

An Assessment of Areal Coverage of Severe Weather Parameters for Severe Weather Outbreak Diagnosis

CHAD M. SHAFER

Department of Earth Sciences, University of South Alabama, Mobile, Alabama, and School of Meteorology, and Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma

ANDREW E. MERCER

Northern Gulf Institute, Mississippi State University, Starkville, Mississippi

MICHAEL B. RICHMAN AND LANCE M. LESLIE

School of Meteorology, and Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma

CHARLES A. DOSWELL III

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

(Manuscript received 18 November 2011, in final form 9 March 2012)

ABSTRACT

The areal extent of severe weather parameters favorable for significant severe weather is evaluated as a means of identifying major severe weather outbreaks. The first areal coverage method uses kernel density estimation (KDE) to identify severe weather outbreak locations. A selected severe weather parameter value is computed at each grid point within the region identified by KDE. The average, median, or sum value is used to diagnose the event's severity. The second areal coverage method finds the largest contiguous region where a severe weather parameter exceeds a specified threshold that intersects the KDE region. The severe weather parameter values at grid points within the parameter exceedance region are computed, with the average, median, or sum value used to diagnose the event's severity. A total of 4057 severe weather outbreaks from 1979 to 2008 are analyzed. An event is considered a major outbreak if it exceeds a selected ranking index score (developed in previous work), and is a minor event otherwise. The areal coverage method is also compared to Storm Prediction Center (SPC) day-1 convective outlooks from 2003 to 2008. Comparisons of the SPC forecasts and areal coverage diagnoses indicate the areal coverage methods have similar skill to SPC convective outlooks in discriminating major and minor severe weather outbreaks. Despite a seemingly large sample size, the rare-events nature of the dataset leads to sample size sensitivities. Nevertheless, the findings of this study suggest that areal coverage should be tested in a forecasting environment as a means of providing guidance in future outbreak scenarios.

1. Introduction

The identification of major severe weather outbreaks has been a primary objective of severe weather forecasting for decades (Schaefer 1986; Johns and Doswell 1992; Doswell et al. 1993; Doswell 2007a). As these events typically are responsible for a substantial portion of high impact severe weather observed in a given year (e.g.,

Doswell et al. 2006, hereafter D06; Verbout et al. 2006; Brotzge and Erickson 2009, 2010), methods that can discriminate major severe weather outbreaks from less significant events need to be demonstrated. As mentioned in Shafer et al. (2010a, hereafter S10a), surprisingly little research has been conducted on the discrimination of severe weather outbreaks based on their relative severity.¹ Most outbreak studies typically are of individual cases

Corresponding author address: Chad Shafer, Dept. of Earth Sciences, Rm. 136, University of South Alabama, 5871 USA Dr. N., Mobile, AL 36688-0002.
E-mail: cmshafer@usouthal.edu

¹ On the other hand, numerous studies have attempted to discriminate convective mode or observed severe weather for individual storms.

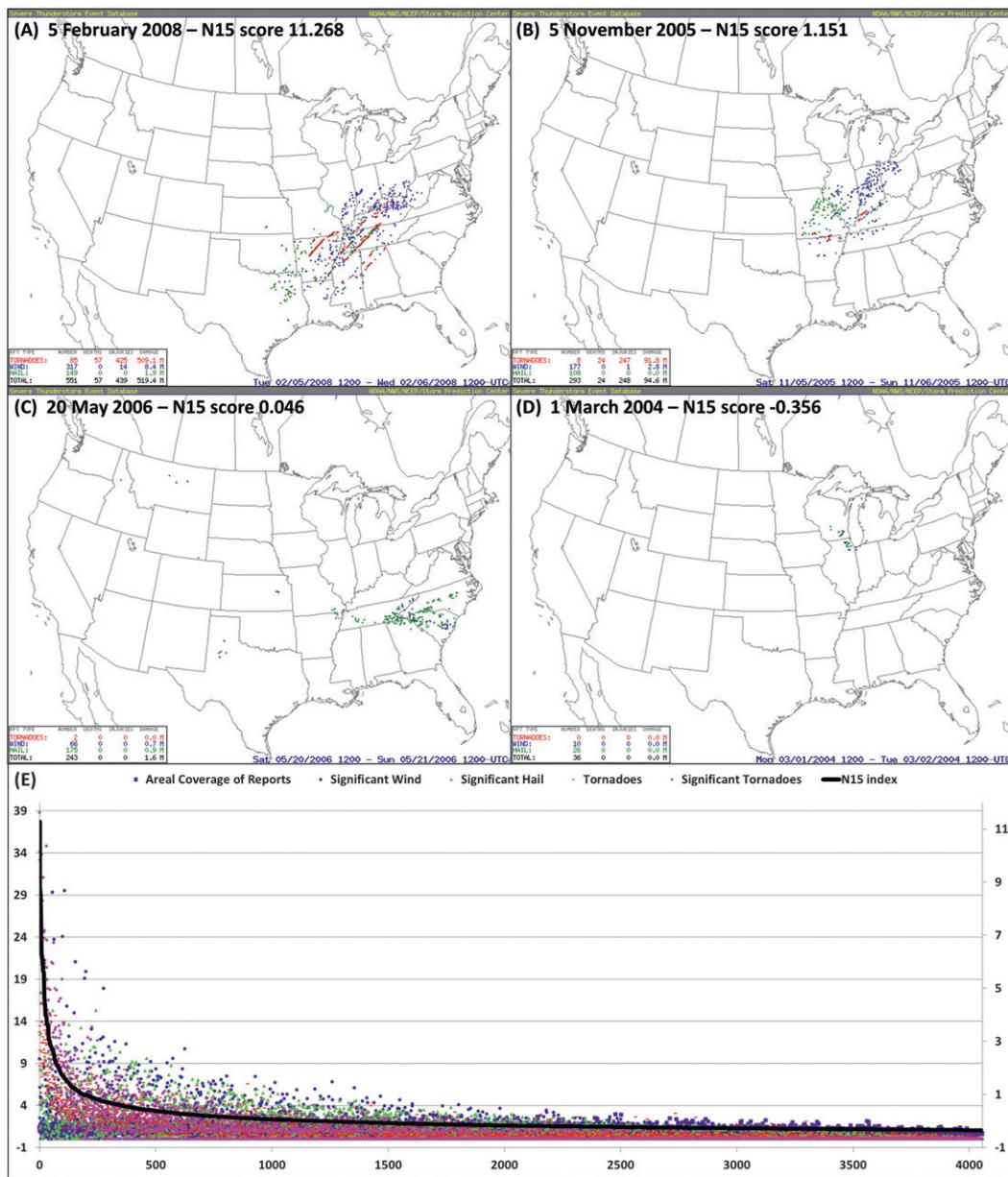


FIG. 1. Severe reports (tornadoes in red, hail in green, winds in blue) from 1200 UTC on the nominal date to 1200 UTC the following day, on (a) 5 Feb 2008, (b) 11 Nov 2005, (c) 20 May 2006, and (d) 1 Mar 2004. N15 ranking index scores (described in text) are indicated. (e) Detrended and standardized severe weather report variables and standardized areal extent of the KDE region (left y axis) for each outbreak as a function of the N15 index score (right y axis), as computed in Shafer and Doswell (2011).

(e.g., Thompson and Edwards 2000; Corfidi et al. 2010) or include a small sample of cases (e.g., Stensrud et al. 1997); hence, the findings may not be generalizable.

In recent years, several studies have introduced reproducible, quasi-objective methods to identify severe weather outbreaks of specific types (D06; Verbout et al. 2006) or of any type (Shafer and Doswell 2010, hereafter SD10; Shafer and Doswell 2011, hereafter SD11) using

observations of severe weather. Although the archive of severe reports is known to have nonmeteorological artifacts (Brooks et al. 2003a; Doswell et al. 2005; D06; SD10), these studies have demonstrated successfully the capability of the severe reports to rank and classify these events in agreement with subjective notions. However, the imperfect observations of severe weather, the limited number of severe weather report variables that are

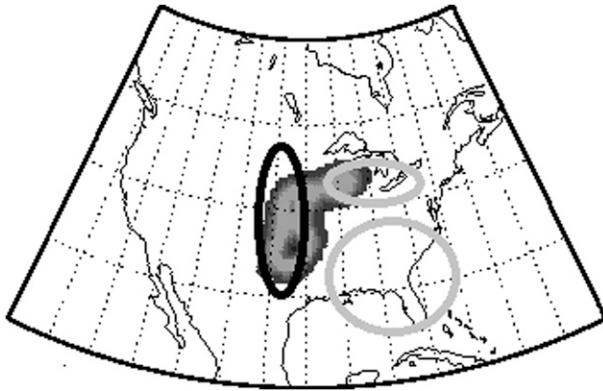


FIG. 2. Illustration of the intersect method. The shaded region indicates the KDE region for a hypothetical event. Each ellipse indicates a hypothetical region in which a selected severe weather parameter exceeds a specified threshold. The black ellipse outline is the region that would be selected for the intersect method in this case, as it is the largest region that also intersects the KDE region.

archived, and the nebulous notion of what comprises a severe weather outbreak suggest that distinction of severe weather outbreaks as major or minor events is subject to uncertainty (see S10a and SD11 for more details).

TABLE 1. Statistical discrimination methods and identification numbers as indicated in the relevant figures.

Statistical method	ID No.
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

Previous studies (e.g., Mercer et al. 2009; Shafer et al. 2009, 2010b) have limited the scope to major tornado and primarily nontornadic outbreaks, in which prototypical cases of each type were identified by D06. However, most severe weather outbreaks fall in between these two categories (SD10; SD11), which prompted additional investigation into the ability of reanalysis data to distinguish the most significant severe weather outbreaks

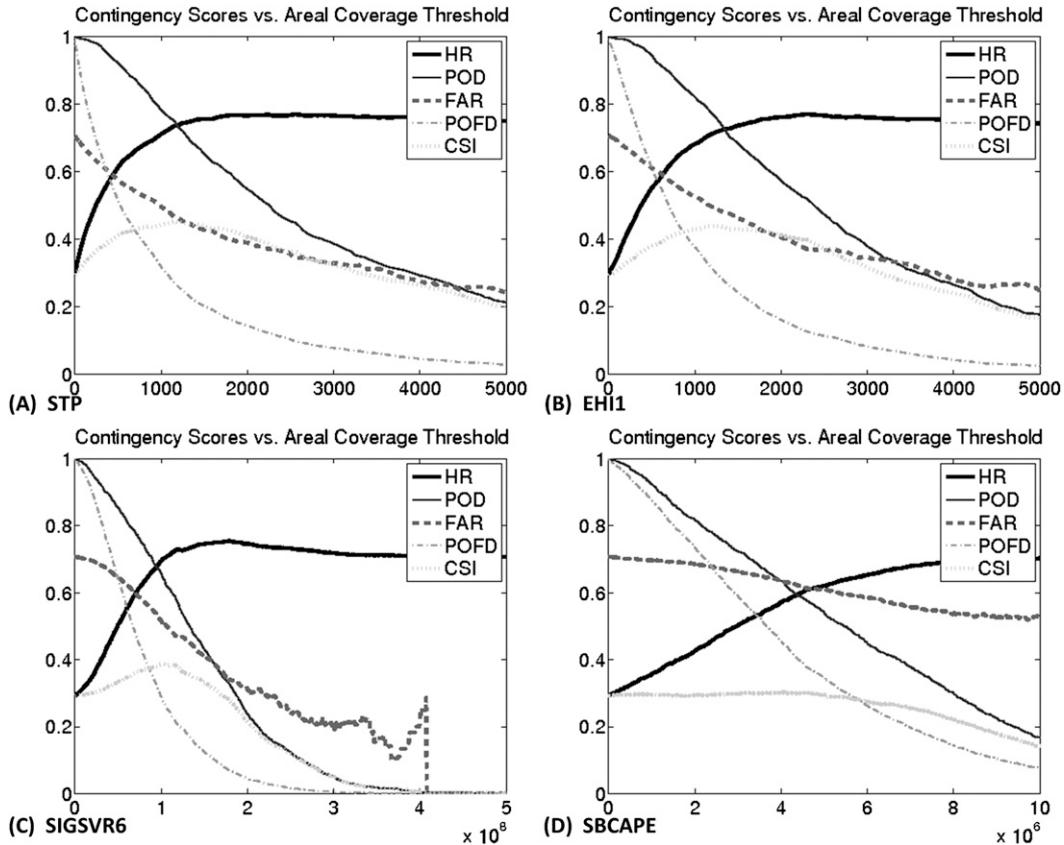


FIG. 3. Contingency statistics (y axis; labeled) as a function of areal coverage threshold (x axis) for the sum values of (a) STP, (b) EHI1, (c) SIGSVR6, and (d) SBCAPE using the KDE method. Cases are identified as major severe weather outbreaks if the N15 index scores equal or exceed the value of 0 and minor outbreaks (null cases) otherwise. See Fig. 1 for the N15 index scores of each outbreak as a function of its rank.

