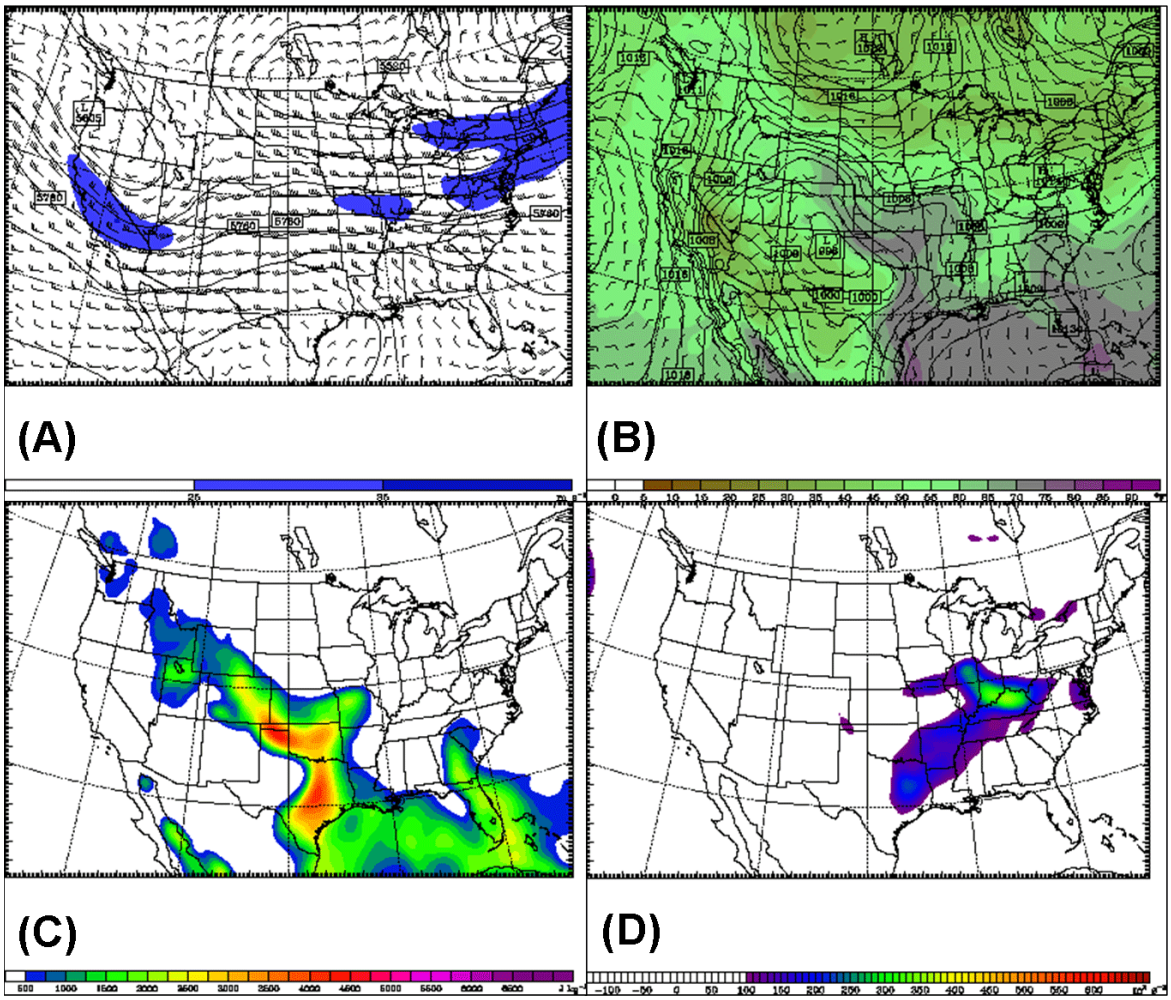


Figure 15: As in Fig. 14, valid at 0000 UTC 3 June 1993. *Click image to enlarge*.

marginal. These outbreak cases tend to be of two types: few or no tornadoes occur, or the geographic area is small. The 9 November 2000 day (Figs. 13a and 14) is an example of a primarily nontornadic outbreak classified as a major severe weather outbreak using the N25 index. This wind-dominant event was associated with 0–1 km AGL SREH >100 m^2 s^{-2} (Fig. 14d), but SBCAPE <500 J kg^{-1}. Thus, EHI and STP fields (not shown) were very small, and this case was diagnosed incorrectly as an intermediate event. When the index used to distinguish major from intermediate and marginal outbreak days was changed to those indices that used all of the tornado variables (N0–N16 and N20; see SD10), the score decreases to below 1 (e.g., 0.93 for N15 versus 1.40 for N25), resulting in a correct areal coverage diagnosis. Not surprisingly, when using parameters specifically developed to distinguish tornadic from nontornadic environments and/or storms (such as STP), indices with all of the tornado variables are more appropriate.

A second example is the 2 June 1993 outbreak day (Figs. 13b and 15). With this event, a large number of significant hail events (32) occurred in portions of Colorado, Kansas, and Oklahoma. This led to a relatively high index value of 1.32 using the N25 index, as this index weighs significant nontornadic reports relatively highly and because the coverage of the severe reports was small (i.e., the reports were very tightly clustered). The latter characteristic contributes to comparably high index scores, because of the incorporation of the so-called "middle-50% parameter" (first introduced in Doswell et al. 2006). This variable is designed to counteract the tendency for days in which a large number of reports are scattered across a large geographic region to be considered major severe weather outbreaks. The variable works

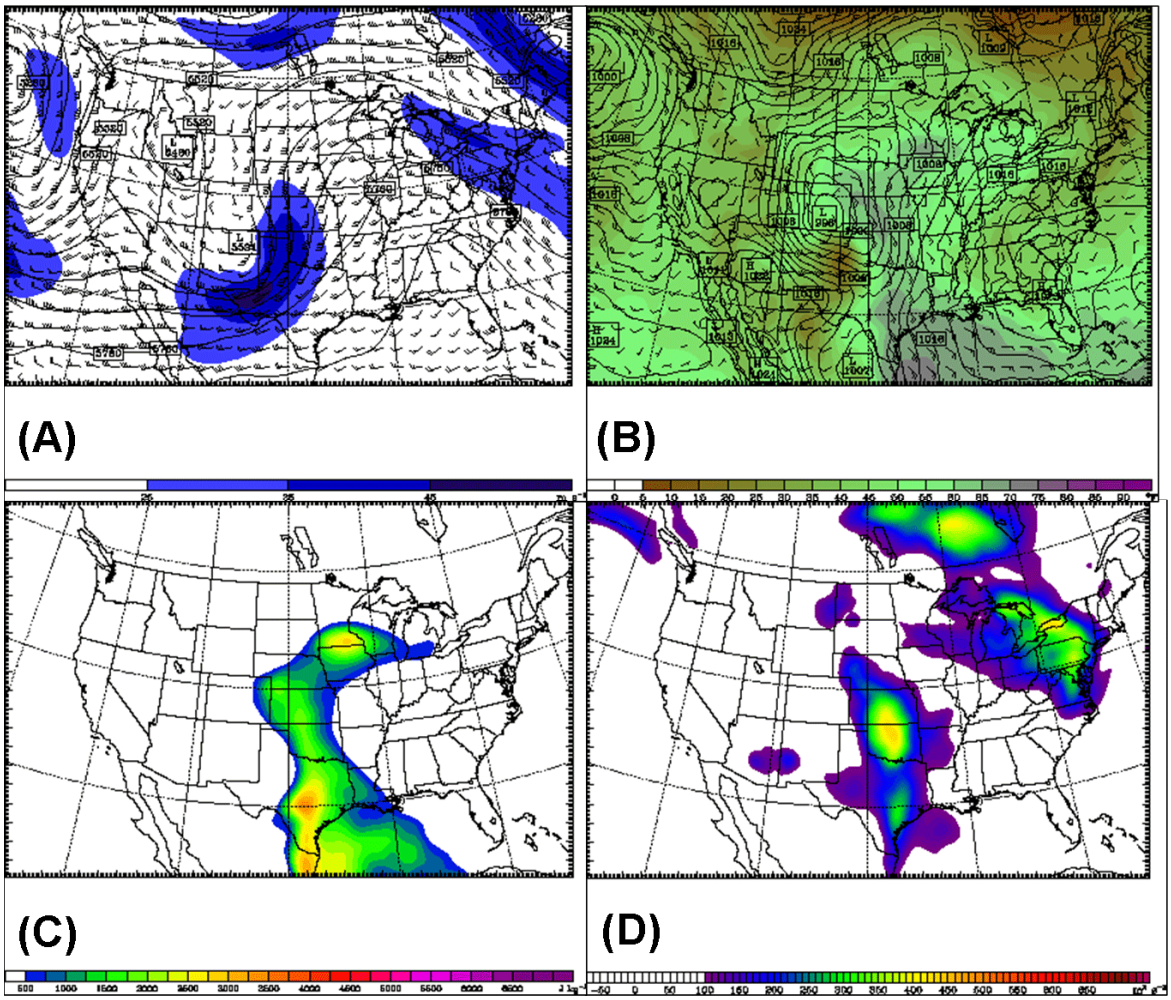


Figure 16: As in Fig. 14, valid at 0000 UTC 8 March 2000. *Click image to enlarge*.

by finding the differences of the 25th and 75th percentiles of the latitudes and longitudes separately, and multiplying these differences together. The result is a latitude-longitude area (see SD10; their Fig. 3). The smaller area the reports encompass, the higher the final ranking index score.

As there were only 11 tornadoes reported on 2 June 1993, the indices in which all of the tornado variables were included had slightly lower scores (e.g., 0.86 for N15). With these types of events, the areal coverage of favorable severe weather parameters was quite small (Figs. 15c,d), and the number of tornadoes was relatively low. Thus, this case would be classified incorrectly as an intermediate outbreak for indices N17–N19 and N21–N25, whereas it would be classified correctly for indices N0–N16 and N20.

This example exposes a drawback of the ranking indices and/or using areal coverage as a means of diagnosing outbreak type. Events with a large number of reports over a very small region are ranked higher than the same number of reports over a larger region. Thus, these cases are more likely to be classified as major events, whereas using areal coverage as a means of diagnosing outbreak type potentially could favor events occurring over a larger area.

Typically, cases classified as false alarms account for over half the diagnoses of major severe weather outbreaks. Many of these cases exhibit certain characteristics that could be used to distinguish these events using modified techniques. However, such modifications typically resulted in the addition of other undesirable effects. A common type of case that was misclassified as a major severe weather

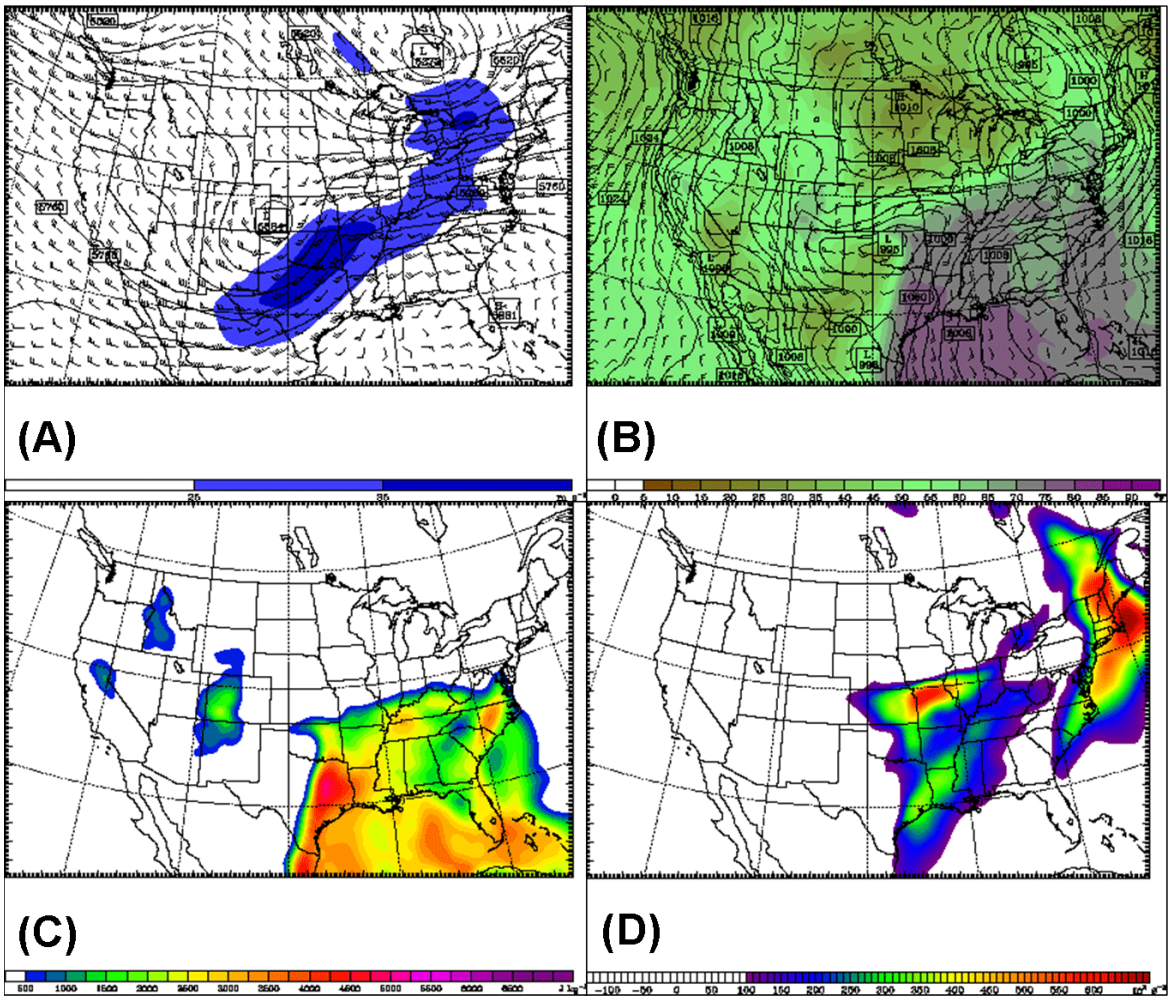


Figure 17: As in Fig. 14, valid at 0000 UTC 18 May 1995. *Click image to enlarge*.

outbreak featured midlevel flow oriented nearly parallel to a surface boundary (or perpendicular to θ_e gradients).

Two map types commonly demonstrate this characteristic. The first type exhibits quasi-zonal midlevel flow atop a warm or stationary front. These events commonly feature a large, unstable warm sector with ample low-level shear near the warm or stationary front. Because of these characteristics, large regions of favorable severe weather parameters exist, typically in an elongated region surrounding the surface boundary.

The second type features highly amplified, nearly meridional midlevel flow atop a meridionally-oriented cold front or dryline. The 7 March 2000 outbreak day (Fig. 13c; Fig. 16) is an example of this type of event. The 500-hPa shortwave trough approaching the region is highly amplified, displaying a negative tilt (Fig. 16a). The surface boundary is oriented northwest to southeast across the southern high plains (Fig. 16b). Thermodynamic instability (Fig. 16c) and SREH (Fig. 16d) typically are situated in a relatively narrow, meridionally oriented region in the warm sector. As the midlevel flow is parallel to this favorable region, pseudo-trajectories calculated within this region would result in long distances and times in which the storm remains in the favorable area. However, the convective mode with these events frequently is linear, which is not favorable for widespread tornado development.

These two types of cases are straightforward to identify visually in a pattern recognition sense but are more difficult to account for using the areal coverage calculations. However, the inclusion of new parameters that describe these

characteristics could be highly beneficial in discriminating these cases correctly. A recent study by Dial et al. (2010) provides examples of variables (e.g., component of cloud-layer shear normal to a boundary) that could be incorporated in future work.

Ultimately, there are cases in which no obvious methods exist to prevent their classifications as false alarms. The 17 May 1995 outbreak day (Fig. 13d; Fig. 17) is one example. A relatively small number of tornadoes, but an appreciable number of wind and hail reports, were observed in much of the southern plains. Shear and instability were both large in magnitude and extensive spatially (over a sizable portion of the southern US, Figs. 17c,d) with a strong shortwave trough approaching the area. Midlevel wind vectors contain a substantial component perpendicular to the surface boundary (Figs. 17a,b). The synoptic environment and the fields of severe weather parameters were quite similar to major severe weather outbreak days, yet the standardized scores of each variable used to create the multivariate index were negative (except for the number of significant tornadoes, which was slightly above zero). These cases, in particular, need to be studied in more detail to identify any synoptic and subsynoptic-scale effects and interactions that prevented the occurrence of a more significant severe weather outbreak.

Not surprisingly, most failures in outbreak classification occur near the threshold used to distinguish the two groups. For example, most major severe weather outbreaks misclassified as intermediate or marginal outbreak days have scores just above the value of 1, which is the value used to distinguish the two groups (not shown). Similarly, most false alarms occur with intermediate outbreak days that fall just below the index values of 1. Classifying most types of meteorological events into groups using specified thresholds is subject to such misclassifications. The use of statistical techniques to account for "close calls" may be a beneficial endeavor (see Barnes et al. 2007 for a conceptual example using tornado warnings), as the classification of these events is subject to uncertainty, given the nature of reporting and archiving severe weather and the inherent challenges in the ranking of these events (SD10).

5. Summary and conclusions

Subjective notions regarding the occurrence of major severe weather outbreaks include the presence of favorable ingredients for severe weather in a relatively large region. With recent studies focused on identifying and ranking these events according to meteorological and societal significance in a relatively rigorous and repeatable way (Doswell et al. 2006; SD10), testing this particular notion with a relatively large sample of cases was possible. The results of this study indicate that there was a tendency for the most severe convective outbreaks (primarily major tornado outbreaks) to have larger regions of severe weather parameters above specified thresholds compared to less severe (intermediate) and marginal outbreak days. The use of areal coverage was tested in two ways: computing the total number of grid points within a specified domain covering the US that exceeded a predetermined threshold, and calculating backward and forward pseudo-trajectories of hypothetical storms originating at each grid point within the region in which specified meteorological parameters exceed a certain threshold.

Preliminary subjective analysis of the areal coverage calculations indicated certain limitations of the technique that required some adjustments. For example, several outbreak days featured very high values of areal coverage, simply because of the inclusion of grid points over water. Moreover, certain fields of meteorological parameters consisted of values considered favorable for significant severe weather, despite the presence of other meteorological conditions perceived to be unfavorable for its development (such as high instability but very low shear, and conversely). In an attempt to counteract these undesirable effects, additional constraints were imposed on a subset of the areal coverage computations, including the elimination of grid points over water and the inclusion of instability and helicity constraints for any grid point considered. However, the incorporation of any constraint typically led to a larger number of major outbreak days misdiagnosed as intermediate and marginal outbreak days.

A large number of intermediate and marginal outbreak days consisted of areal coverage values similar to major outbreak days. These large numbers of false alarms are similar to past studies attempting to discriminate storm and

severe weather types. Subjective analysis of these false alarms suggested that a subset could be identified as such because of the synoptic environment observed at the time of the outbreak. Specifically, events in which midlevel wind vectors were oriented parallel to a surface boundary were commonly misclassified as major outbreaks, as the conditions observed allowed for large regions of severe weather parameters favorable for severe storms.

The vast majority of marginal outbreak days, characterized by large geographic scatter or multiple regional clusters of severe reports, featured small values of areal coverage. However, some days appeared to have particularly large, contiguous regions of favorable values. This suggests that the "middle-50% parameter" method of accounting for geographic scatter in the severe reports, used by D06 and SD10, may be an inadequate way of accounting for these cases. Instead, a method to eliminate these days entirely or to identify multiple outbreaks on a given day (for days in which a large number of reports are clustered in multiple regions of the country) appears to be necessary. One method, currently under investigation, is the use of kernel density estimation to identify regions in which a specified threshold regarding coverage of severe reports is exceeded.