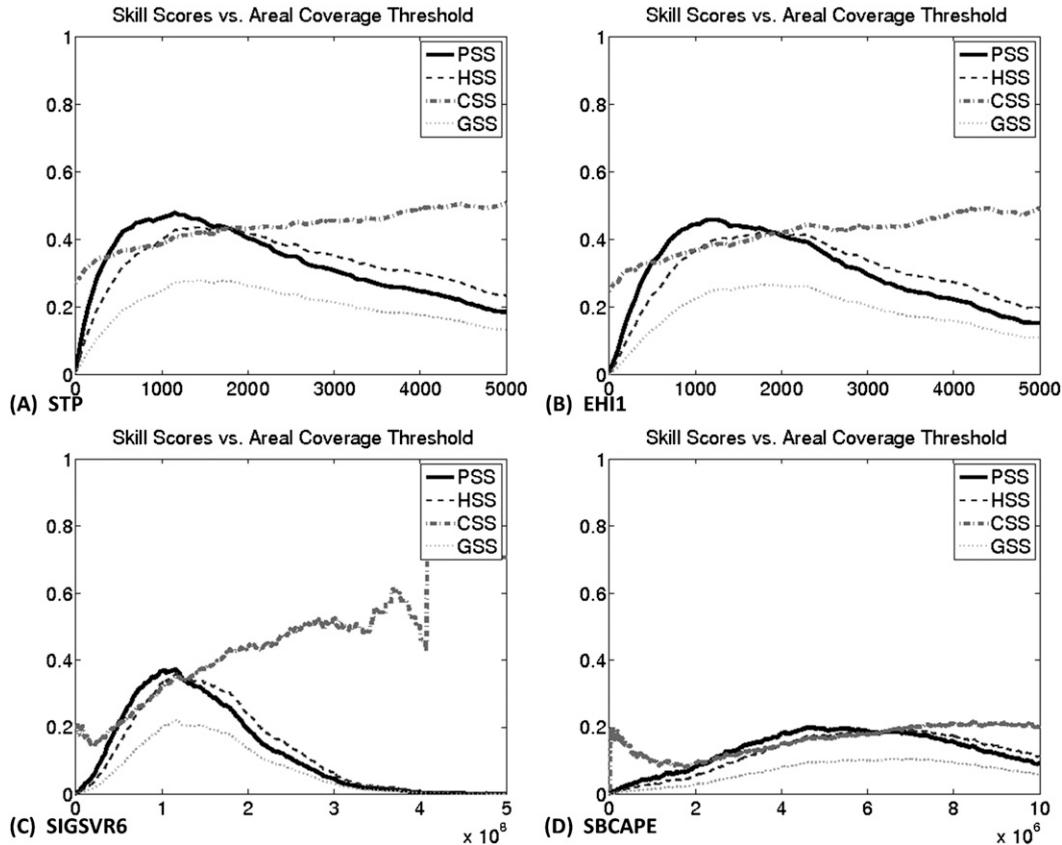


FIG. 4. As in Fig. 3, but with skill scores.

from intermediate and marginal events (S10a). S10a used a necessarily arbitrary threshold to classify outbreaks as major or minor, based on the scores of the multivariate index used by SD10 to rank the events. Any study attempting to classify severe weather outbreaks inevitably must choose arbitrary thresholds/criteria to distinguish the events.

A limitation of outbreak discrimination studies is the small sample of cases available for study (Doswell 2007b). S10a included 840 cases for outbreak discrimination and discovered sample size issues remained, even though this sample size far exceeded those of predecessor outbreak studies. Moreover, the sample used by S10a included cases with geographically dispersed severe reports and/or multiple spatially distinct clusters of reports. As neither of these characteristics agrees with preconceived notions of what constitutes an outbreak, SD11 introduced a new method to identify, rank, and classify severe weather outbreaks using two-dimensional kernel density estimation (KDE; Bowman and Azzalini 1997) with Gaussian kernel functions. Outbreaks were identified based on the number and density of the severe reports. A region where probability density function (PDF) threshold criteria

were exceeded (based on the clustering of the severe reports) was determined to be the area associated with a single outbreak. Then, regions in which a threshold number or density of reports was not exceeded were excluded automatically. Not only did this method eliminate cases with large geographic scatter and include distinct severe weather report clusters as separate events on the same day, but it also greatly increased the sample size of cases from that of previous work (~ 1400 in SD10 to ~ 6000 in SD11). Classification of cases as major or minor events in this study is based on the outbreak ranking indices developed by SD11. The reader is referred to that study for details regarding the characteristics and limitations of these indices.

As operational models do not resolve tornadoes explicitly and are not expected to do so soon, outbreak (and storm) discrimination studies necessarily have focused on using meteorological covariates (Brown and Murphy 1996). That is, they have used meteorological variables associated with severe weather as a means of diagnosing and/or predicting outbreak type. For the outbreak discrimination studies mentioned above, two

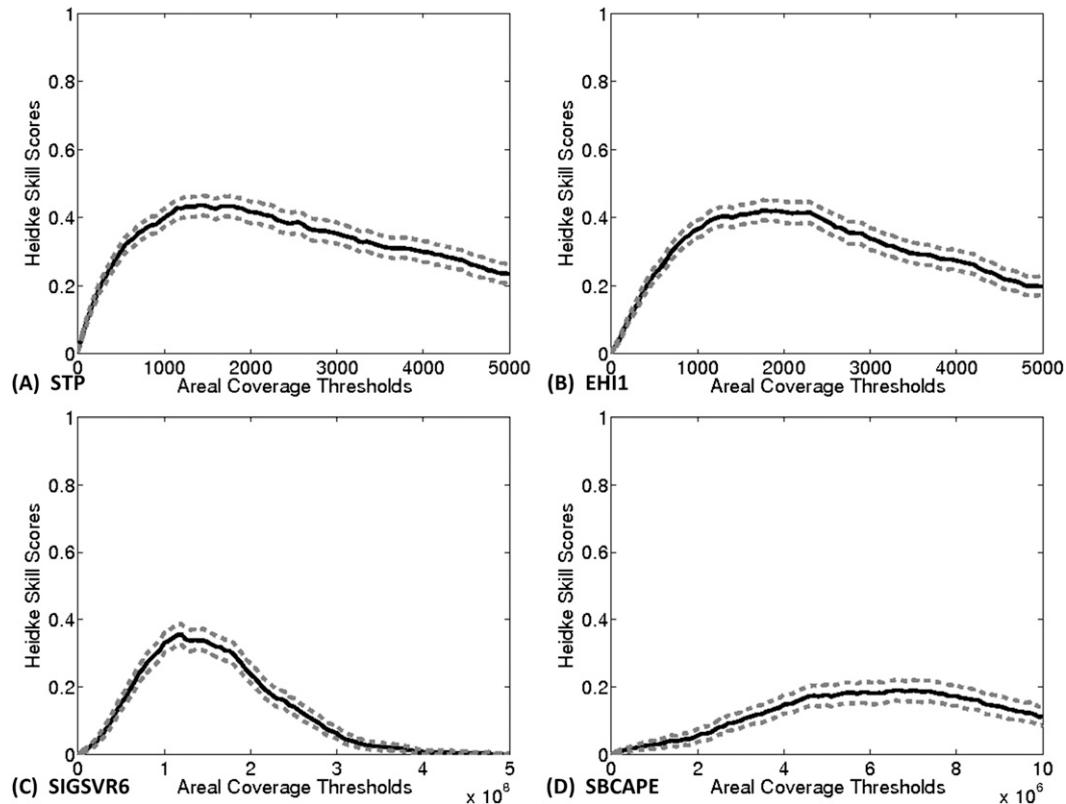


FIG. 5. (a)–(d) Bootstrap medians (black line) and 95% confidence intervals (dashed gray lines) of HSSs of the 4057 outbreaks from 1979 to 2008 as a function of areal coverage threshold for the variables labeled (the same as those in Fig. 3). The N15 index score of zero is used to classify events as major or minor outbreaks.

methods have been introduced. The first method uses principal component analysis of gridded fields as a means of data mining, and the principal component scores are used to train and test statistical models. This method was introduced by Mercer et al. (2009) and Shafer et al. (2010b) to discriminate tornadic and nontornadic outbreaks using mesoscale model output. The second method uses information regarding the areal coverage and magnitudes of a severe weather parameter (or a combination of parameters) as a means of distinguishing outbreaks. Areal coverage describes the geographic extent to which severe weather parameters favorable for significant severe weather exist. Areal coverage was used subjectively by Shafer et al. (2009) and was incorporated in an objective manner by S10a. The areal coverage and magnitudes of severe weather parameters have been used in past studies to identify the potential for significant severe weather (e.g., Brooks et al. 2003b; Hamill et al. 2005). As the areal coverage and principal component methods have been shown to be similarly skillful (cf. Mercer et al. 2009; Shafer et al. 2009, 2010b) and the areal coverage method is simpler to compute and interpret, it will be used as the sole means of outbreak discrimination herein.

The objectives of this study are 1) to test areal coverage as a means of diagnosing major severe weather outbreaks from less significant events (as determined by the so-called outbreak ranking index threshold introduced by SD11) using reanalysis data valid near the median time of the outbreaks, 2) to determine the outbreak ranking index threshold in which a particular severe weather diagnostic variable discriminates major and minor outbreaks with the highest skill, 3) to identify a set of meteorological variables that appear to be most capable of discriminating outbreaks, 4) to assess any sample size and temporal limitations in the dataset, and 5) to compare the areal coverage method to operational short-term forecasts of severe weather outbreaks (i.e., Storm Prediction Center convective outlooks). As the objective techniques developed in these past studies have not been compared to current operational forecasts attempting to determine the overall severity of the events to this point, the fifth objective will be a primary focus of the present work. However, a caveat is that the areal coverage technique is tested diagnostically, whereas Storm Prediction Center (SPC) convective outlooks are short-term

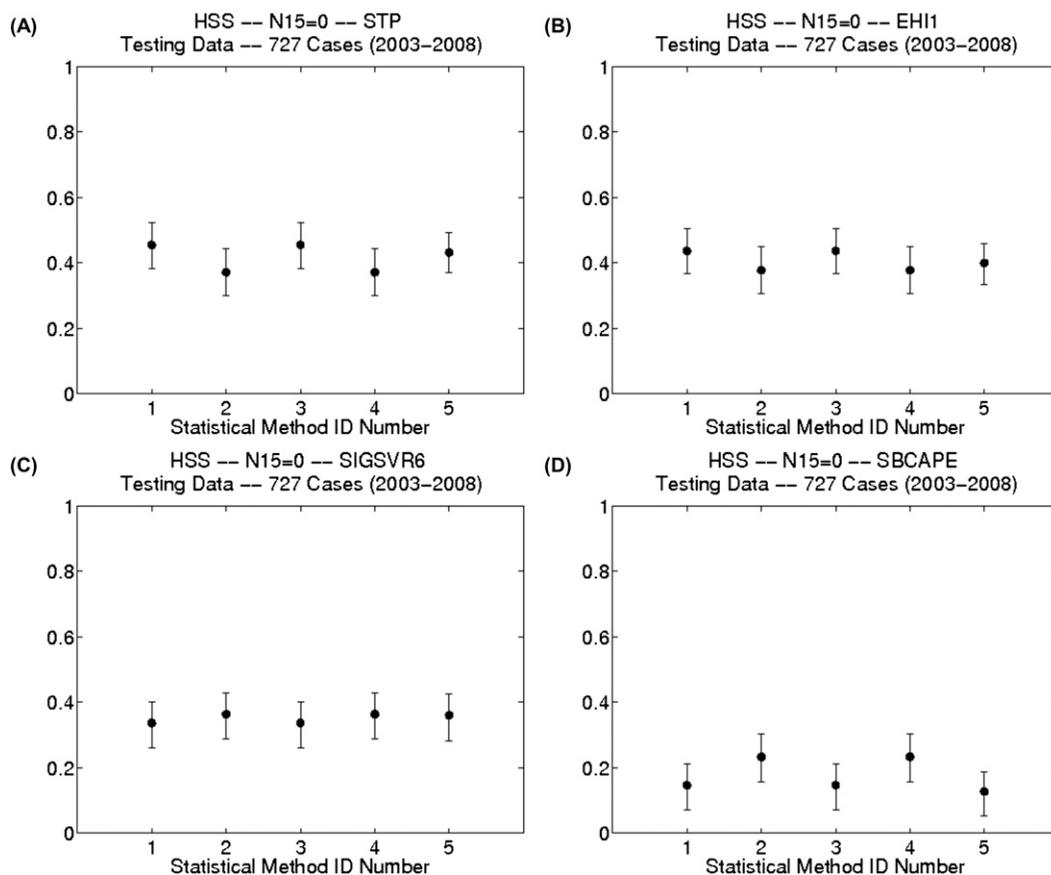


FIG. 6. (a)–(d) HSS values (95% bootstrap confidence intervals, with the medians indicated by the points) of the training models using the 3129 cases from 1979 to 2002, when evaluated on 727 cases from 2003 to 2008. Severe weather variables are as indicated (same as those in Fig. 3). The N15 index threshold of zero is used to classify events as major or minor outbreaks.

prognoses. Nevertheless, we see this comparison as an effective means of assessing the relative utility of the areal coverage technique in the identification of major severe weather outbreaks.

In section 2, we discuss the data and methods used in the study. We present results of areal coverage diagnoses for all of the cases included in the study in section 3 and compare the results of the areal coverage technique to SPC convective outlooks in section 4 for a subset of the cases. Section 5 addresses the use of multiple variables and sample size concerns. In section 6, we summarize the findings and conclude with some ideas for future work.

2. Data and methods

SD11 ranked over 6000 severe weather outbreaks based on their relative severity from the period 1960–2008. Each outbreak was defined as a region associated with a cluster of severe reports occurring in a 24-h period from 1200 UTC on the nominal date to 1159 UTC the

following day. Outbreaks identified on adjacent days were considered separately. The valid time of each event was considered to be the time of the analysis available before the median time of the reports. The fields of meteorological covariates were available in 3-h increments (1200, 1500, 1800, . . . , 0900 UTC; see below). For example, if the median time of the reports was 2230 UTC, the 2100 UTC analysis was used as the valid time of the event.