Other sources of error appeared to depend on the index used to rank the outbreaks in terms of severity (SD10). For example, indices that used a larger number of tornado variables to rank the outbreaks appeared to be classified more accurately and skillfully using the areal coverage technique, though this improvement was seldom statistically significant. As most severe weather parameters analyzed in this study were intended to describe environments favorable for tornadoes, this result was not particularly surprising. Furthermore, most false alarms occurred near the threshold of the index used to classify outbreaks as "major" versus those that were "intermediate or minor". Use of a "close call" technique, as proposed in Barnes et al. (2007), may be appropriate in future work to account for this tendency.

A large number of false alarm cases showed no readily apparent pattern or parameter differences between the major outbreak days and those classified as intermediate and marginal. This result demonstrates that: (1) scientific understanding involving our ability to discriminate the severity of outbreaks is limited; (2) past investigations of outbreaks or events may not have investigated null events as thoroughly as necessary; (3) the methods implemented to rank outbreak days are subject to the difficulties and limitations in observing and archiving severe weather, and likely do not counteract these undesirable characteristics completely; and (4) we likely are not observing and assessing the parameters that determine the relatively widespread occurrence of tornadoes and significant severe weather. Furthermore, the technique would require modification prior to its implementation as a forecasting tool, where many more false alarms would arise (i.e., the Bayesian inversion problem).

ACKNOWLEDGMENTS

Funding was provided by NSF Grant AGS-0831359. Figure 13 was obtained by using the online version of SVRPLOT V3.0, maintained by John Hart and Jared Guyer. The authors would like to thank forecasters from the Storm Prediction Center for suggesting this topic for investigation. We thank Paul Roebber, Michael Coniglio, and Corey Mead for their helpful reviews.

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.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone, 1993: *Classification and Regression Trees*. Chapman and Hall, 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., 2004: 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, 1994: On the environments of tornadic and nontornadic mesocyclones. *Wea. Forecasting*, **9**, 606–618.

——, ——, 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.

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

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

Cristianini, N., and J. Shawe-Taylor, 2000: *Support Vector Machines and Other Kernel-Based Learning Methods.* Cambridge University Press, 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.*, Vol. 79, Amer. Geophys. Union, 573–582.

Davies-Jones, R., D. W. Burgess, and M. P. 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.

Dial, G. L., J. P. Racy, and R. L. Thompson, 2010: Short-term convective mode evolution along synoptic boundaries. *Wea. Forecasting*, **25**, 1430–1446.

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

——, 2007: 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 Storms 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.

——, 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.

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

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

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

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.

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

Johns, R., 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.*, Vol. 79, Amer. Geophys. Union, 583–590.

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

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

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. E. 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., 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.

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

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

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.

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*, **25**, 921–930.

Rasmussen, E. N., 2003: Refined supercell and tornado forecast parameters. *Wea. Forecasting,* **18,** 530–535.

——, 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.

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

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

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.

——, A. E. Mercer, 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.

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

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, 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.]

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., 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.

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. *Meteor. Appl.*, **9**, 455–459.

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

REVIEWER COMMENTS

[Authors' responses in *blue italics.*]

REVIEWER A (Paul J. Roebber):

Initial Review:

Recommendation: Accept with minor revision.

General Comment: This is an interesting paper that seeks to evaluate the notion that some measure of the "size" of environmental conditions supportive of severe weather should correlate to the magnitude of the event.

Substantive Comments: In my view, these include all comments regarding scientific content, regardless of how substantial the implied revisions might be.

1) Section 2, first paragraph: What deficiencies exist in these data (NARR)? The advantage [with respect] to other reanalyses is apparently resolution. What benefits are there, though, [with respect to] RUC data, which are often used in studies of this kind? I'd like to know more about what dictated this choice. I suspect it was the period of record and otherwise the authors might have preferred to use RUC data rather than NARR. This being the case, I'd like to know what problems exist with NARR data and how these were circumvented or otherwise were not an issue.

We refer the reviewer to the following PowerPoint presentation, presented at the 85th Annual AMS Meeting for details:

http://www.emc.ncep.noaa.gov/mmb/rreanl/narr.ppt
or http://www.ejssm.org/ojs/public/vol5-7/narr.ppt

Notably, the regional reanalysis (RR) seemed to perform better than global reanalysis (GR) with near-surface temperatures and winds, important obviously for shear/CAPE calculations. Furthermore, lower grid spacing, particularly in the vertical (45 total vertical levels in the NARR, versus 17 in the NNRP), leads to more vertical levels sampling the boundary layer, another desirable characteristic of RR.

We note in the manuscript that we want to look at mesoscale fields for a large number of cases. The grid spacing of NNRP (e.g.) does not permit this (see Shafer et al. 2009). The RUC was not used because the NNRP permits usage of many more years of data. Given the sample size issues that remain even when using 28 years of cases, using a dataset with a smaller number of years available (such as the RUC) is not preferable.

Some limitations of the NARR, also in the presentation linked above, generally do not affect the research we are conducting. For example, we do not incorporate precipitation into the analyses, and we do not have any outbreaks in the Southwest (see slide 46).

We have included a link to this presentation in the article, and we have emphasized wording regarding the necessity for a large dataset in the same paragraph.

2) Second full paragraph of section 2: This is a standard means of splitting data. For meteorological information, however, where there is often temporal dependence between days, is this a factor? It is probably minor, since most outbreaks are probably separated in time but there may be a few events where this could be an issue.

This is a good point, but multi-day outbreaks are not guaranteed to be in the same category. For example, in this study, 3 May 1999, 4 May 1999, and 5 May 1999 were all included. The 3 May and 4 May outbreak

days were major, whereas the 5 May outbreak day was intermediate. Similarly, from 8-10 May 2003, the 8 May and 10 May outbreak days were major, whereas the 9 May outbreak was intermediate.

Additionally, consecutive outbreak days usually have diurnal maxima, indicating these events (though possibly associated with the same synoptic-scale system) are subject to different synoptic- and subsynoptic-scale environments and could be considered separate events. A subjective analysis of these cases (such as those listed above, and others) suggests that this is predominantly the case.

Of greater concern are the outbreaks that tend to fall near the boundaries of the 24-h periods, in which one event occurs over two days considered separately. There are no such cases included in this study. Future work will be conducted to account for and include these cases in subsequent discrimination studies.

3) Fourth full paragraph of section 2: SVMs are subject to poor performance from noisy data. Was this evaluated at all?

We believe this is occurring, as well as known problems with SVM discrimination of imbalanced datasets – though we have not investigated this fully. However, the point of using multiple statistical algorithms was to determine if one particular algorithm exhibited consistently better or worse performance.

4) [Section 3a], third bullet point: It seems like additional discrimination is provided by looking at the total grid points comprised of the intersection of multiple parameters exceeding thresholds. For example, although you state that you use STP for analysis of the areal coverage technique, you later state: "Only grid points in which SBCAPE ≥1000 J/kg and 0-1 km SREH ≥100 m^2 s^{-2} were considered in some subsequent calculations…" I think the procedure needs to be clarified a little. This goes back to the question of whether it is better to present the process of discovery which is necessarily iterative or to clean it up for the purposes of presenting the final results. I have more to say about this in the technical comments.

We have cleaned up some wording in Section 3a. Specifically, we've added a footnote specifically identifying the definitions of "constrained" and "unconstrained" STP in this paper, at the point referenced in the reviewer's comment. We have also modified the next paragraph to provide additional reasoning for the use of both "unconstrained" and "constrained" STP in subsequent calculations.

The "process of discovery" was presented in this paper for three primary reasons. First, we wanted to have a discussion devoted to describing ***the data*** *– i.e., that areal coverage is noisy for outbreaks of various ranks; that some cases were prone to unrealistically high values of areal coverage because of water coverage, biases in STP (and other index) calculations, etc. The (admittedly subtle) point of this section is that a naïve implementation of the grid point sums is not appropriate. Second, there are shortcomings when accounting for the observed noisiness, STP biases, etc. That is, implementing additional constraints in the data inevitably leads to a side effect. This side effect is typically an increase in the number of "misses" of the major outbreak cases. As a separate reviewer suggested additional/modified constraints, it is likely we have understated this point, and we have emphasized in the revised manuscript. Third, this section immediately exposes the big problem with outbreak discrimination: an excessive number of false alarms – which is a primary focus of the rest of the paper.*

5) [Section 3b]: I like the approach of using multiple skill statistics, recognizing the need to consider multiple measures which have different properties.

Thank you.

6) [Section 3b]: You raise the issue of considering the cost of misses versus the cost of false alarms, but then end the discussion rather abruptly. It may well be beyond the scope of this work, but you can provide more information about what the tradeoffs are. In particular, false alarms are high as you have noted. In the context of severe weather, it is probably true that one prefers to suffer higher false alarms to avoid missing events, but I think further elaboration on this important point is warranted and given some of the authors' long experience, well within their capability. Again ("…numerous false alarms remain, no matter which

combination of thresholds is used."). Incidentally, I'd be interested to see how this might play out for the two days of 2 May 1999 and 3 May 1999.

We have added some discussion to this paragraph regarding the false positive/false negative tradeoff (and have added Doswell [2004] as a reference) to address this issue, but we refrained from introducing a lengthier discussion. This topic is certainly worthy of elaboration and consideration, but as this is not a forecasting study, it is somewhat tangential to our objectives in this paper.

More relevant to the study is the observed "duality of error" of our findings, addressed in Doswell (2004) and discussed with the topic of tornado warnings in Brooks (2004). We have added some text on this topic later in Section 3b, when discussion of modifying the classification thresholds (e.g., new Fig. 7) is presented.

We suspect the reviewer means 3 May 1999 and 4 May 1999 – both of which were major outbreak days, based on the indices developed in Shafer and Doswell (2010). (There is no consideration of 2 May 1999 in our study. It did not qualify as a top-30 day in 1999.) If we use the "constrained" STP as the calculation, the 3 May 1999 outbreak day has 2858 grid points exceeding the STP threshold of 1, whereas the 4 May 1999 outbreak day has 1523 grid points exceeding the threshold. Of course, this would mean that 4 May 1999 was subject to a miss much more so than 3 May 1999. However, both values are far above 1000, which is near the value of highest PSS (see new Fig. 4d). Interestingly, pseudo-trajectories have a mean distance of 767.9 km for 3 May 1999 and 983.8 km for 4 May 1999 – both of which easily exceed a threshold of ~500 km near the peak PSS (new Fig. 6d). The higher value for 4 May 1999 is a result of midlevel flow oriented slightly more parallel to the axis of high STP.

7) "It is emphasized that the preceding analysis with all 840 cases was not used in any way to conduct the training/testing analysis discussed below." This is obviously critical to having independence between the training and testing data.

Agreed. We wanted no such confusion when we switched topics to training/testing.

8) [End of Section 4]: This discussion seems to be leading to the notion that this information might be better posed in the form of probabilities. Can the authors comment on how a probabilistic approach might work?

For a forecasting project, a probabilistic approach is absolutely appropriate. As the ability to discriminate major from intermediate/marginal outbreak days is imperfect, a probabilistic approach is preferable. However, this project is not intended to be a forecasting study; instead, our objective is to show how imperfect the discrimination is. As a result, we have included no such discussion in our revised version.

One way to conduct a probabilistic approach would be to use frequencies of training data as a probabilistic diagnosis of an independent case. For example, given a grid point sum and a specified upper bound and lower bound of grid point deviations, training data could be used to obtain a frequency with which that grid point sum (within lower and upper bounds) corresponded with a major outbreak.

We performed a quick example. Using mean hypothetical storm distances within the fields of unconstrained STP ≥1 and a lower (upper) bound 50 km below (above) the value for a particular outbreak test case (frequencies obtained using the training set), we found a Brier score of 0.0804. The Brier score of climatology was 0.0938, leading to a Brier skill score of 0.1432. Although we did not perform confidence interval calculations for this example, these certainly can be computed as well.

A limitation to this probabilistic approach is sample size concerns, particularly with cases featuring large values of areal coverage. Frequencies of past cases may not correspond very well to an independent set because few examples exist. Also, the frequency with which days featuring large areal coverage of parameters resulting in no major outbreak may be larger, perhaps substantially so, as we only look at the top 30 days of each year. As a result, the approach above would be a conditional probability. Finally, the frequency approach becomes more complicated if multiple parameters are computed. Using kernel density

estimation to compute probabilities of multidimensional data may be appropriate here (as discussed in Doswell and Schultz 2006).

9) Section 5, second full paragraph: You discuss the issue that removing the water points (and other constraints) leads to less distinction between major and intermediate or marginal outbreak days. But doesn't this reflect the true task? This also relates to point #4 above. Why start with using the water points at all? I can see this if you are desiring to present the process, but otherwise I would take that out from the beginning.

The removal of water points did not lead to less distinction between major and intermediate or marginal outbreak days, and that is not what the manuscript says (said). Removing water points led to a reduction of a large number of intermediate cases with widespread "favorable regions" over water, but it also led to some major outbreak days having a substantially smaller number of grid points. The point of this discussion was to suggest that any effort to mitigate undesirable properties (such as substantially large areal coverage for intermediate days because of their proximity to water) leads to a side effect (a larger number of misses).

However, we have taken much of the discussion of water points out of the updated manuscript, particularly in Section 3a. We have also modified the wording in Section 5 accordingly.

[Minor comments omitted...]

Second review:

Recommendation: Accept.

General Comments: I find that the [authors] have been thoughtful in their responses: they have made a number of changes that improve the paper, and where they have elected to not make suggested changes, have provided valid reasons for not doing so. Accordingly, I recommend publication of the revised version. Thank you for the opportunity to participate in this process.

REVIEWER B (Michael C. Coniglio):

Initial Review:

Reviewer recommendation: Accept with major revisions (mostly dealing with the presentation).

General Comments: This paper summarizes work on the discrimination of severe weather outbreaks using the areal coverage of severe weather parameters that builds on recent work by Shafer and Doswell on the discrimination of tornado and primarily non-tornadic outbreaks. A tremendous amount of careful work that is certainly relevant to the severe storms forecasting community was performed by the authors. It's nice to see a rigorous attempt to quantify the subjective view that a good predictor of outbreaks is the amount of real estate that is expected to have sufficient CAPE/helicity/etc. on a given day.

However the paper is too long, overly dense in many places, and the main points of the paper I believe are hard to discern as a result. Although my specific knowledge of the statistical algorithms and techniques to quantify uncertainty is limited, I do have some experience applying exploratory techniques to model analyses, and I couldn't help get the feeling that the authors were trying to swat a fly with a sledgehammer, so to speak. I give some specific examples of this in my comments below.

Furthermore, I was surprised that little to no attention was given to the problems of the storm reports data base, given that the characteristics of the database have changed significantly during the period examined in your study (1960-2006).

A clarification: The period of study is actually 1979-2006, as NARR data are only available from 1 January 1979 to the present. The scores used to rank the indices, however, are based on storm reports from the period 1960-2006.

This particular problem is discussed at length in Shafer and Doswell (2010 – hereinafter SD10) and the many studies referenced within SD10 (e.g., Brooks et al. 2003; Doswell et al. 2005; Verbout et al. 2006). For example, attempts to account for the numerous well-known nonmeteorological artifacts in the data are discussed throughout that paper – particularly in terms of detrending the variables. We could add a lot of material on this topic to this paper, but it serves little purpose, not only because it is already discussed in SD10, but also because our findings from that paper show that the categorization of outbreaks as major or intermediate/marginal are quite robust to altering the weights used to rank the outbreaks (a method of accounting for the uncertainty of the severe reports). The only exceptions to this occur right around the threshold used to categorize the events, which is unsurprising – as the atmosphere provides a spectrum of events rather than distinct bins, and when several of the tornado variables were removed from the indices. This latter characteristic was desirable – as it was intended to upgrade significant events with few or no tornadoes. One method proposed in our paper to account for this uncertainty is to use various indices – not just one – and to vary the threshold used to distinguish major events from intermediate/marginal cases (e.g., new Fig. 7).

We have added a few sentences in Section 2 discussing the nonmeteorological artifacts in the dataset, providing the references listed above.

The first time anything related to this problem is mentioned in the paper is in the very last paragraph of the paper (statement #3).

This is not entirely true, though our wording was admittedly subtle. For example, in Section 2, we state:

> *... because the cases were not completely rank-invariant (leading to some cases being classified as major outbreaks for some indices and intermediate for others; see SD10), the value of 1 initially was selected to separate **major** outbreaks from **intermediate** or **marginal** outbreak days, where the latter included cases with scores below the value of -1. We by no means are stating that these values are the most appropriate, however. Indeed, selecting various thresholds to examine differences in diagnosing outbreak severity is appropriate.*

The fact that the cases are not completely rank-invariant suggests there is uncertainty in the severe reports (as well as a lack of outbreak ranking "truth," of course).

In Section 1, we have also added wording to refer to SD10 for discussion on limitations of the data used to rank the outbreaks.

I'm sure the authors are intimately aware of this problem, so the exclusion of any substantial discussion of the potential limitations of these uncertainties on this study is curious. Are the authors assuming that the problems with the storm reports are obvious to the readers of this journal and go without saying; or was there not a more careful consideration of the potential effects on this study?

This is not an either/or question. These were discussed at length in SD10.

Although I don't think it's a reason to stop publication of this paper, I wonder if the accuracy of the storm reports is sufficient to warrant such a sophisticated and complex treatment of the data. In other words, what would the uncertainty associated with the observations be relative to the uncertainty associated with different test data and statistical algorithms (Figs. 13-15)?

*Uncertainty with individual observations is certainly present, but with **outbreak categorization** is another matter. If the reports are detrended as in Doswell et al. (2006 – hereinafter D06) and SD10, outbreaks are ranked/categorized in a way that agrees with subjective notions and are relatively robust when modifying the weights of the variables (types of severe reports) used to rank the indices. The fact that the same tornado outbreaks seem to appear at the top of the rankings in D06 and the same severe weather outbreaks seem to appear at the top of the rankings in SD10 implies that the uncertainty of individual storm reports is not enough to change the classifications of the outbreaks substantially. The uncertainty of the **rankings** is*

high enough to consider prognosis of an outbreak's position in the rankings using observations or model simulations questionable, but this is not necessarily true for an outbreak's ***classification****.*

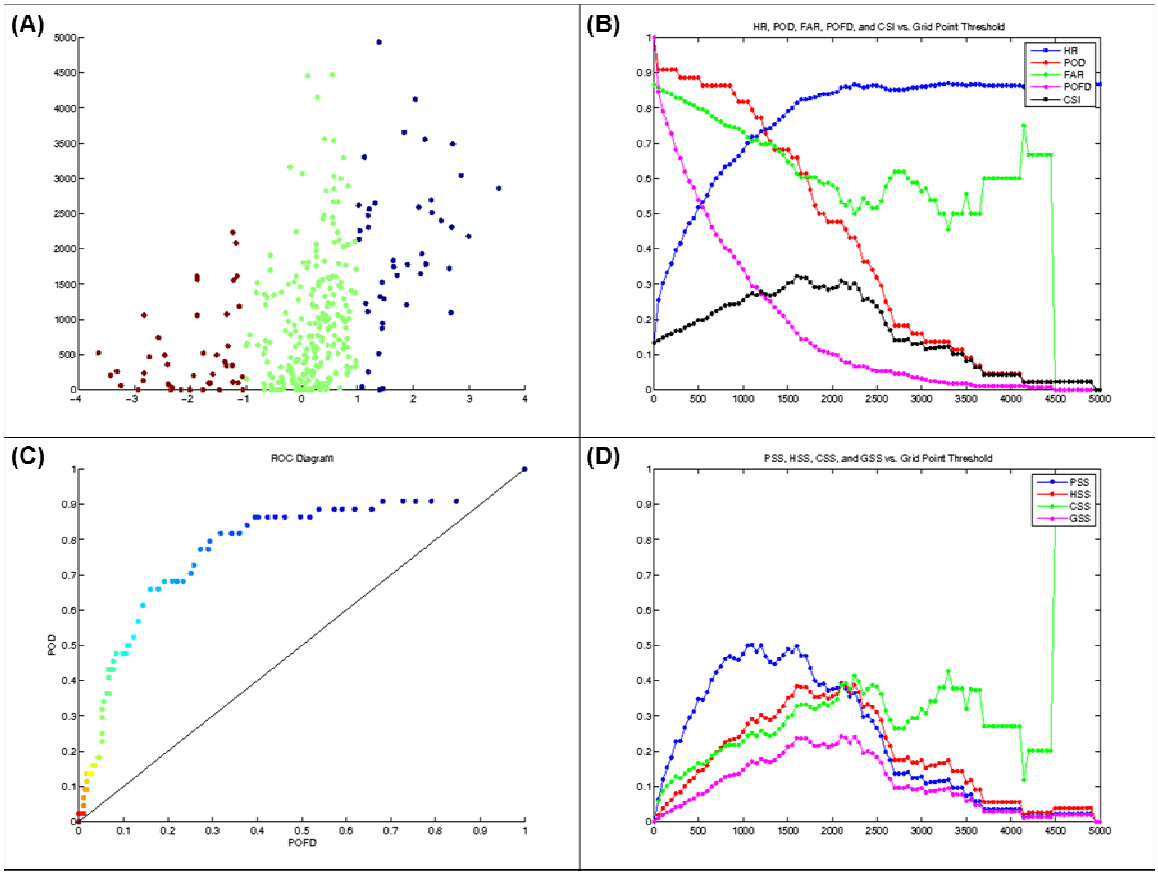
Would the observational uncertainty overwhelm anything meaningful that can be obtained from this analysis? Admittedly, quantifying uncertainty with the storm reports and analyses is a very difficult problem. But it's somewhat strange to see such sophisticated techniques used to quantify the uncertainty with the classifications, with little to no effort placed on quantifying the uncertainty in the observations themselves, or even any discussion on how the report vulgarities might affect interpretation of the results.