The North American Regional Reanalysis (NARR; Mesinger et al. 2006) dataset was used for this study, as in S10a. The NARR dataset is available from 1979 to the present, permitting a large number of cases (4057) to analyze for the period of record. NARR data have a horizontal grid spacing of 32 km, and 45 staggered layers in the vertical from the surface to 100 hPa. These data were converted via bilinear interpolation to an 18-km Lambert-conformal grid with 31 vertical levels, which is the grid spacing used in previous outbreak discrimination studies (including Mercer et al. 2009; S10a). Such conversion permits simple comparison of the diagnostic results herein

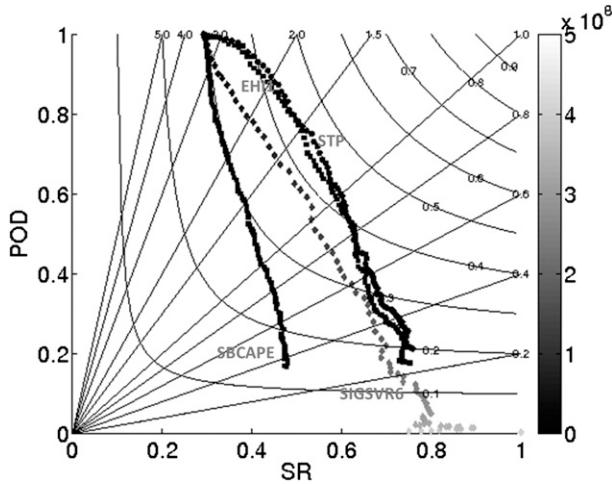


FIG. 7. Performance diagrams (as in Roebber 2009) for the areal coverage iteration technique, with severe weather parameters (labeled). The N15 index threshold of 0 is used for classification of major and minor outbreaks (as in Fig. 3). POD is plotted on the y axis, SR is plotted on the x axis, lines of equal bias are diagonal from bottom left to top right, and curves of equal CSI appear from top left to bottom right. Areal coverage thresholds (used for diagnosis) are indicated as the shaded dots in the diagrams.

to future modeling studies with the same objectives. Examples of the domain are shown in S10a (e.g., their Fig. 3).

Each case is ranked based on the magnitude of one of the outbreak ranking index scores developed by SD11. To remain consistent with previous research (i.e., S10a), the N15 index described in SD11 is selected for this study. The N15 index weights events with multiple significant tornadoes (i.e., $\geq F2$) highest, with moderately high weights given to significant nontornadic reports [i.e., wind speeds $\geq 33.4 \text{ m s}^{-1}$ (65 kt) and hail size $\geq 5 \text{ cm}$].

This allows for major tornado outbreaks to be weighted highest (e.g., Fig. 1a), and events with a large number of significant nontornadic reports to be weighted moderately high (e.g., Fig. 1b). These cases correspond roughly to N15 index scores of ≥ 2 and between 0.5 and 2, respectively. As the N15 scores approach and then fall below 0, the events become somewhat smaller in size and feature few or no tornadoes and few significant nontornadic reports (e.g., Figs. 1c–e).

Areal coverage is used to diagnose each outbreak’s severity, which is determined by the N15 index score for each case. An event is classified as a major outbreak if the N15 index score is equal to or exceeds a predetermined threshold and is a minor outbreak (null case) otherwise. As there is no obvious definition of what a major severe weather outbreak is, this study seeks to determine the threshold of the N15 index score (to within 0.01) in which the areal coverage distinguishes the major and minor severe weather outbreaks most skillfully. This is referred to as the *ranking iteration* method. Alternatively, one could define major and minor severe weather outbreaks using a single N15 index threshold, and determine the areal coverage threshold that most skillfully discriminates outbreaks at the selected N15 index threshold. This is referred to as the *areal coverage iteration* method.

The areal coverage method utilized in S10a is modified in this study. In S10a, areal coverage was computed as the total number of grid points that exceeded a predetermined threshold value for a selected severe weather diagnostic variable. This method was susceptible to the inclusion of locations not associated with the outbreak of interest. In this study, there are two methods by which areal coverage is computed to mitigate this problem: the KDE and intersect methods (IM). For the KDE method,

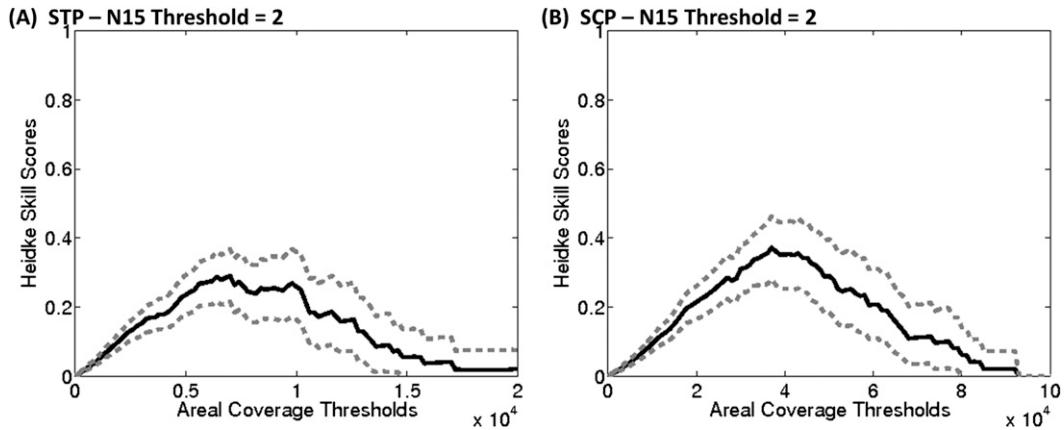


FIG. 8. As in Fig. 5, but for (a) STP and (b) SCP, using the N15 index threshold of 2 to separate major severe weather outbreaks from less significant events.

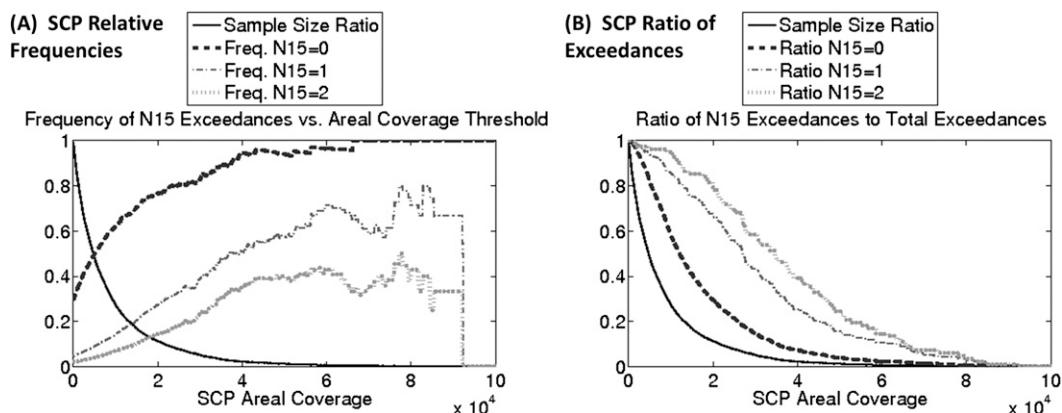


FIG. 9. (a) Ratio of the number of cases that exceeded the N15 index scores for a given SCP areal coverage sum threshold (x axis) to the total number of cases in which the areal coverage threshold is exceeded. (b) Ratio of the number of cases that exceeded the N15 index scores (labeled) for a given SCP areal coverage sum threshold (x axis) to the total number of cases in which the N15 index threshold is exceeded. In each plot, the sample size ratio also is indicated, which is the ratio of the number of cases that exceed the areal coverage threshold indicated on the x axis to the total number of cases.

areal coverage is obtained by computing the value of the severe weather parameter of interest at each grid point within the region associated with the outbreak as determined via KDE (e.g., Fig. 2; see also SD11). These values then are summed, and the mean, median, or sum value of the reports is used as a diagnosis of the outbreak's severity. If the mean, median, or sum value exceeds a predetermined threshold value, the event is diagnosed as a major outbreak.

The second method, IM (Fig. 2), begins by finding all regions in which a severe weather parameter exceeds a specified threshold. The largest contiguous region of the parameter exceedance that also intersects the KDE region associated with the outbreak is selected for analysis. Each grid point within the parameter exceedance region (e.g., within the black ellipse in Fig. 2) is included in the areal coverage computation. (In contrast, the shaded region in Fig. 2 would be used for the KDE method areal coverage computation.) As with the KDE method, the magnitude of the severe weather parameter is tabulated for each grid point, and the mean, median, or sum value is used as a diagnosis of the outbreak's severity. This method was developed in addition to the KDE method because the outbreak's precise location is not known before the event occurs. Thus, a method that incorporates the field of the severe weather parameter that is in proximity to the outbreak was deemed a desirable alternative that may match a forecasting scenario more closely.

After the areal coverage values are computed for each outbreak, binary contingency statistics are used to evaluate the diagnoses. These statistics include the hit rate (HR), probability of detection (POD), false alarm ratio (FAR), probability of false detection (POFD), success ratio (SR), critical success index (CSI), Peirce skill score

(PSS), Heidke skill score (HSS), Clayton skill score (CSS), and Gilbert skill score (GSS). Equations for these statistics are provided in S10a, and many of these variables are discussed in Doswell et al. (1990), Murphy (1996), Wandishin and Brooks (2002), and Wilks (2006), among numerous other studies. The four elements of the binary contingency table (e.g., see the appendix of Mercer et al. 2009) are identified as a (correct hit), b (false alarm), c (missed event), and d (correct null) in this paper.

The severe weather variables analyzed include 0–1-, 0–3-, and 0–6-km bulk shear (BULK1, BULK3, BULK6); surface-based CAPE (SBCAPE); 0–1- and 0–3-km storm-relative environmental helicity (SREH1 and SREH3; Davies-Jones et al. 1990); 0–1- and 0–3-km energy-helicity index (EHI1 and EHI3; Hart and Korotky 1991); the product of SBCAPE and 0–1- or 0–6-km bulk shear (SIGSVR1 and SIGSVR6); the supercell composite parameter (SCP; Thompson et al. 2003); and the significant tornado parameter (STP; Thompson et al. 2003). Many of these parameters have been found to be useful in the discrimination of convective mode and observed severe weather (e.g., Davies and Johns 1993; Johns et al. 1993; Brooks et al. 1994; Rasmussen and Blanchard 1998; Doswell and Evans 2003; Markowski et al. 2003; Thompson et al. 2003; Mercer et al. 2009; Shafer et al. 2009; S10a).

3. Results (1979–2008)

a. The areal coverage iteration technique

Examples of the areal coverage iteration technique, using the N15 index score of zero as the threshold for

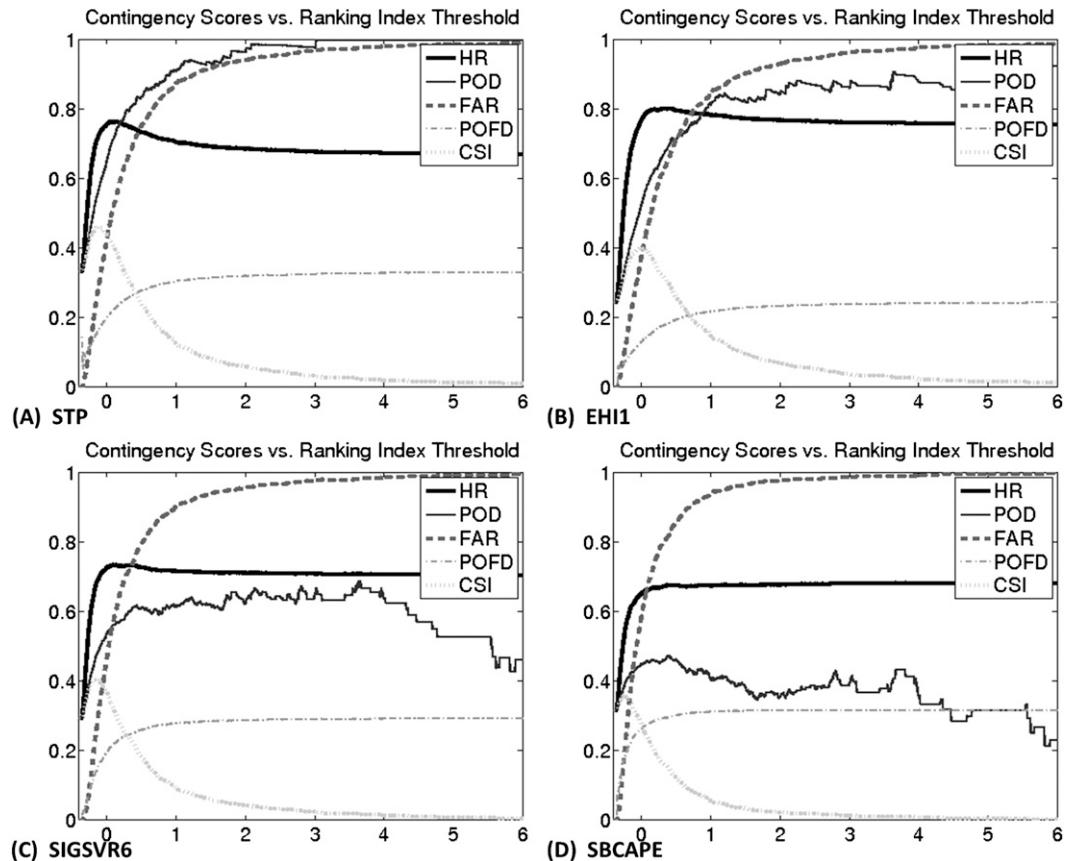


FIG. 10. As in Fig. 3, but for the ranking iteration method, using an areal coverage sum threshold of (a) 1500 for STP, (b) 2250 for 0–1-km EHI, (c) $1.25 \times 10^8 \text{ m}^3 \text{ s}^{-3}$ for the product of SBCAPE and 0–6-km bulk shear, and (d) $6 \times 10^6 \text{ J kg}^{-1}$ for SBCAPE. N15 index thresholds from -0.4 to 6 are shown on the x axis.

classifying outbreaks as major or minor events,² are an effective means of 1) identifying the meteorological covariates that are more accurate and skillful in distinguishing the two classes of outbreaks and 2) identifying the areal coverage threshold that distinguishes these two classes best. The generalization of the results was also tested by separating the 4057 cases into training and testing data. The training data are composed of the cases from 1979 to 2002 (3189 cases), and the testing data are 727 cases from 2003 to 2008 that are compared directly to SPC convective outlooks in section 4. Statistical models are developed using discrimination methods (Table 1). Linear and quadratic discriminant analysis (Seber 1984; Krzanowski 1988) and decision trees (Breiman et al. 1993) were used to train the models, and these were tested on the 727 cases. Other methods were incorporated [e.g., support

vector machines; Cristianini and Shawe-Taylor (2000)] but were not found to improve significantly upon the techniques discussed below (not shown). The test cases were bootstrapped using a bias-corrected-and-accelerated technique (Efron and Tibshirani 1993), as the computed 95% confidence intervals provide insight into whether a particular meteorological variable is significantly better or worse in identifying major severe weather outbreaks.