See above responses.

For example, instead of examining the effects of the different statistical algorithms used to make the classifications, why not look at the uncertainty associated with the periods over which the outbreak events are obtained? For example, I'd be very curious to see discrimination results on the data obtained from the report database after 1995, when a significant change in the reporting of severe convective winds was implemented (see Weiss 2002 Severe Storms [Conference] paper for more details).

Any effort to look at certain periods within the 1979-2006 period is doomed to have sample size issues. Because there are few major events to be considered with the whole dataset already – leading to sample size issues already addressed in the manuscript—limiting the period to 12 years is subject to substantial uncertainty. Any results we find here may not be associated with the differences in the period of record but simply with an inadequate number of samples from which to interpret results. See Doswell (2007) for more discussion.

Using the same methods as in (new) Fig. 4, here are the results when looking at cases only from 1996-2006 (330 cases):



Comparing the above to Fig. 4, the differences are actually quite minor. As expected, the differences are somewhat more noticeable for higher grid point thresholds (a result of small sample size of cases with relatively large areal coverage) – though even here, differences are minor. Given the techniques employed in D06 and SD10 to account for secular trends in the severe reports and the relatively minor differences we see when looking at shorter time periods (which are bound to have more substantial effects from small sample size), we are not compelled to add more material on this topic to the paper.

I've looked at SeverePlot output for a number of significant severe weather events over the same period examined in this study, and I suspect that some events classified as primarily tornado outbreaks or marginal outbreaks prior to 1995, or from earlier in the period, might have many more wind reports associated with the event if it occurred closer to 2006 and be classified differently as a result. Problems like these would seem to be a much larger source of uncertainty than the particular algorithm used to make the classifications, or the indices used to classify the outbreaks, the results of this study would be more amenable to application if issues like these were addressed much more so than in the current version of the paper.

This is why the variables used in the ranking indices were detrended – see SD10.

Finally, SD10 is cited frequently throughout the paper. I review the current paper from a perspective of someone who only did a cursory reading of SD10, and therefore might represent a reader that learns of the material from SD10 for the first time. I was frustrated by the many citations to SD10 and what I felt was an over-reliance on discussions from that paper to explain the methods and results from the current paper.

A cursory read of SD10 is probably not adequate, given the reviewer's comments. We have added wording in Section 1 to suggest a thorough read of SD10 is highly recommended. We understand the frustration, but as this study clearly is based on the work of SD10, frequent citations are to be expected and a more comprehensive read of that study is critically important.

This is especially troublesome in section 2. It might help to warn the reader up front that a reading of SD10 would help with the understanding of the current paper, but it still wouldn't solve the problem of what I feel are confusing explanations of the data and methods, and insufficient descriptions of the indices and various terms borrowed from SD10. I give some specific examples of this below.

We'll respond to these examples point-by-point below.

Specific Major Comments:

[Section 2]: The reasons for using several classification algorithms and three different techniques to interpret the statistical results are not clear. This is the sledgehammer I'm talking about. Was it necessary? If so, what did using so many different tools contribute to the results?

The use of multiple algorithms permits investigation of whether a specific algorithm appears to perform consistently better or worse than others, or if the results among all of them are consistent. This does not mean that one result is better than another – i.e., if one algorithm performs worse than all of the others, it is best not to use that algorithm; on the other hand, if all of them perform consistently, the results are relatively robust and the use of a single algorithm may be good enough – but using a single algorithm naively is inappropriate, especially if that algorithm performs noticeably worse than most others. We have added some wording in this paragraph to explain.

[End of Section 2]: Because "N15" and "N25" are not clearly defined up front, it took me a little time to realize that they are not predictive indices, but indices that are used to classify outbreaks. Some reorganization of Section 2 is needed to help with comments #7-10 above.

This is an excellent point. As a result, we have added a paragraph summarizing the development of indices in SD10, introducing the labels N0-N25 (per response to comment 5), including the dataset used to develop the indices (with added reference to Schaefer and Edwards 1999), and adding brief mentions of converting

the variables to standard normal (per response to comment 6) and detrending (per response to overall comments). We are hoping this additional information will make this description of the data and methods, particularly with regard to how the outbreak classifications were developed, clearer.

I'm not sure what the purpose of Figure 2 is and I don't find it to be particularly helpful. We don't know which case is which, and even if we did, it would be too hard to see it on this figure.

We have modified Fig. 2 substantially, per comments from a separate reviewer. We have used a scatter plot instead, labeled major outbreaks and intermediate/marginal outbreaks with different colors, and have modified some of the labeling. Note that the outbreak rank is on the x-axis, which has nothing to do with the date the cases occurred (there is no such bias in the rankings, as the severe reports were detrended). What should be noted is that there is a lot of scatter in the grid point sums, and the STP and CAPE/SREH figures have a noticeable upward trend with the major outbreaks, whereas the CAPE/BULK6 and CAPE/SRFL products have no such trend. Individual identification of cases is not possible for such a figure, aside from their corresponding ranking.

Also, I find it curious that the curve seems to bend upwards for the lower-numbered cases for STP and CAPE/SREH1. Any ideas why? If the lower-numbered cases are the ones closer to 1960, is this a problem with report biases or analysis biases in some way?

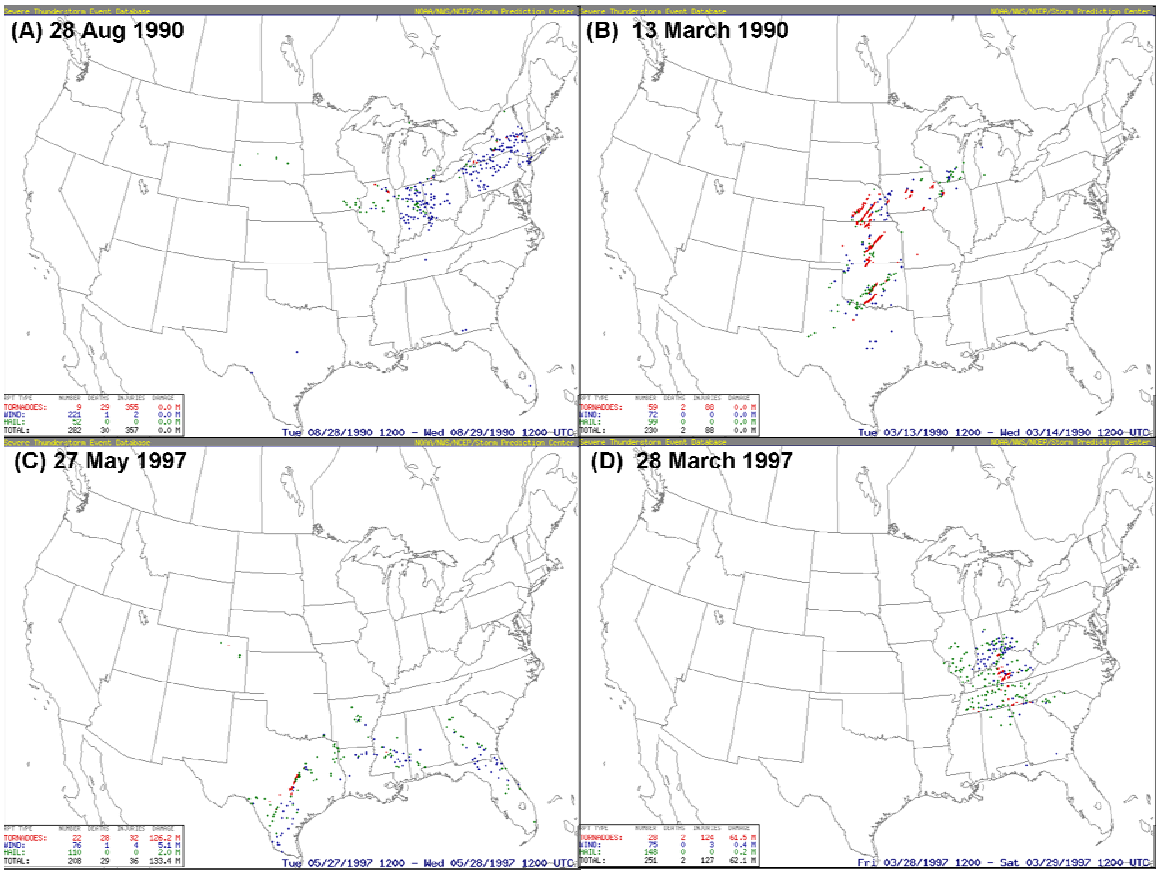
We are not sure what the reviewer is seeing here. Certainly with the modified figure, and even with the old figure, we do not see such an obvious trend (especially comparing this to the major outbreak days). In fact, the dots in the modified figure clearly cluster toward grid point sums of 0, as one would expect. There are several individual exceptions obviously, which illustrate the false alarm problem. We do note the CAPE bias with several of the marginal outbreak days in the text, which tend to be summer events with widespread regions of ample thermodynamic instability.