From the results of the areal coverage diagnoses of the sum values of STP, EHI1, SIGSVR6, and SBCAPE using the KDE method (Figs. 3–7), STP and EHI1 are superior to SIGSVR6 and statistically significantly more skillful than SBCAPE in discriminating major and minor outbreaks. Whereas observed skill scores (excluding CSS) tend to be below 0.25 for the best areal coverage threshold for SBCAPE, skill scores for SIGSVR6 approach 0.4 and for STP and EHI1 exceed 0.4 (Fig. 4). Training models developed using the four parameters clearly suggest the significantly higher discrimination skill of STP and EHI1 to SBCAPE, with the linear discriminant analyses and decision trees of STP and EHI1

² The N15 score of zero is selected as the threshold because evaluation of SPC convective outlooks results in maximum skill scores near this value. See section 4.

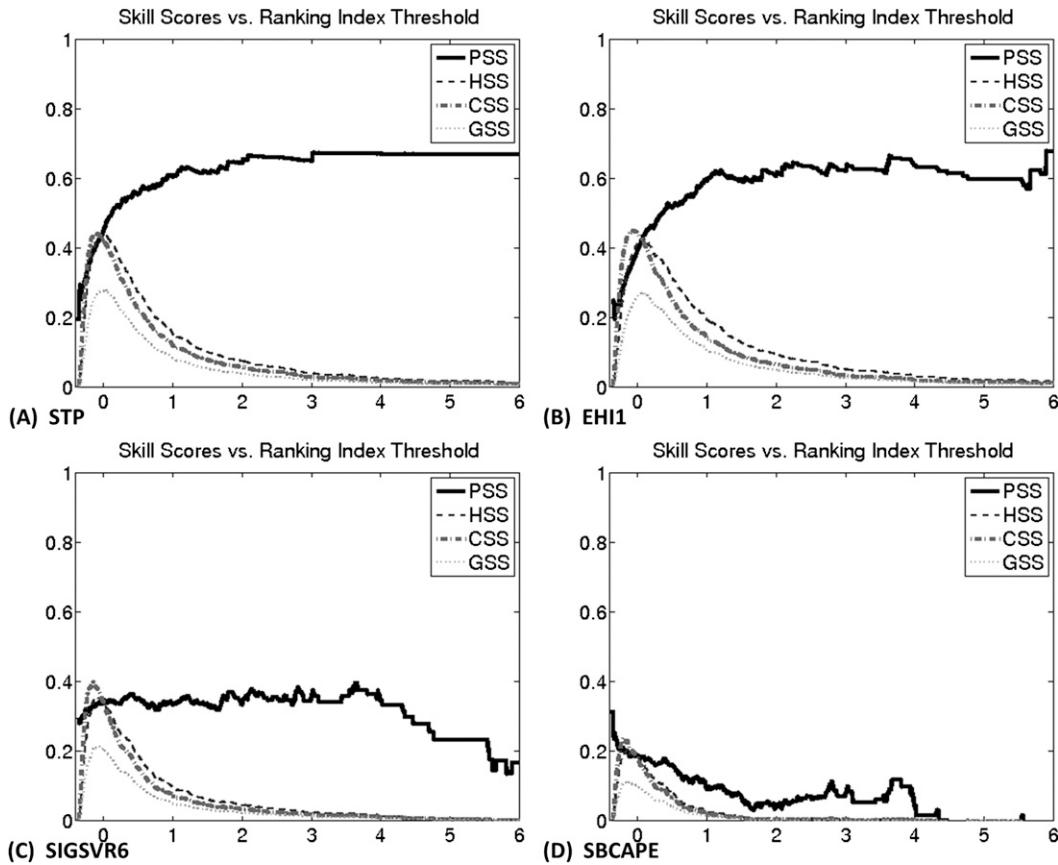


FIG. 11. As in Fig. 4, but using the ranking iteration technique, with variables and thresholds as in Fig. 10.

significantly more skillful than SIGSVR6 to 95% confidence (Fig. 6).

The relative inability of CAPE to distinguish outbreak environments is consistent with previous studies (e.g., Monteverdi et al. 2003; Shafer et al. 2009). As the areal coverage threshold is increased, the FAR does not decrease substantially (Figs. 3d and 7), whereas the POD decreases in a manner similar to other meteorological covariates (Figs. 3a–c and 7). The result is a near-constant value of CSI for a large range of SBCAPE values that begins to decline above areal coverage sums of $5 \times 10^6 \text{ J kg}^{-1}$. The other parameters shown in Fig. 3 exhibit a maximum in CSI (~ 0.4 for SIGSVR6; >0.4 for STP and EHI1).

The CSS trends differently (as a function of increased areal coverage threshold) from the other skill scores using the areal coverage iteration technique. This is because the CSS [=SR – detection failure ratio; see Wandishin and Brooks (2002)] is sensitive to a small number of false alarms and a large number of correct nulls in a rare-events dataset. That is, it is easy to generate diagnoses with few false alarms (by using very high areal coverage thresholds) while correctly identifying a large number of

null cases. Therefore, the CSS should be interpreted with caution for the areal coverage iteration technique.

One of the advantages of the areal coverage iteration technique is that there can be a quick determination of

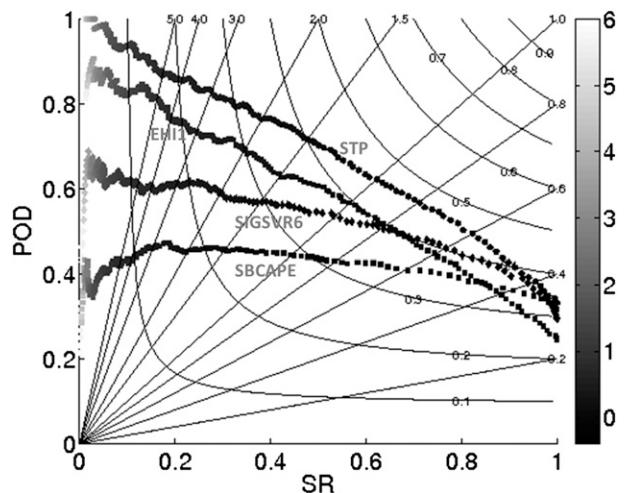


FIG. 12. As in Fig. 7, but using the ranking iteration technique, with the contingency scores for each N15 index score from -0.4 to 6 shown.

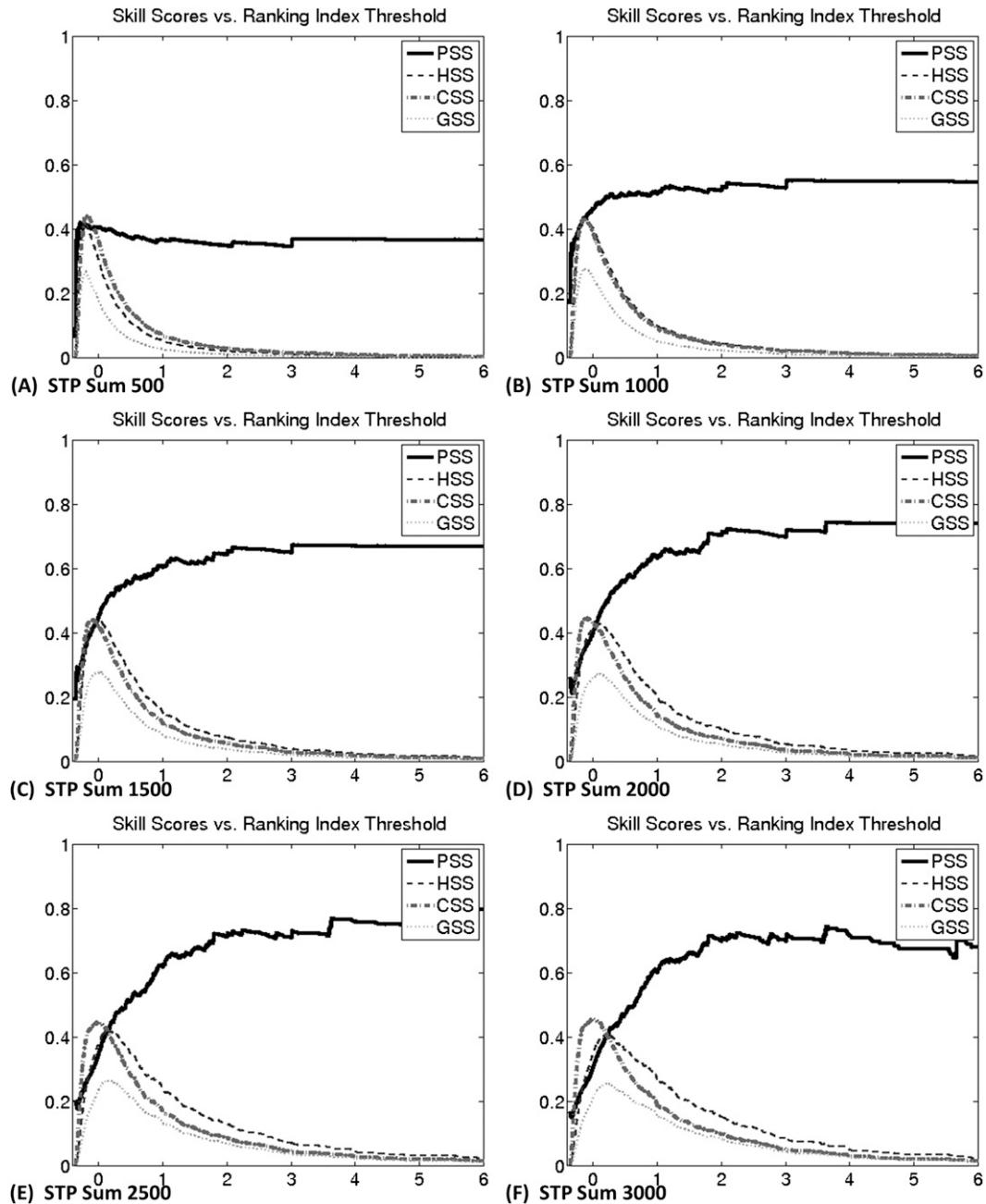


FIG. 13. As in Fig. 11, but using STP sum areal coverage thresholds of (a) 500, (b) 1000, (c) 1500, (d) 2000, (e) 2500, and (f) 3000 as the diagnosis.

the best value to use as a discriminator for any outbreak ranking index threshold. For example, SD11 found that N15 index thresholds of ~ 2 separated major tornado outbreaks from other types of events reasonably well. The areal coverage iteration method indicates that, for STP and SCP sum values respectively, the best thresholds to use (for the gridded domain used in our study) are ~ 7000 and $\sim 37\,000$ (Fig. 8). Additionally, the median

skill score of SCP was larger (~ 0.38) than that of STP (~ 0.3) for these values, though these results are not statistically significant at the 95% confidence level (owing to a smaller sample size of cases with N15 index scores ≥ 2).

Probabilistic analyses provide considerable insight into the utility of the areal coverage technique. Using the SCP sum areal coverage as an example, the number of cases that exceed the N15 index threshold for a given

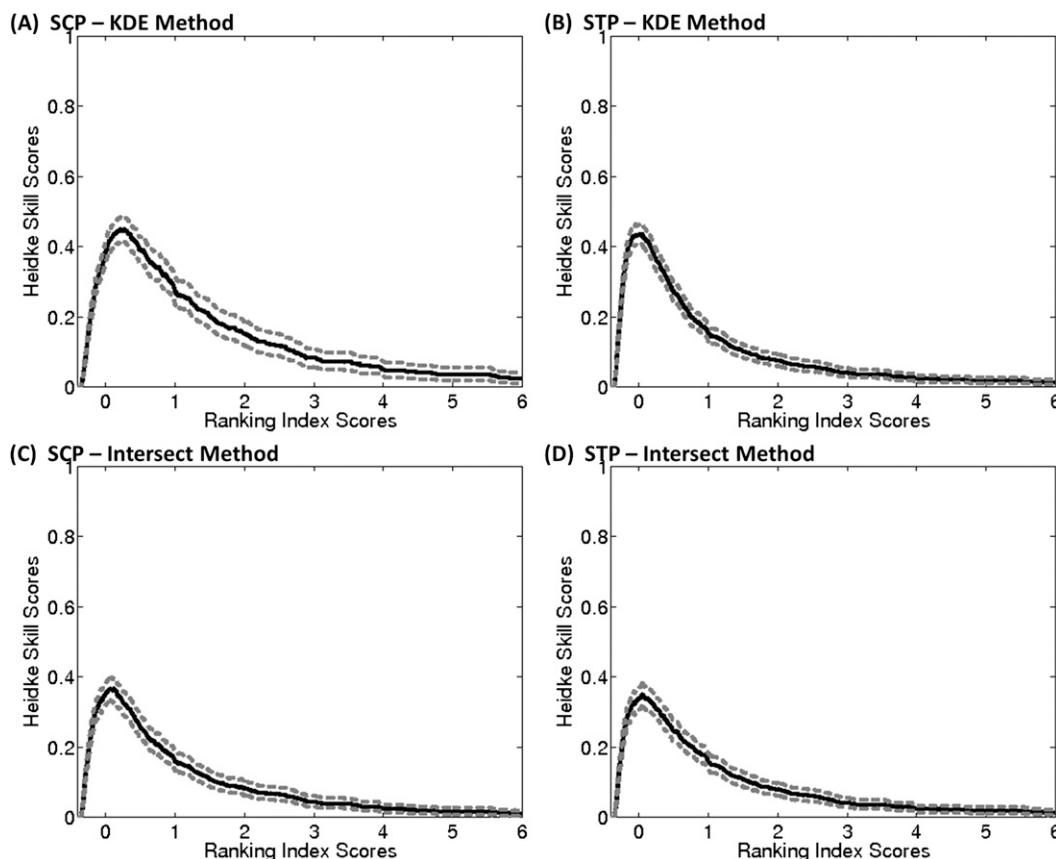


FIG. 14. As in Fig. 5, but here skill scores are shown as a function of N15 index scores (ranking iteration method), for (a) SCP using the KDE method, (b) STP using the KDE method, (c) SCP using the intersect method (for regions with $SCP \geq 1$), and (d) STP using the intersect method (for regions with $STP \geq 1$). The areal coverage sum threshold is 15 000 for (a) and (c), and 1500 for (b) and (d).

areal coverage threshold divided by the total number of cases that exceed the specified areal coverage threshold (Fig. 9a) generally increases with increasing areal coverage threshold and decreasing N15 index threshold. For example, greater than 95% ($\sim 35\%$) of the cases in which the SCP sum areal coverage threshold equals or exceeds 40 000 have N15 scores >0 (>2). Similarly, the number of cases that exceed a selected N15 index threshold for a specified areal coverage threshold divided by the total number of cases that exceed the same N15 index threshold (Fig. 9b) increases with decreasing areal coverage and increasing N15 index thresholds. For example, approximately 10% (40%) of events with N15 scores >0 (>2) featured SCP sum areal coverage values ≥ 40 000.

b. The ranking iteration technique

Differences in the areal coverage iteration and ranking iteration techniques can be observed by the distinct behavior of the contingency statistics as a function of areal coverage (Figs. 3–7) versus as a function of N15

index score (Figs. 10–12). For example, the selection of the lowest areal coverage threshold for a given N15 index score that classifies events as major or minor means that every event is a forecast of a major outbreak (i.e., $c = d = 0$). Thus, the POD and POFD are equal to unity, and the FAR is large (SR is small). As areal coverage increases, the POD, POFD, and FAR trend downward (see Figs. 3 and 7). On the other hand, selecting the lowest outbreak ranking index score as the threshold for classifying events as a major outbreak means that every event is a major outbreak (i.e., $b = d = 0$). Thus, the FAR for any selected areal coverage threshold is zero, the POD is low, and the POFD is undefined. Thus, for the ranking iteration method, the tendency is for increasing POD, POFD, and FAR with an increasing outbreak N15 index score threshold (Figs. 10 and 12). For the areal coverage iteration method, the CSI maximizes at low areal coverage thresholds because $POD < FAR$ at high thresholds. From inspection of Figs. 3 and 7, the CSI tends to be largest when $POD > SR$ (i.e., at biases >1 , or

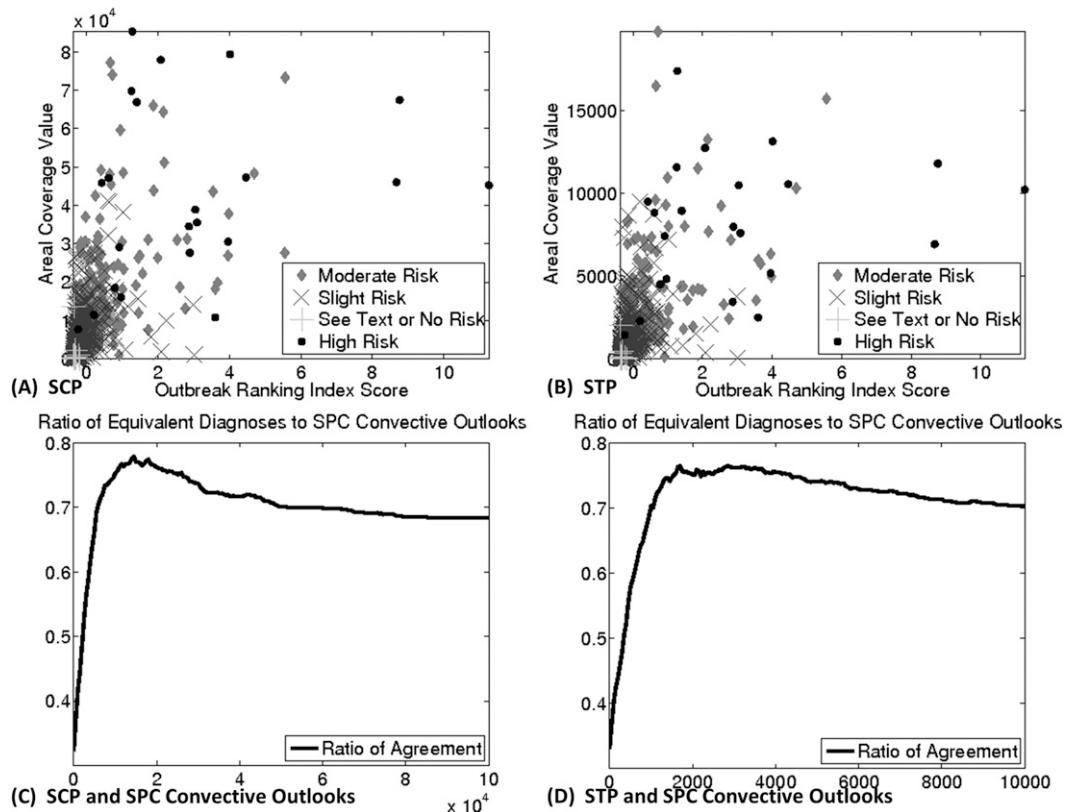


FIG. 15. Scatterplots of areal coverage (y axis) vs N15 index score (x axis) for each of 727 cases from 2003 to 2008 for (a) SCP and (b) STP, using the KDE method. The highest risk of the 1630 UTC day-1 SPC convective outlook associated with the event is indicated. (c) The ratio (y axis) of SPC convective outlooks (day 1, 1630 UTC) that agree with areal coverage diagnoses based on the threshold indicated (x axis) for SCP using the KDE method. (d) As in (c), but using STP.

$c < b$). For the ranking iteration method, the CSI maximizes at low outbreak ranking index thresholds because $POD < FAR$ at high thresholds. From inspection of Figs. 10 and 12, the CSI tends to be largest when $SR > POD$ (i.e., at biases < 1 , or $b < c$).

Moreover, rather than the CSS becoming much larger than the other skill scores at high areal coverage thresholds using the areal coverage iteration technique (Fig. 4), the PSS becomes large at high outbreak ranking index thresholds using the ranking iteration technique (Fig. 11). This is because the $PSS (= POD - POFD)$ is sensitive to a small number of misses and a large number of correct nulls in rare-events datasets, which occur for very high outbreak ranking index thresholds. That is, it is easy to identify the most extreme events and still correctly identify a large number of null cases. This tendency is discussed further in Doswell et al. (1990) and S10a.

The differences in the two techniques allow for a more thorough investigation of a particular variable's ability to discriminate major and minor outbreaks. For example, the ability of SIGSVR6 and SBCAPE to detect the

most significant severe weather outbreaks (i.e., events with increasingly large N15 index scores) for the same areal coverage threshold is considerably lower than that of STP and EHI1. This is obvious from inspection of Fig. 12, which indicates that, at lower SRs, STP and EHI1 have higher PODs for increasing N15 index scores. Moreover, for a given N15 index score, SIGSVR6 and SBCAPE have considerably lower SRs than STP and EHI1 for the same POD at increasing areal coverage thresholds (Fig. 7). The combination of these results suggests that SIGSVR6 and SBCAPE are less accurate and skillful in the discrimination of major and minor outbreaks for any areal coverage threshold for virtually all N15 index scores used to classify the events.

The areal coverage thresholds selected for each of the variables in Figs. 10–12 result in maximum skill scores at N15 index thresholds of around zero. Selection of other thresholds results in displacements of these maxima from zero (e.g., Fig. 13). A reasonable objective is to find the N15 index score with the highest HSS and GSS, but given the uncertainty associated with the N15 index

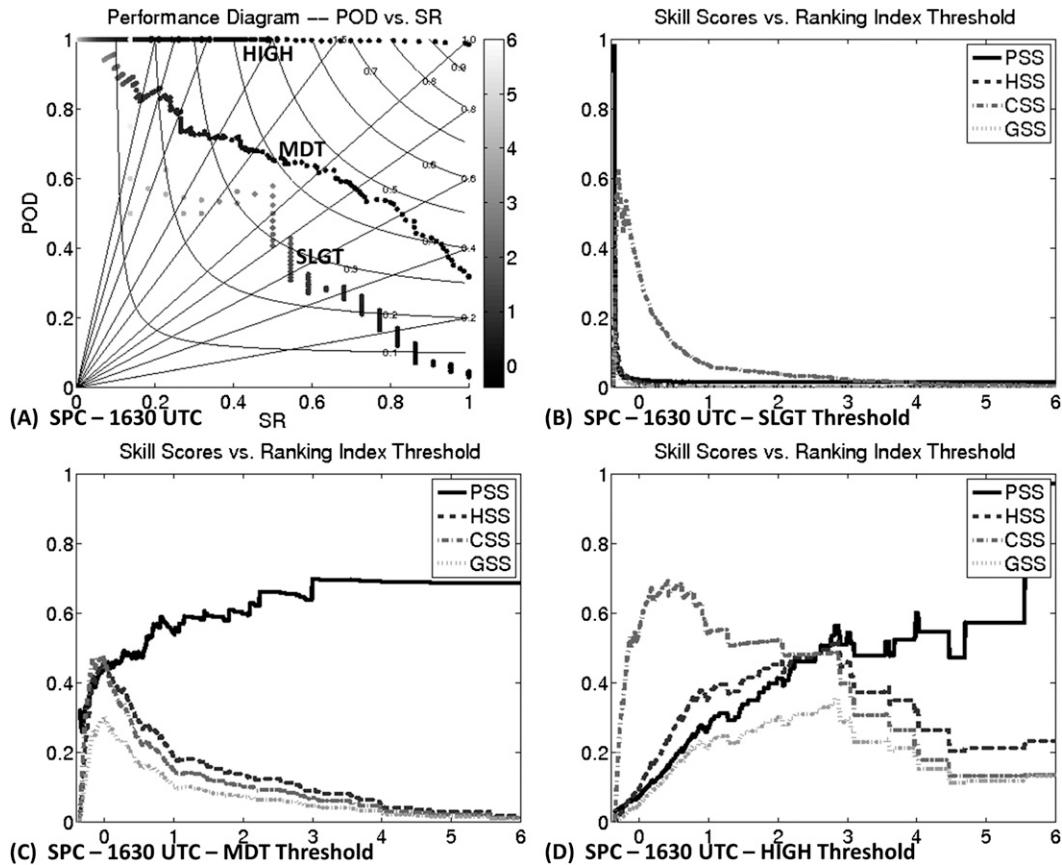


FIG. 16. (a) As in Fig. 12, but using SPC 1630 UTC day-1 convective outlooks with the slight (SLGT), moderate (MDT), and high (HIGH) risks used as the threshold in forecasting major severe weather outbreaks. (b)–(d) As in Fig. 11, but using SPC convective outlook categories of slight, moderate, and high (respectively) as the threshold in forecasting a major severe weather outbreak.

scores (see SD11), one may instead wish to find the highest N15 index score in which the HSS and GSS are not statistically significantly different than the maximum skill observed. This permits identification of the smallest number of most significant outbreaks (i.e., those cases most likely to be major tornado outbreaks) without a substantial decrease in the probability of their detection. For example, STP areal coverage sum thresholds of ~2000–3000 appear to be a desirable choice for the KDE method (Fig. 13) because 1) the maxima in skill scores (HSS and GSS) occur at relatively high N15 index thresholds (compared to using a lower diagnostic areal coverage threshold), which means fewer cases are classified as major severe weather outbreaks without loss in skillfully discriminating those cases from less significant events and 2) the probability to detect these cases does not decrease substantially (relative to selecting a lower areal coverage threshold) while simultaneously reducing the number of false alarms (since fewer cases would be diagnosed as major outbreaks).

c. Comparing the KDE and intersect methods

As the KDE method uses a posteriori knowledge of the outbreak location, it was expected that the KDE method would be somewhat more skillful in distinguishing major severe weather outbreaks from less significant events. This result was universally true for the severe weather parameters tested. For example, comparison of SCP and STP using the KDE and intersect methods (Fig. 14) demonstrates the reduction in skill for the intersect method. The maximum HSS decreases by ~7%–10%. This result was consistent for other variables (e.g., EHI, SREH, and SIGSVR; not shown) and was a function of the magnitude of the maximum scores (i.e., the higher the overall skill, the more discrepancy between the KDE and intersect methods). For HSS, these results were generally statistically significant to 95% confidence (e.g., SCP and STP; Fig. 14). Importantly, however, the intersect method exhibits considerable skill in distinguishing major and minor outbreaks; this is discussed further in section 4.