Too much space is wasted describing the effects of the "water problem." Why not just make this obviously necessary constraint from the start and be done with it? You'd thereby avoid introducing unnecessary jargon (constrained STP) and shorten the paper.

The point of this section was to discuss the data and why a naïve implementation of it without incorporating additional constraints is not wise. We have reduced, but have not completely eliminated, discussion of the water points for this reason. Additionally, inclusion of any constraint, even those that seem "obvious", resulted in an increased number of major outbreaks misdiagnosed as intermediate/marginal outbreak days, a point certainly worthy of discussion. "Unconstrained" variables do not consider water points, but they also contain no additional constraints. "Constrained" refers to the CAPE/SREH thresholds, which are open to more serious questions – see below. We have clarified the text accordingly.

Also, it's not clear why you'd want to constrain the analysis to the grid points with both sufficient CAPE and helicity >100. What about days like Plainfield, IL on 28 August 1990, or Jarrell, TX [27 May 1997], where the extreme CAPE seemed to compensate from a lack of low-level shear?

The reviewer curiously mentions two cases with isolated significant tornadoes with an otherwise relative lack of tornadoes across the country—both events are classified as "intermediate" by all 26 of the indices developed by SD10. We have attached the reports for these two cases below. As major severe weather outbreaks are primarily ***major tornado outbreaks****, this is not a compelling case for modifying the constraints.* ***Isolated high-impact tornadoes are not the focus of this investigation.*** *Refer to our discussion of the differences of outbreak versus storm discrimination in Section 1. We have also provided two examples of cases from the same year that do qualify as major severe weather outbreaks, for comparison.*



The comparison of the 1990 cases shows an outbreak with a large number of (significant) tornadoes (13 March 1990) compared to 28 Aug 1990, whereas the 1997 cases show an outbreak with more reports clustered in a smaller region on 28 March 1997 compared to the more isolated/scattered reports observed on 27 May 1997.

The reviewer also alludes to the limitation of adding constraints to the areal coverage variables – increasing the number of "misses" – for which we have also added some commentary in Section 3a. "Constrained" does not always produce "better" results – note the results of "constrained STP" versus "unconstrained pseudo-trajectories" (i.e., cf. [new] Figs. 4-6).

[Section 3b]: The authors mention that the CSS can pinpoint when forecasts no longer have value, but value in terms of what, exactly? What application does determining the point at which value is lost have to this particular study?

The Clayton skill score provides the ability to show the range of users who find value from a forecast (as shown in Wandishin and Brooks 2002), one of the characteristics considered to be desirable for verification (Murphy 1996). (Here, value would be assessed using a cost-loss model.) The reviewer is referred to these two manuscripts, mentioned in the text, for additional details. The point of using this statistic, as well as several others, is to investigate different properties of various contingency statistics – absolutely necessary in studies like this. If it is found that no users find value in the discrimination, then there would be no need to incorporate the method of discrimination in an operational setting.

I wonder if some of the major severe weather outbreak days that are classified incorrectly as intermediate/marginal outbreak days are events that occurred in the cool season in the eastern U.S. Models/forecasters seem to have trouble with low CAPE/high shear events.

In general, this was not the case. Most "misses" were primarily nontornadic outbreaks, or outbreaks spanning a relatively small area, as discussed in Section 4.

I was looking forward to some subjective explanations for the incorrect classifications to supplement the statistics-heavy section 3, but was disappointed with section 4. It is not clear how the examples provided in section 4 "provide valuable insight into the weaknesses of the methods."

For example, for the 9 November 2000 event, since it was a primarily nontornadic outbreak, and the N25 index identifies it as a major severe weather outbreak and the STP was low, where does the misclassification occur?

The N25 index classifies it as a major event, but the areal coverage of the STP was very low – hence, it was diagnosed (incorrectly) as an ***intermediate*** *event. Remember, the index is used to classify the event (as truth) and the areal coverage is used to determine what type of event it is (as a diagnosis). As the goal is to associate larger areal coverage with major events and lower areal coverage with intermediate/marginal events, this is a misclassification.*

Also, doesn't it go without saying that "when using parameters specifically developed to distinguish tornadic from nontornadic environments and/or storms, using the indices with all the tornado variables is more appropriate?" When would it ever be less appropriate?

We pose the reviewer an alternative question. What if the results suggested otherwise? To our knowledge, only two studies have focused on using STP (and other variables) to distinguish tornado outbreak environments from other types of events: Shafer et al. (2010; MWR) and this study. The former specifically examined tornado outbreaks and primarily nontornadic outbreaks. This study focuses on major severe weather outbreaks and less significant events. Even though the variable was designed to discriminate storm environments, that does not mean it would discriminate ***outbreak*** *environments—please refer to our comments on this in Section 1. Although we would expect this consistent behavior and subjective notions suggest this should be true, tests must be conducted to confirm it!*

Furthermore, a major severe weather outbreak is not always a tornado outbreak. This is mentioned in the text (e.g., Section 1).

What purpose is there to examining the ability of indices with fewer tornado variables to discriminate the tornadic from the nontornadic outbreaks?

That is not the focus of our study. We are discriminating major outbreaks from intermediate outbreaks. Although major severe weather outbreaks are predominantly major tornado outbreaks, ***that is not always the case****.*

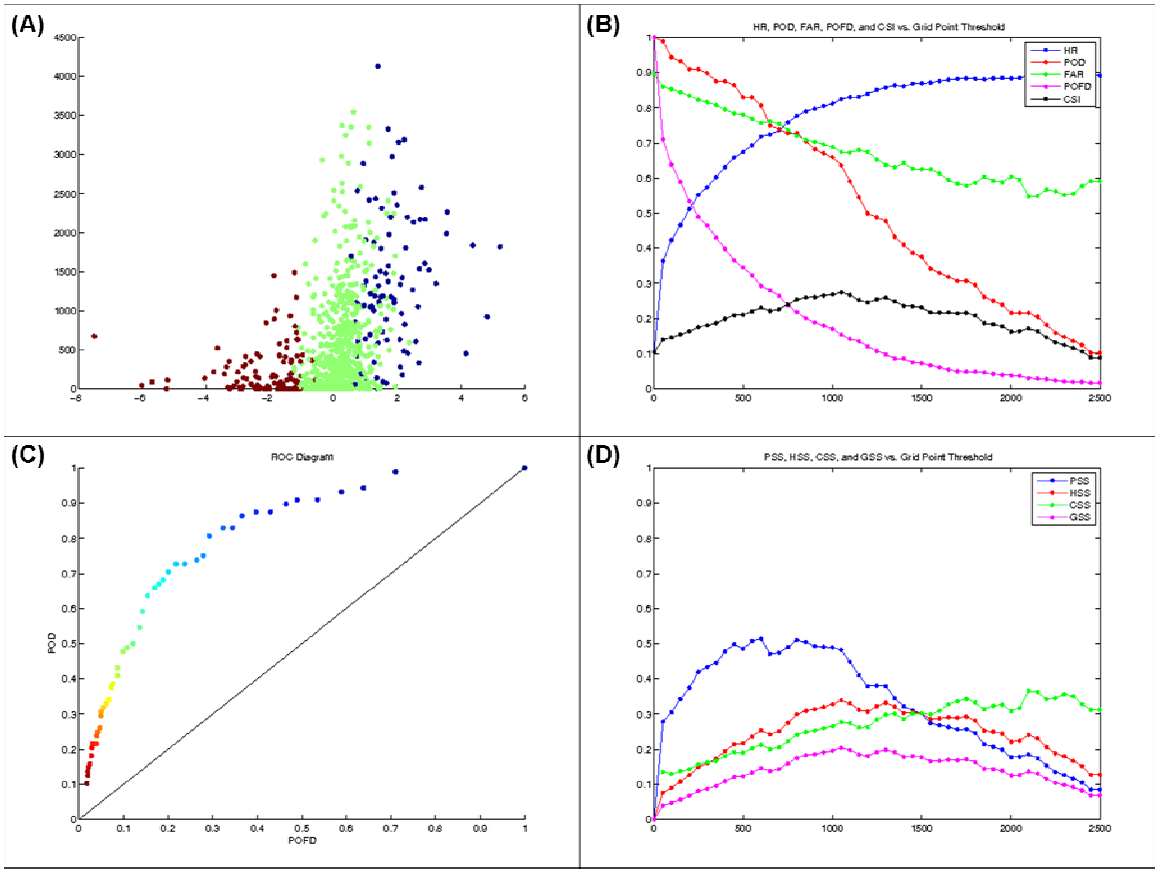
[Section 4]: Was a parameter that represents the capping inversion or convective inhibition included in the analysis? This seems like a simple way to factor in the potential for fewer storms than a large, unstable warm sector would suggest is favorable for severe weather.

We show the results below of using SBCIN as a constraint in addition to CAPE and SREH. In this example, the constraint for SBCIN was 50 J kg^{-1}, though we did test other thresholds and found no appreciable differences. The N15 index is used to classify outbreaks (e.g., compare to new Fig. 6).

The results suggest little difference whether CIN is included as a constraint or not, at least in general. Part of the problem with CIN may be associated with the fact that an accurate diagnosis of CIN requires high vertical resolution. Reanalysis data may not have sufficient vertical resolution to make it a useful parameter, even if it is physically relevant.

For the parallel midlevel flow-surface boundary map types, the problem is not necessarily with regions away from the boundary (the typically capped warm sector) but with the area in proximity to the boundary.

These areas tend to be large (elongated), and pseudo-trajectories within the region extend the length of the favorable region rather than crossing the much shorter width.



I don't see how the calculation of pseudo-trajectories within regions of large equivalent potential temperature gradients would help here.

Equivalent potential temperature gradients tend to be large near surface boundaries. If an enclosed region of these gradients (using some threshold) is calculated, the pseudo-trajectory can be drawn for each hypothetical storm at every grid point within this region. The result would be a mean value (distance). Longer distances imply midlevel flow oriented more parallel to the surface boundary. A threshold distance could be determined to identify these cases, signaling the possibility of this case as a "false alarm" based on the synoptic pattern in place.