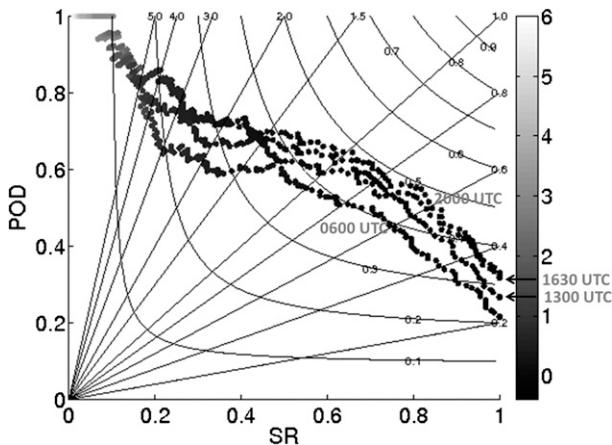


FIG. 17. As in Fig. 12, but for the SPC day-1 convective outlooks issued at 0600, 1300, 1630, and 2000 UTC on the nominal dates for the 727 cases from 2003 to 2008.

4. Results (2003–08)

a. SPC convective outlooks

The SPC issues day-1 convective outlooks several times daily (0600, 1300, 1630, 2000, and 0100 UTC) valid from issuance time to 1200 UTC on the day following the nominal date. SPC convective outlooks can be compared to the areal coverage diagnoses to determine the relative utility of the areal coverage method as a means of diagnosing outbreak severity. However, as mentioned in section 1, caution is advised when making such comparisons. First, the SPC convective outlooks are forecasts of the events, whereas the areal coverage method discussed herein is diagnostic. Second, the SPC convective outlooks are not designed for the outbreak rankings proposed by SD11. However, the levels of risk (i.e., high, moderate, slight, and “see text” or “no organized severe weather”) indicated in the outlooks are associated strongly with the N15 index scores (see below).

SPC day-1 convective outlooks are available from 2003 to 2008, so only those cases (868 total) in the 1979–2008 dataset are shown in this section. However, multiple severe weather events can occur in the same 24-h period. As the highest categorical risk issued by the SPC for a particular day almost always is associated with the most significant event in the 2003–08 sample, only the most significant severe weather outbreak is selected on a given day for simplicity, leaving 727 cases for comparison. The sample size is quite small compared to the 30-yr period used in section 3, and there is evidence that this small sample size affects the results (see section 5). In the following analysis, convective outlooks featuring moderate or high risks of severe weather are evaluated as forecasts of major severe weather outbreaks, and convective

outlooks featuring only slight risks (or “lower” risks) are evaluated as forecasts of minor severe weather outbreaks (null cases), unless otherwise specified. Moreover, this study does not attempt to verify the location of the SPC convective outlooks.

Scatterplots of areal coverage values for the events of interest, with the convective outlooks indicated, suggest substantial agreement in diagnosed–forecast outbreak classification (Figs. 15a,b). Most days with relatively large areal coverage of variables associated with significant severe weather are days with moderate- and high-risk convective outlooks. The ratio of areal coverage diagnoses that “agree” in this way with the convective outlooks (Figs. 15c,d) confirms this. The areal coverage value with the highest ratio typically has a ratio magnitude of 0.725–0.8. Note that, owing to the rare-event nature of the dataset, extremely high areal coverage thresholds always lead to diagnoses of null cases, resulting in at or slightly below 70% of the diagnoses agreeing with SPC convective outlooks (as ~70% are slight risks or lower). In addition, it is apparent that events with relatively high N15 scores generally are diagnosed as major severe weather outbreaks. However, there are occurrences of slight-risk days and events with substantial areal coverage of favorable parameters in which N15 scores are around or below zero (approximately 10%–15% of the 727 cases for the areal coverage method and SPC convective outlooks), which is indicative of a false alarm problem.

Considering the 1630 UTC SPC day-1 convective outlooks (Fig. 16) and using the slight risk as the threshold for forecasting major severe weather outbreaks, nearly every case is forecast as a major severe weather outbreak. This results in a POD of ~1 for every N15 index threshold selected (with decreasing SR for increasing N15 threshold; Fig. 16a), and the highest skill scores for the lowest thresholds (Fig. 16b). Selecting such a threshold for major severe weather outbreaks does not agree with preconceived notions of these events (e.g., see Fig. 1d). Thresholds of about -0.4 may be reasonable, however, in the discrimination of severe weather events from null cases (i.e., “outbreak” versus “no outbreak”).

Using the moderate risk as the major severe weather outbreak threshold results in Roebber (2009) performance diagrams and skill score plots that appear qualitatively similar to the areal coverage techniques discussed in section 3 (cf. Figs. 12 and 16a; Figs. 11 and 16c). Thus, comparing the areal coverage method to the moderate- and high-risk outlooks is appropriate. The skill scores for the moderate-risk or higher outlooks peak at N15 index thresholds of ~0 (Fig. 16c). As discussed in Shafer (2010) and SD11, events with N15 scores ≥ 0 tend to feature multiple significant tornadoes and/or an anomalously

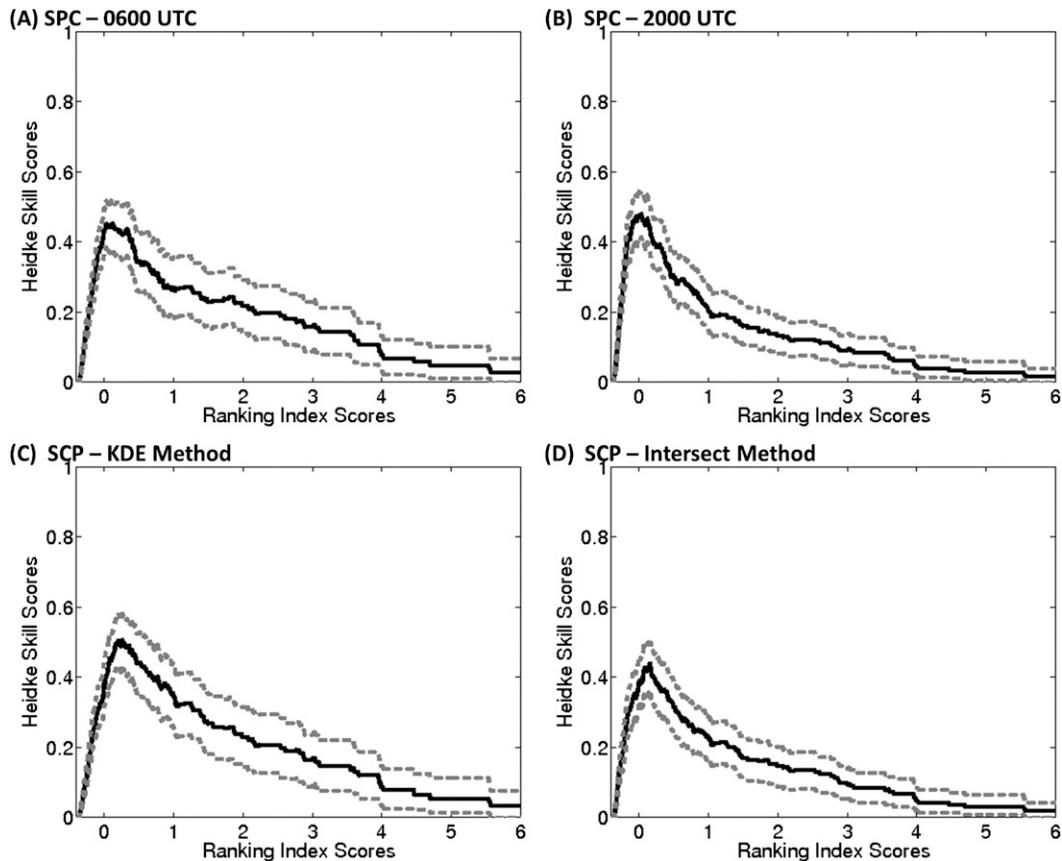


FIG. 18. (a),(b) As in Fig. 16c, but for the 0600 and 2000 UTC SPC day-1 convective outlooks evaluated with the 727 cases from 2003 to 2008. (c),(d) As in (a),(b), but using the KDE and intersect methods, with SCP (areal coverage threshold sum of 15 000) for the 2003–08 cases.

large numbers of significant nontornadic reports. The high-risk threshold for major severe weather outbreaks has a lower POD for the same SR compared to moderate-risk thresholds (or lower SR for the same POD; Fig. 16a), and maximum skill scores at higher N15 index thresholds (~ 2.75 ; Fig. 16d). As mentioned above, events with N15 index thresholds >2 almost always are major tornado outbreaks.

The trends in the performance diagrams (Fig. 17) and skill score plots (Figs. 18a,b) as a function of N15 index threshold from 0600 to 2000 UTC SPC outlooks indicate subtle, but not statistically significant (95% confidence), improvement throughout the day as forecast lead time decreases.³ For example, for a bias of unity, the CSI increases from the 0600 to 2000 UTC outlooks (0.38–0.47). Skill scores also increase from the 0600 to 2000 UTC

outlooks (0.44 to 0.47). A considerable portion of this improvement comes from the 0600 to 1630 UTC outlooks, as 1) 1200 UTC runs of the operational models become available, 2) morning upper-air observations become available, and 3) the evolution of the preconvective environment is established (see Davis et al. 2010 for additional discussion). Between 0600 and 1630 UTC, approximately 82% of the convective outlooks (595 out of 727 of the events) had the same maximum risk. Between 1630 and 2000 UTC, this increased to 95% (693 out of 727 events).

b. Comparison of SPC convective outlooks and areal coverage diagnoses

The areal coverage diagnoses of the 727 cases from 2003 to 2008 are similar to the results for the 1979–2008 period, in general (e.g., cf. Figs. 14a,c and 18c,d). The bootstrapped median contingency statistics of the areal coverage diagnoses using the intersect method (Fig. 18d) are also quite similar to the 0600 UTC (Fig. 18a) and 2000 UTC (Fig. 18b) SPC convective outlooks. As

³ The maximum skill scores also trended toward lower N15 index scores from 0600 to 2000 UTC SPC day-1 convective outlooks, though this trend also was not statistically significant to 95% confidence.

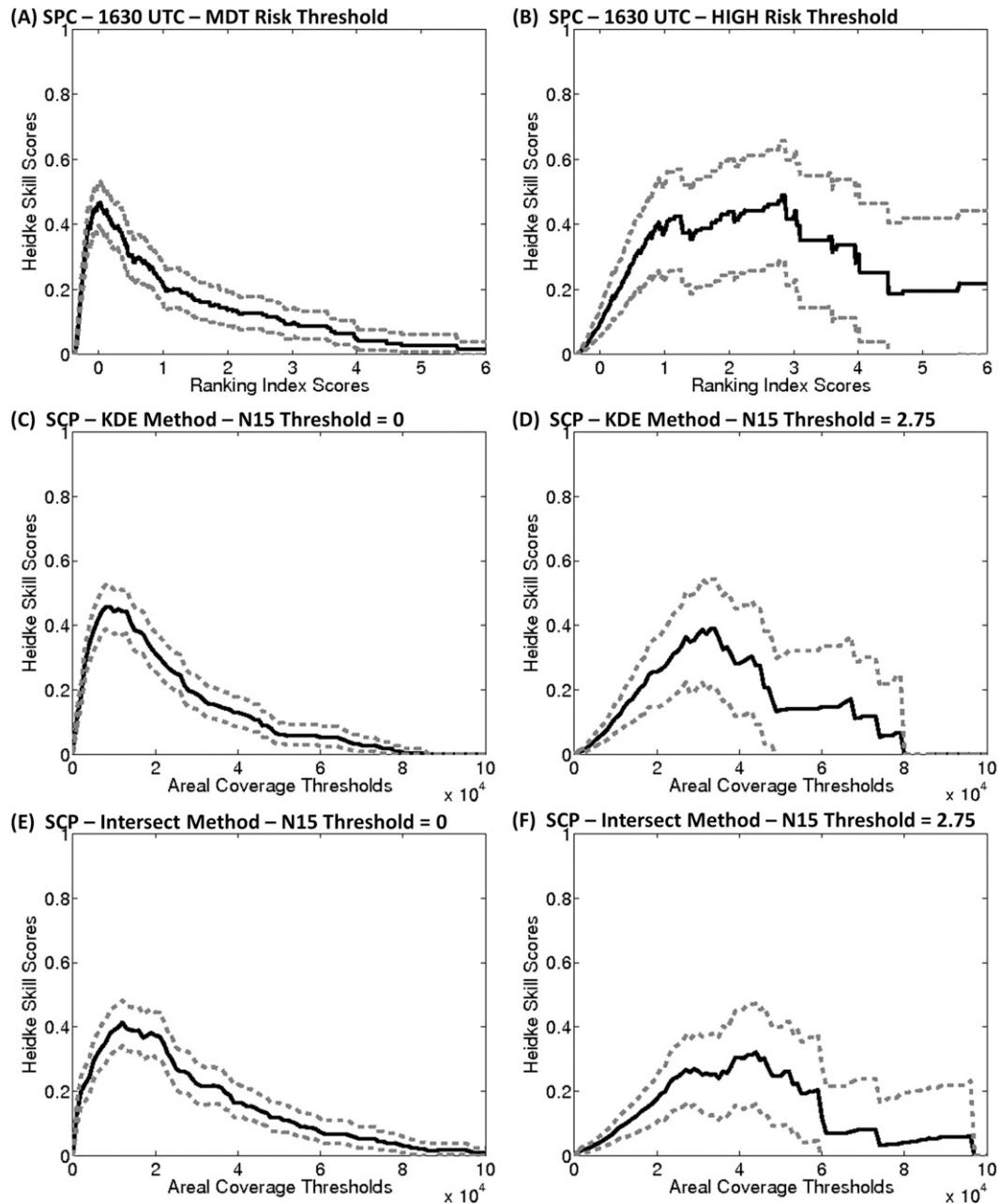


FIG. 19. (a),(b) As in Fig. 18a, but for the SPC 1630 UTC day-1 convective outlooks, with moderate- and high-risk outlooks (respectively) as the thresholds for forecasts of major severe weather outbreaks. (c),(d) As in Fig. 8b, but using SCP and the KDE method (areal coverage threshold sum of 15 000) for the 727 cases during 2003–08, with N15 index thresholds of 0 and 2.75, respectively. (e),(f) As in (c),(d), but using the intersect method (areal coverage threshold sum of 15 000 and a minimum grid point value of 1).

expected, the KDE method is superior to any of the other forecasts and diagnoses (Fig. 18c). However, the bootstrap results suggest no method is statistically significantly superior to any of the others. This result is true for SCP, STP, EHI, and SREH. For other variables (BULK1, BULK3, BULK6, SBCAPE, SIGSVR1, SIGSVR6, etc.), the areal coverage method is similar

to or worse than the SPC convective outlooks (not shown).