Per suggestion from a separate reviewer, we have added Dial et al. (2010) as a reference here – as they investigate the use of additional variables and have removed the suggestion we previously offered.

Major outbreaks can still occur under a regime of midlevel wind vectors oriented parallel to surface boundaries- this is a common large-scale pattern for derechos. What is it about the meteorology that these set-ups tend not to be major outbreak days in your study, despite the large regions of severe weather parameters favorable for severe storms?
Because most major outbreaks, as defined in SD10 and our study, are not derechos – such a pattern would actually be ***unfavorable****. The tendency for these events to be (primarily) nontornadic, generally because of the resultant convective mode, is the reason these cases tend to be diagnosed incorrectly.*

[Minor comments omitted...]

Second review:

Recommendation: Accept.

General Comments: I just finished reading through the revised paper and have no further comments or concerns large enough to note.

REVIEWER C (Corey M. Mead):

Initial Review:

Recommendation: Accept with minor revisions.

The "Rasmussen table" below summarizes my evaluation of this study. General and specific comments follow the table.

Criterion	**Satisfied**	**Deficient, but can be remedied**	**Deficient; *cannot* be remedied by modifying the paper**	**Deficient, *not known* if it can be remedied by modifying the paper**
1. Does the paper fit within the stated scope of the journal?	X			
2. Does the paper 1) identify a gap in scientific knowledge that requires further examination; 2) repeat another study to verify its findings; or 3) add new knowledge to the overall body of scientific understanding?	X			
3 Is the paper free of errors in logic?	X			
4. Do the conclusions follow from the evidence?	X			
5. Are alternative explanations explored as appropriate?	X			
6. Is uncertainty quantified?	X			
7. Is previous work and current understanding represented correctly?	X			
8. Is information conveyed clearly enough to be understood by the typical reader?	X			

General Comments: This paper explores the ability to diagnostically discriminate major severe weather outbreaks from less significant events using an objective areal coverage approach. This areal coverage approach utilizes grid point summation of meteorologically favorable parameters and the calculation of backward and forward "pseudo-trajectories" of hypothetical storms within this region of favorable parameters to determine outbreak severity.

Though the particular method used in this work has some shortcomings, the prospect of potentially applying a similar type of approach in a prognostic sense is exciting. Indeed, being able to differentiate high impact, widespread and significant severe weather episodes from those of lesser consequence is an overarching goal of operational forecasters. And as such, this type of research is well warranted and worthy of publication in EJSSM.

Substantive Comments:

1) In Section 1, I would like to comment on the author's assertion that since operational models are not able to explicitly resolve tornadoes, their value then lies in "the forecast of meteorological parameters." There is no doubt that evaluating the NWP forecasts of meteorological parameters is certainly a large component of present day severe weather forecasting. However, I would also add that with the advent of high resolution (~5 km or less) convection allowing models, forecasters now have more detailed insight into possible convective modes which is a critical aspect of determining outbreak type and severity. In fact, the manuscript acknowledges this by stating "The convective mode has profound implications on the type of severe weather that is observed…".

Good point. We have included some wording here to indicate this.

Further along in Section 1, I found myself a bit confused by the manuscript apparently making the distinction between many studies which "investigated the utility of a variety of severe weather parameters to distinguish storm modes, significance of severe weather, or types of severe weather" and the present work which "uses various severe weather parameters in the identification of a particular type of outbreak." I suspect you are differentiating between research efforts on individual storm environments versus outbreak characteristics. But, isn't the former a subset or determinant of the latter?

This is true, but it does not equate the two types of research. Consider an event in which a single, significant tornado occurs on an otherwise nontornadic day, whereas a separate event has dozens of significant tornadoes. The storm environment of the lone tornado in the former event could be quite similar to the storm environment of any of the significant tornadoes occurring with the latter event. The synoptic and mesoscale environments, however, may be very different in the regions in which severe weather occurred – which is where the idea of areal coverage distinguishing outbreak events comes from, of course.

2) In section 2, it was noted that the NARR dataset was used which features 32 km horizontal grid spacing with 45 vertical layers which "provided a convenient means of analyzing mesoscale fields (of meteorological parameters) for a large number of outbreak days". This is in contrast to the companion work referenced in this manuscript (specifically, S09 and M09) which initialized high-resolution WRF model forecasts with the NCEP-NCAR reanalysis datasets having a horizontal grid spacing of 2.5 degrees latitude by longitude and 17 vertical levels. The stated intent was to determine the extent to which synoptic-scale processes influence outbreak type. The results of S09 and M09 indicate that even with the coarse nature of the input data, the WRF forecasts were skillful in discriminating outbreak type out to three days.

Given the success of the above-mentioned methodology, why has this study adopted a different approach? Specifically, why was there a shift in focus from synoptic-scale input data to that of mesoscale? For consistency, couldn't a similar method to S09 and M09 have been employed whereby 00-hr WRF forecasts (valid at the time of the outbreak) were evaluated with the model initialized by the 2.5 degree NCEP-NCAR reanalysis data? These forecasts would use S09's Domain 3 (18 km) which would be expanded from a 121×121 grid to 300×200. This would allow for an "apples to apples" comparison with the previous model simulations. Moreover, all of the cases in SD10 (1410 days—1960-2006) could then be used, affording you larger training and testing sets.

There are three important differences between this study and those of S09 and M09. The first is those two studies were modeling studies – this one is not. A true "companion piece" would be to run 1410 model

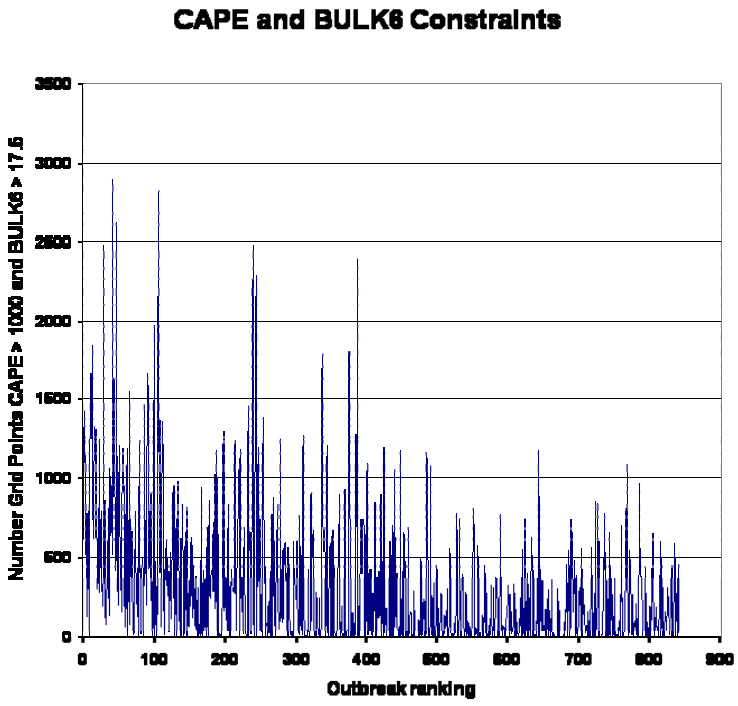
simulations for 1-day, 2-day, and 3-day forecasts of each of the cases. Obviously, that's computationally intractable.

The second is that the type of outbreak discrimination is different. In the former studies, we sought to determine the ability of the WRF to discriminate tornado and nontornadic outbreaks. Here, we are attempting to distinguish major outbreaks from intermediate/minor events – a more difficult task. Although we hypothesized that synoptic-scale processes could be used to distinguish tornado/nontornadic outbreaks, we are much less certain of that for discriminating major and intermediate/marginal outbreaks.

Finally, our goal here is not to determine if synoptic-scale processes could distinguish these events. We want to know if analyses of meteorological fields can distinguish major outbreaks from less significant events. A primary motivation was to consider specifically if enhanced resolution would prove useful at this somewhat more sophisticated discrimination task. As a first step, if the initial, analyzed fields were of little or no value, there would be no point to going on to consider model forecast fields.

3) In section 3a, I have a couple of comments and a question. First, it was stated that "some areal coverage parameters showed negligible capability distinguishing outbreak days (e.g., the product of SBCAPE and 0-6 km bulk shear exceeding 60 000…". As opposed to the product of SBCAPE and 0-6 km shear, I would suggest a grid point check of SBCAPE ≥1000 J kg^{-1} and 0-6 km shear ≥17.5 or 20 m s^{-1}. That way, the baseline threshold is an instability/shear combination that is supportive of supercells and other organized modes of convection. As it stands, the SBCAPE term can dominate the product, highlighting grids in which the 0-6 km shear might be rather weak.

We have attached this figure below.