Using the N15 index thresholds of 0 and 2.75 (the approximate values of the peak HSS for the moderate-risk and high-risk thresholds; see Figs. 16c,d), the areal coverage iteration method identifies the areal coverage threshold that maximizes HSS for the KDE and intersect

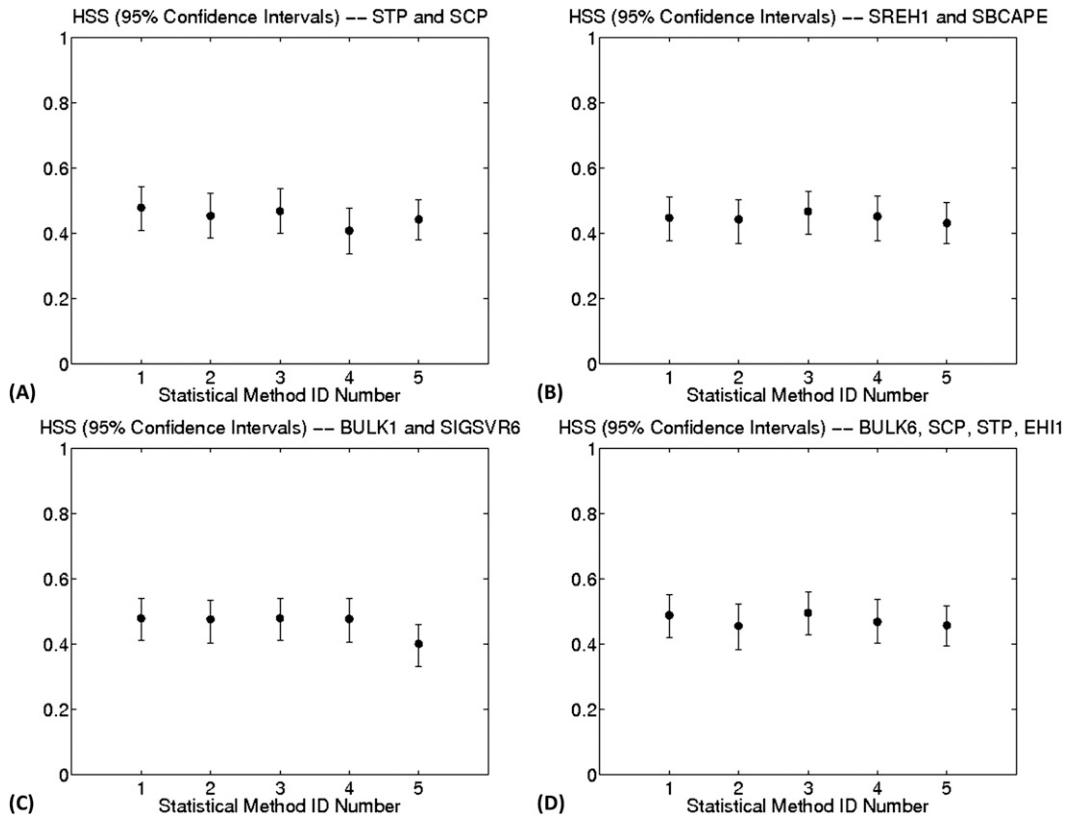


FIG. 20. As in Fig. 6, but using (a) SCP and STP, (b) 0–1-km SREH and SBCAPE, (c) 0–1-km bulk shear and the product of SBCAPE and 0–6-km bulk shear, and (d) 0–6-km bulk shear, SCP, STP, and 0–1-km EHI.

methods (e.g., SCP in Fig. 19). For the N15 index threshold of zero, the bootstrap median HSS of the KDE method is higher than that of the intersect method, and the HSS of the intersect method is similar to that of the 1630 UTC outlook (cf. Figs. 19a,c,e); however, the 95% confidence intervals overlap for all three methods. For the N15 index threshold of 2.75, the SPC convective outlook has a higher bootstrap median HSS than either areal coverage technique (Figs. 19b,d,f); however, the small sample of outbreaks exceeding the N15 index threshold of 2.75 (21 out of 727) results in large confidence intervals and substantial overlap.

5. Extensions of the areal coverage method

a. Multiple covariates

In sections 3 and 4, the utility of the areal coverage method was examined using a single variable (which itself may be a combination of other variables, such as SCP). Next, multiple variables are used to determine if there is increased ability to discriminate major and minor severe weather outbreaks. For this analysis, the 4057 cases are separated into training and testing data, as discussed in

section 3a. One might expect improvement in the discrimination of major and minor severe weather outbreaks if multiple variables are used in the diagnosis. However, bootstrap confidence intervals of the HSSs suggest little or no improvement when using multiple variables (cf. Figs. 6 and 20). There is a reduced range of the bootstrap confidence intervals from previous work [$\sim 20\%$ – 25% in S10a (their Fig. 12f) and $\sim 10\%$ – 15% here, owing to an increased sample size of the test cases (210 in S10a and 727 in this study)].

Why is there virtually no improvement when adding more variables to the analysis? As Doswell and Schultz (2006) discussed, many of the severe weather parameters used in severe weather research or operational forecasting are indices or derived variables, computed by combining meteorological variables associated with thermodynamic instability and/or vertical wind shear in some manner. As a result, many of the severe weather variables are moderately to highly correlated (Table 2). Thus, it appears that little additional skill is gained by adding variables to identify major severe weather outbreaks, using the techniques presented here.

If the N15 index threshold used to separate events into major or minor outbreaks is increased, the contingency

TABLE 2. Correlations of the areal coverage sums for the severe weather parameters labeled for the 4057 cases from 1979 to 2008.

	STP	SCP	EHI1	EHI3	BULK1	BULK6
STP	1.0000	0.9036	0.8607	0.8831	0.5535	0.6429
SCP	0.9036	1	0.9348	0.8441	0.5945	0.5779
EHI1	0.8607	0.9348	1	0.9203	0.5831	0.594
EHI3	0.8831	0.8441	0.9203	1	0.4446	0.6163
BULK1	0.5535	0.5945	0.5831	0.4446	1	0.8019
BULK6	0.6429	0.5779	0.594	0.6163	0.8019	1
SBCAPE	0.2334	0.2334	0.38	0.5049	-0.0323	0.2854
SREH1	0.5877	0.5979	0.541	0.4104	0.9306	0.7345
SREH3	0.6970	0.6255	0.5845	0.5496	0.8728	0.8369
SIGSVR1	0.6835	0.7754	0.9007	0.8512	0.4743	0.5279
SIGSVR6	0.5703	0.5492	0.6711	0.8013	0.1819	0.5297
	SBCAPE	SREH1	SREH3	SIGSVR1	SIGSVR6	
STP	0.2334	0.5877	0.697	0.6835	0.5703	
SCP	0.2334	0.5979	0.6255	0.7754	0.5492	
EHI1	0.38	0.541	0.5845	0.9007	0.6711	
EHI3	0.5049	0.4104	0.5496	0.8512	0.8013	
BULK1	-0.0323	0.9306	0.8728	0.4743	0.1819	
BULK6	0.2854	0.7345	0.8369	0.5279	0.5297	
SBCAPE	1	-0.1668	-0.0806	0.6281	0.8738	
SREH1	-0.1668	1	0.9406	0.3273	0.0708	
SREH3	-0.0806	0.9406	1	0.3715	0.2014	
SIGSVR1	0.6281	0.3273	0.3715	1	0.8058	
SIGSVR6	0.8738	0.0708	0.2014	0.8058	1	

statistics for a particular combination of variables move away from their optimal values (Fig. 21). Moreover, the range of the 95% confidence intervals increases, since the sample size of the major severe weather outbreaks decreases. These larger confidence intervals can be decreased by reducing the size of the training dataset and increasing the size of the testing data; however, the training models may not be as skillful because of the reduced sample size of the training data. For example, if the N15 index threshold of 2 (as in Fig. 21d) is selected to classify events as major or minor outbreaks, only 82 of the 4057 cases from 1979 to 2008 have scores equal to or higher than this threshold (only 28 of the 727 test cases). Therefore, sample size concerns are present no matter the size of the training and testing cases for high thresholds of the outbreak ranking indices.

b. Temporal and sample size sensitivities

To determine if the results of the 1979–2008 period are consistent throughout the time period, all cases for assessment in 6-yr periods were selected. Comparisons of contingency statistics and skill scores of the entire dataset to those of five 6-yr periods (e.g., Fig. 22) indicate the results are not always consistent. The 1985–90 time period has skill scores that are relatively low compared to the other 6-yr periods and to the whole dataset. Of note was this period's relative lack of major tornado outbreaks. The number of severe weather outbreaks considered in

this period was 690, of which 190 cases featured N15 index scores ≥ 0 (27.5% of all cases in this 6-yr period) and 6 cases had N15 index scores ≥ 2 (0.87% of all cases in this period). For the whole dataset, these percentages are 29.2% and 2.02%. The observed skill scores for the cases from 1985 to 1990 are lower than those of the entire dataset (using SCP, maximum HSS of 0.36 versus 0.43; cf. Figs. 22c and 22a). Conversely, the 2003–08 period comprised 868 cases, of which 264 feature N15 index scores ≥ 0 and 30 with N15 scores ≥ 2 . The corresponding ratios are 30.4% and 3.46%, respectively, which are higher than those of the entire dataset. The skill scores for these cases are higher than those for the whole dataset (maximum HSS of 0.52 versus 0.43; cf. Figs. 22e and 22a).

These findings may be demonstrating time sensitivities within the dataset. However, similar results are found for nonconsecutive years featuring the lowest numbers of cases in which the N15 index score is at least 0 and/or 2 (not shown). As forecasting accuracy is known to be related to the frequency of events (e.g., Doswell et al. 1993), it is possible these findings are merely indicative of sample size sensitivities. That is, for subsets of cases in which relatively few (many) major severe weather outbreaks are included, contingency statistics tend to deteriorate (improve) from those of the whole dataset.

Sample size sensitivities are observed in other ways. The skill scores at higher N15 index thresholds are quite variable in the 6-yr periods (e.g., the PSS from 1991 to 1996 versus 1997 to 2002; the HSS from 1997 to 2002 versus 2003 to 2008, etc.). The only solution to this problem is to increase the sample size, but some challenges arise. For example, if more SPC convective outlooks were available for comparison, an underlying assumption of the present work is that these forecasts feature no temporal trends in outbreak discrimination skill. This is not true in reality and is in more danger of being violated with a larger number of years considered. If SPC convective outlooks were compared from 1979 to 2008, for example, the skill with which major severe weather outbreaks were identified would be lower in the first decade than the last decade, because of the increased physical understanding of these events during this period.⁴

6. Summary and future work

The present study is an extension of the work done by S10a, in which the areal coverages of various severe weather parameters are computed for over 4000 severe

⁴ Obviously, this does not imply that an increased sample size would be harmful. This example merely illustrates that the comparison is not straightforward when increasing the sample size.

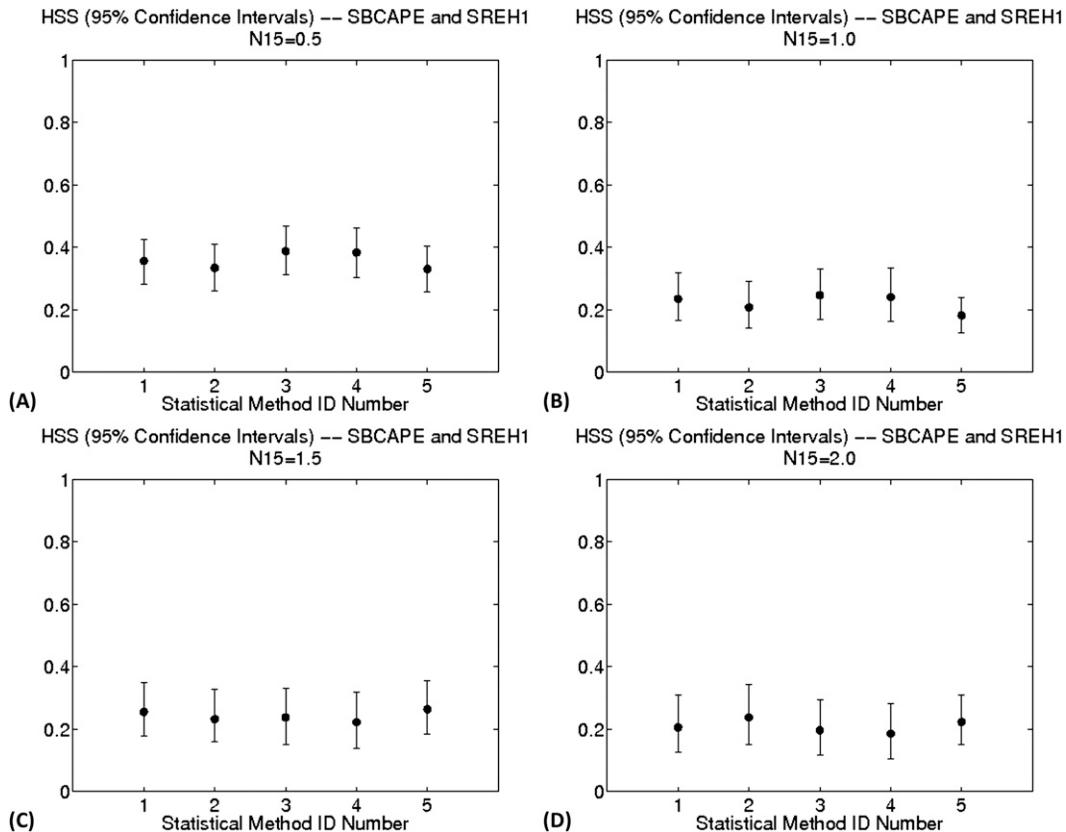


FIG. 21. As in Fig. 20b, with the N15 index threshold for outbreak classification set to (a) 0.5, (b) 1, (c) 1.5, and (d) 2.

weather events to discriminate major severe weather outbreaks from less significant events. Here, the sample size of analyzed cases was increased by nearly a factor of 5 compared to S10a, as a result of the work by SD11 to include multiple severe weather events on the same day. Additionally, rather than using a subjectively selected threshold of the ranking index scores developed by SD10 and SD11 (as conducted in S10a), the thresholds were incremented to determine which thresholds the areal coverage method seemed to discriminate events most skillfully.