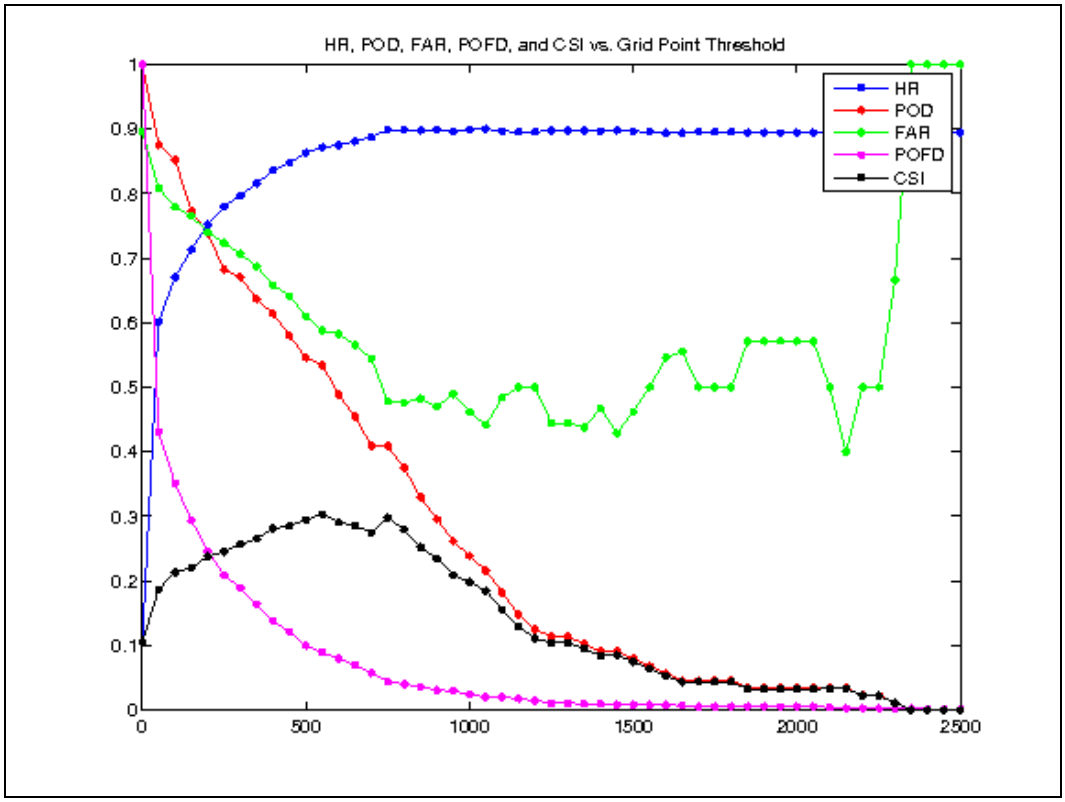
Indeed, a signal for the major outbreaks is seen here, similar to that of CAPE/SREH and STP. However, false alarms remain a problem, suggesting this variable is no better than those discussed in the text. (Also, the point of showing the product of SBCAPE and 0-6 km bulk shear was that there were combinations of parameters that performed poorly – our examples in the text were by no means the only combinations tested, as a footnote in Section 3a explains.)

I am curious as to what the cause(s) for the noncontiguous regions of parameters exceeding thresholds were in the majority of cases; perhaps the effects of model convective precipitation? Going into this article, I had anticipated that the major outbreak days might have large, contiguous areas of favorable parameters when compared to lesser events.

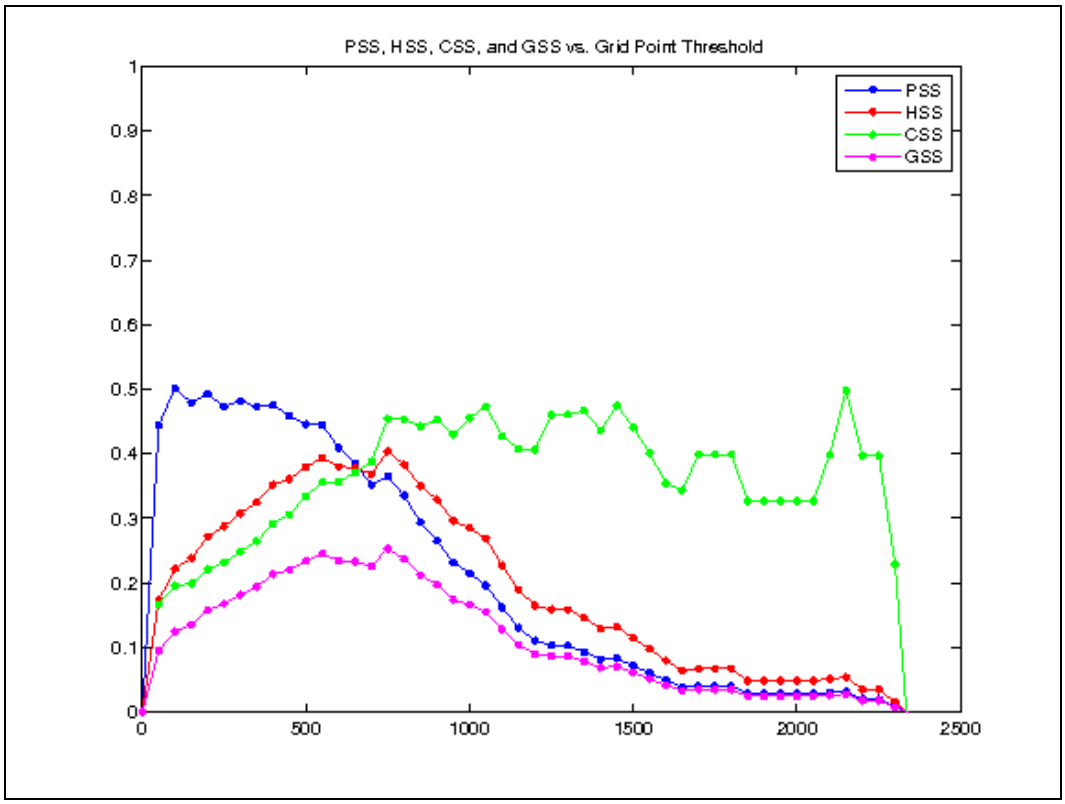
The manuscript states that "[v]ery few cases exhibited noncontiguous regions of parameters exceeding thresholds". Those cases that did exhibit these noncontiguous regions were primarily intermediate/marginal days, not major outbreak days, and generally were associated with multiple synoptic-scale systems (e.g., a shortwave trough approaching the High Plains and a separate shortwave trough nearing the East Coast). Additionally, model convective precipitation is not involved in our work. We are looking at analysis fields—not model forecast fields.

Due to the large values of CAPE dominating the EHI, SCP and STP, a constraint of SBCAPE ≥1000 J kg^{-1} and 0-1 km SREH ≥100 m^2 s^{-2} was used in some of the calculations of areal coverage thresholds. As mentioned above, because of the way these indices are constructed, the large CAPE can mask environments deficient in sufficient deep-layer shear for supercells and other organized types of convection. As such, I would recommend modifying the constraint to SBCAPE ≥1000 J kg^{-1}, 0-1 km SREH ≥100 m^2 s^{-2} **and** 0-6 km shear [magnitude] ≥17.5 or 20 m s^{-1}.

Two comments: (1) STP involves all three of these parameters already (Thompson et al. 2003), and (2) the addition of deep-layer bulk shear could come at a cost of increasing the number of misses (remembering that each additional constraint is expected to increase the number of major outbreaks misclassified – see Section 3a). We have attached the modified constrained STP contingency statistics below.



The results show that grid point thresholds are very small for this particular version of constrained STP (i.e., with the deep-layer shear constraint included). The false alarm problem remains, as FAR>POD starting at low thresholds. The N15 index is used in the figure above.



As expected, skill scores are generally higher at lower thresholds. The peak of PSS is comparable using the extra constraint (cf. new Fig. 4), showing little advantage of using this modified statistic.

4) In Section 4, it was noted that, "a common type of case that was misclassified as a major severe weather outbreak featured midlevel flow oriented nearly parallel to a surface boundary." Further in the discussion it is stated, "the inclusion of new parameters that describe these characteristics could be highly beneficial in discriminating these cases correctly." And finally in Section 5, it was noted that, "events in which midlevel wind vectors were oriented parallel to a surface boundary were commonly misclassified as a major outbreak…". The authors are referred to Dial et al. (2010) "Short-Term Convective Mode Evolution along Synoptic Boundaries" (http://journals.ametsoc.org/doi/pdf/10.1175/2010WAF2222315.1) for possible ways of addressing the orientation of the deep-layer wind field to the initiating boundary.

We had recently become aware of this article when working on a separate manuscript and were going to include in the revised version of this one anyway, as it is quite relevant to the discussion in Section 4. We have done so, and we will consider some of these parameters in future work.

5) In section 5, several possible reasons are listed to explain false alarm cases where the large-scale environment appeared quite similar to that of major outbreak days. One possible cause not listed and not likely to be found given your current approach is dominant convective mode, which can have a profound impact on how a given severe weather event will unfold. To this end, what if one were to consider conducting WRF simulations where a model field like updraft helicity (UH) was used as a surrogate to supercell/discrete mode? A similar grid point summation could be done where UH exceeded a specific threshold. This information could then be used in conjunction with your grid point summation of meteorologically favorable parameters to help determine outbreak type and severity. This is certainly beyond the scope of this work, but something that might be interesting to consider in future research.

This is obviously something that deserves consideration in future work, though there are challenges and limitations inherent in such research. Any modeling study is limited computationally, and simulating 840

cases is a very time-consuming task. Furthermore, it is quite possible the model would not develop simulated convection in a subset of cases.

Convective mode is a consideration, of course, and is alluded to when mentioning the cases involving midlevel flow oriented parallel to a surface boundary (which, in the meridional case, typically develops a squall line). As we mention in Section 1, meteorological fields of parameters are not likely to be associated strongly with convective mode, as previous studies have discovered.

6) Figures. I'm wondering if perhaps the utilization of the N15 or N25 indices to highlight the major outbreaks (i.e., values ≥1) with a different color in figures 2, 5 and 6 might give the reader a better perspective on how the various discriminates performed. I would suggest the use of the N15 index since it more heavily weights the tornado-dominant outbreaks, which are suggested to be the primary, major outbreak type.

Based on other reviewers' comments, former Figs. 5 and 6 were removed entirely. Fig. 1 (the N25 index) is also available for comparison with Fig. 2. We very much like the suggestion the reviewer provides, and we have incorporated in Fig. 2. The N15 index is incorporated, as suggested.

Why was the unconstrained STP used for Fig. 6? Since you mention that "the unconstrained initial calculations of favorable areas were susceptible to a number of undesirable characteristics," I would think you would want to use the constrained STP results in the figure.

In our first version, we failed to illustrate the side effects of adding constraints to the areal coverage calculations. We have added some material in Section 3a to explain these drawbacks. Specifically, for every constraint added, the number of "misses" of the major outbreaks increased. So, despite the obvious bias of high-CAPE environments in weak shear, adding the CAPE and SREH constraints typically lowered the areal coverage of cases with relatively low CAPE or relatively low shear. Even though using constraints led to elimination of cases with clearly unfavorable conditions, it also led to an elimination of some cases that actually produced significant outbreaks of severe weather. Thus, we decided to use the unconstrained and constrained variables in subsequent analyses.

[Minor comments omitted...]

Second review:

Recommendation: Accept.

General Comments: The authors have satisfactorily addressed all of the comments and suggestions offered in the initial review process. I want to commend them on their hard work and valuable contribution to EJSSM.