The results of this analysis suggest that areal coverage of severe weather parameters can be an effective means of discriminating major severe weather outbreaks from less significant events. The areal coverage technique appears to work best for ranking index thresholds in which the major events consist of a dense cluster of severe reports, and an anomalously large number of significant nontornadoic reports and/or multiple significant tornadoes. Although knowledge of the outbreak's actual location (i.e., the KDE method) results in improved accuracy and skill, using fields of severe weather parameters exceeding a specified threshold in proximity to the outbreak's location (i.e., the intersect method) is still a useful means of discriminating events. The discrimination capability of the

intersect method is particularly important, as it has utility as a prognostic technique, whereas the KDE method as developed in this study could not be implemented in a forecast setting. Comparisons of the KDE and intersect methods to short-term SPC convective outlooks for cases from 2003 to 2008 indicate statistically similar skill, in part because of the relatively small sample of cases available. Sample size sensitivities also are observed for the most significant severe weather outbreaks for the entire dataset, as these events are extremely rare. Large uncertainty is observed when using high outbreak ranking index thresholds to separate major and minor severe weather outbreaks (see, e.g., Figs. 8 and 19).

The results of this study suggest the potential utility of areal coverage in severe weather forecasting. However, as the methods proposed here are diagnostic, the technique needs to be tested as a forecasting tool in subsequent studies. Model simulations for a large number of severe weather events need to be conducted and analyzed. Furthermore, as the location of an outbreak is not known a priori, development of techniques to forecast an outbreak region is necessary. A recent study by Sobash et al. (2011) suggests one approach. In this study, "surrogate" severe weather reports are generated from

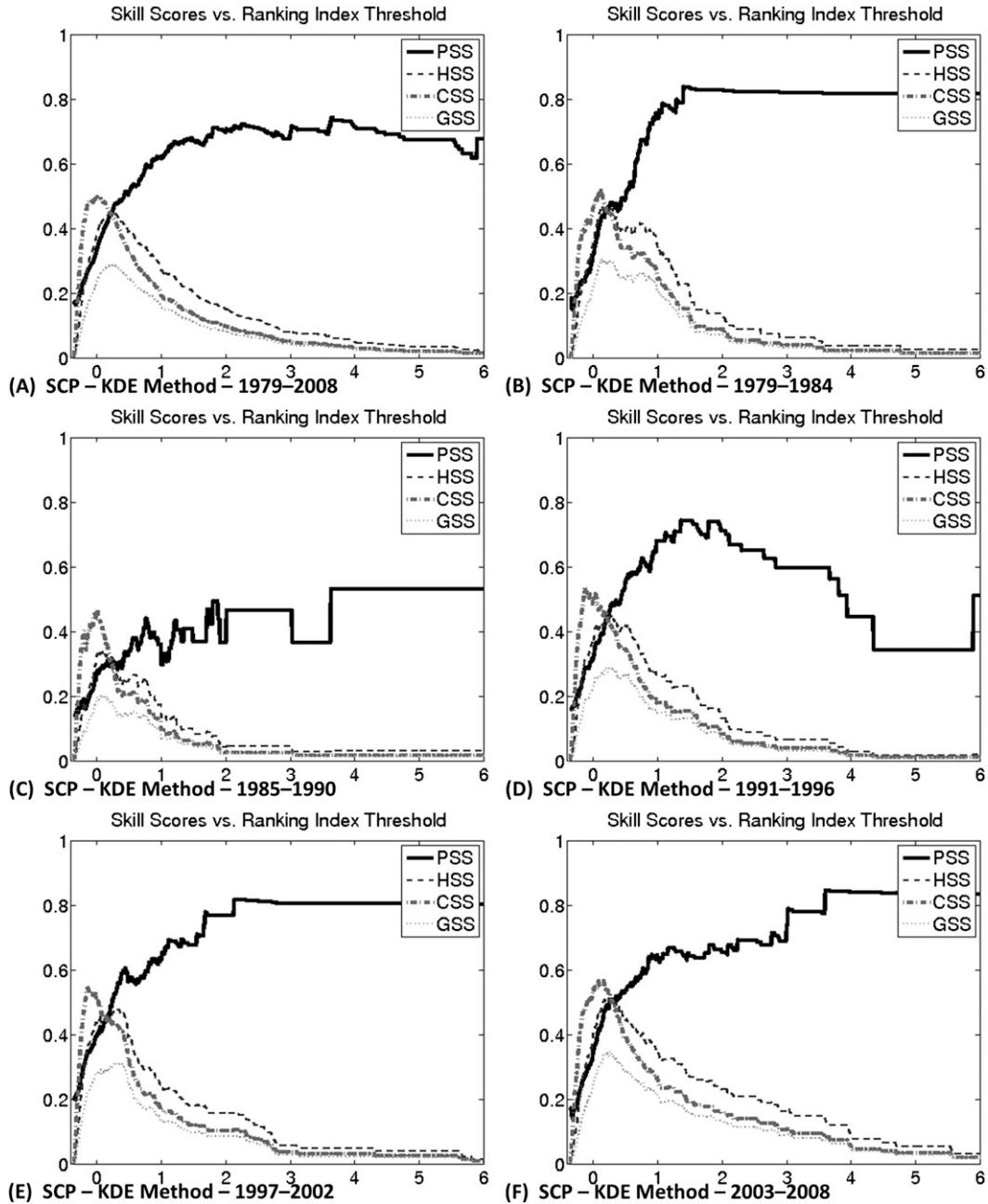


FIG. 22. Skill scores as a function of outbreak ranking index, using SCP and the KDE method for (a) 1979–2008, (b) 1979–84, (c) 1985–90, (d) 1991–96, (e) 1997–2002, and (f) 2003–08.

convection-allowing models. The density of the surrogate severe reports could be used to identify outbreak regions, perhaps using KDE.

Additionally, comparison of the areal coverage method to SPC probabilistic outlooks would be beneficial, particularly when assessing an event’s potential severity for a specific type of severe weather. Conversion of this work to probabilistic guidance also is desired, with emphasis on

associating areal coverage magnitudes with the frequency of threshold exceedances for past events (as discussed in section 3a). However, as the sample size of the most significant severe weather outbreaks and events with very large areal coverage magnitudes is small, event frequencies associated with large areal coverage values become more variable. Thus, many more years of events may be required before reliable and accurate probabilistic forecasts of the

most significant severe weather outbreaks are attained. Finally, assessing the uncertainty of areal coverage in a forecast setting is desirable; thus, the investigation of areal coverage in model ensemble studies is encouraged.

Acknowledgments. The authors thank Steve Weiss, Corey Mead, Roger Edwards, and Rich Thompson for insightful discussions during the beginning stages of this project. Jared Guyer and John Hart have developed and maintained Severe Plot Version 3.0, which was used to create Fig. 1. We thank three anonymous reviewers for their insightful comments, which have improved the manuscript substantially. This project was funded by NSF Grant AGS-0831359.

REFERENCES

- Bowman, A. W., and A. Azzalini, 1997: *Applied Smoothing Techniques for Data Analysis: The Kernel Approach Using S-Plus Illustrations*. Oxford University Press, 208 pp.
- Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone, 1993: *Classification and Regression Trees*. Chapman and Hall, 358 pp.
- Brooks, H. E., 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.
- Brotzge, J., and S. Erickson, 2009: NWS tornado warnings with zero or negative lead times. *Wea. Forecasting*, **24**, 140–154.
- , and —, 2010: Tornadoes without NWS warning. *Wea. Forecasting*, **25**, 159–172.
- Brown, B. G., and A. H. Murphy, 1996: Verification of aircraft icing forecasts: The use of standard measures and meteorological covariates. Preprints, *13th Conf. on Probability and Statistics in the Atmospheric Sciences*, San Francisco, CA, Amer. Meteor. Soc., 251–252.
- 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. Burgess, and M. Foster, 1990: Test of helicity as a tornado forecast parameter. Preprints, *16th Conf. on Severe Local Storms*, Kananaskis Park, AB, Canada, Amer. Meteor. Soc., 588–592.
- Davis, J. M., A. R. Dean, and J. L. Guyer, 2010: A detailed analysis of SPC “high risk” outlooks, 2003–2009. Preprints, *25th Conf. Severe Local Storms*, Denver, CO, Amer. Meteor. Soc. [Available online at <http://ams.confex.com/ams/91Annual/webprogram/Paper179346.html>].
- Doswell, C. A., III, 2007a: Historical overview of severe convective storms research. *Electron. J. Severe Storms Meteor.*, **2** (1), 1–25.
- , 2007b: Small sample size and data quality issues illustrated using tornado occurrence data. *Electron. J. Severe Storms Meteor.*, **2** (5), 1–16.
- , 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. *Electron. J. Severe Storm Meteor.*, **1** (3), 1–14.
- , R. Davies-Jones, and D. L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting*, **5**, 576–585.
- , R. H. Johns, and S. J. Weiss, 1993: Tornado forecasting: A review. *The Tornado: Its Structure, Dynamics, Prediction and Hazards, Geophys. Monogr.*, Vol. 79, Amer. Geophys. Union, 557–571.
- , 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.
- Hamill, T. M., R. S. Schneider, H. E. Brooks, G. S. Forbes, H. B. Bluestein, M. Steinberg, D. Meléndez, and R. M. Dole, 2005: The May 2003 extended tornado outbreak. *Bull. Amer. Meteor. Soc.*, **86**, 531–542.
- Hart, J. A., and W. Korotky, 1991: The SHARP workstation v1.50 users guide. NOAA/National Weather Service, 30 pp. [Available from NWS Eastern Region Headquarters, 630 Johnson Ave., Bohemia, NY 11716.]
- Johns, R. H., and C. A. Doswell III, 1992: Severe local storms forecasting. *Wea. Forecasting*, **7**, 588–612.
- , 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.
- Krzanowski, W. J., 1988: *Principles of Multivariate Analysis: A User's Perspective*. Oxford University Press, 563 pp.
- Markowski, P. M., C. Hannon, J. Frame, E. Lancaster, A. Pietrycha, R. Edwards, and R. L. Thompson, 2003: Characteristics of vertical wind profiles near supercells obtained from the Rapid Update Cycle. *Wea. Forecasting*, **18**, 1262–1272.
- Mercer, A. E., C. M. Shafer, C. A. Doswell III, L. M. Leslie, and M. B. Richman, 2009: Objective classification of tornadic and nontornadic outbreaks. *Mon. Wea. Rev.*, **137**, 4355–4368.
- Mesinger, F., and Coauthors, 2006: North American Regional Reanalysis. *Bull. Amer. Meteor. Soc.*, **87**, 343–360.
- Monteverdi, J. P., C. A. Doswell III, and G. S. Lipari, 2003: Shear parameter thresholds for forecasting tornadic thunderstorms in northern and central California. *Wea. Forecasting*, **18**, 357–370.
- Murphy, A. H., 1996: The Finley affair: A signal event in the history of forecast verification. *Wea. Forecasting*, **11**, 3–20.
- Rasmussen, E. N., and D. O. Blanchard, 1998: A baseline climatology of sounding-derived supercell and tornado forecast parameters. *Wea. Forecasting*, **13**, 1148–1164.

- Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608.
- Schaefer, J. T., 1986: Severe thunderstorm forecasting: A historical perspective. *Wea. Forecasting*, **1**, 164–189.
- Seber, G. A. F., 1984: *Multivariate Observations*. Wiley Press, 686 pp.
- Shafer, C. M., 2010: On the limitations of discriminating outbreaks of severe convection. Ph.D. dissertation, University of Oklahoma, 146 pp.
- , and C. A. Doswell III, 2010: A multivariate index for ranking and classifying severe weather outbreaks. *Electron. J. Severe Storms Meteor.*, **5** (1), 1–39.
- , and —, 2011: Using kernel density estimation to identify, rank, and classify severe weather outbreak events. *Electron. J. Severe Storms Meteor.*, **6** (2), 1–28.
- , 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.
- , C. A. Doswell III, L. M. Leslie, and M. B. Richman, 2010a: On the use of areal coverage of parameters favorable for severe weather to discriminate major outbreaks. *Electron. J. Severe Storms Meteor.*, **5** (7), 1–44.
- , A. E. Mercer, L. M. Leslie, M. B. Richman, and C. A. Doswell III, 2010b: Evaluation of WRF model simulations of tornadic and nontornadic outbreaks occurring in the spring and fall. *Mon. Wea. Rev.*, **138**, 4098–4119.
- Sobash, R. A., J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011: Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme phenomena in convection-allowing model forecasts. *Wea. Forecasting*, **26**, 714–728.
- 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 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.
- 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., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Academic Press, 627 pp.