Objective Classification of Tornadic and Nontornadic Severe Weather Outbreaks

ANDREW E. MERCER, CHAD M. SHAFER, AND CHARLES A. DOSWELL III

Cooperative Institute for Mesoscale Meteorological Studies, Norman, Oklahoma

LANCE M. LESLIE AND MICHAEL B. RICHMAN

School of Meteorology, University of Oklahoma, Norman, Oklahoma

(Manuscript received 2 December 2008, in final form 19 June 2009)

ABSTRACT

Tornadoes often strike as isolated events, but many occur as part of a major outbreak of tornadoes. Nontornadic outbreaks of severe convective storms are more common across the United States but pose different threats than do those associated with a tornado outbreak. The main goal of this work is to distinguish between significant instances of these outbreak types objectively by using statistical modeling techniques on numerical weather prediction output initialized with synoptic-scale data. The synoptic-scale structure contains information that can be utilized to discriminate between the two types of severe weather outbreaks through statistical methods. The Weather Research and Forecast model (WRF) is initialized with synoptic-scale input data (the NCEP–NCAR reanalysis dataset) on a set of 50 significant tornado outbreaks and 50 nontornadic severe weather outbreaks. Output from the WRF at 18-km grid spacing is used in the objective classification. Individual severe weather parameters forecast by the model near the time of the outbreak are analyzed from simulations initialized at 24, 48, and 72 h prior to the outbreak. An initial candidate set of 15 variables expected to be related to severe storms is reduced to a set of 6 or 7, depending on lead time, that possess the greatest classification capability through permutation testing. These variables serve as inputs into two statistical methods, support vector machines and logistic regression, to classify outbreak type. Each technique is assessed based on bootstrap confidence limits of contingency statistics. An additional backward selection of the reduced variable set is conducted to determine which variable combination provides the optimal contingency statistics. Results for the contingency statistics regarding the verification of discrimination capability are best at 24 h; at 48 h, modest degradation is present. By 72 h, the contingency statistics decline by up to 15%. Overall, results are encouraging, with probability of detection values often exceeding 0.8 and Heidke skill scores in excess of 0.7 at 24-h lead time.

1. Introduction

Major tornado outbreaks are of great concern to those living in areas prone to severe weather and to those who forecast the events. Such outbreaks typically are associated with strong synoptic-scale weather systems, but it can be difficult to anticipate the degree of tornadic activity with such systems 24 h or more in advance. Generally, more than 10 major tornado outbreaks affect the United States each year (Doswell et al. 2006, hereinafter D06). Given that an outbreak of severe weather is likely to occur with a particular synoptic-scale system, prior knowledge of impending major tornado outbreaks as opposed to primarily nontornadic events would be ideal, and this study presents an objective method for this discrimination. It is hypothesized that the antecedent synoptic signal possesses information that can be utilized in this outbreak-type classification, so purely synoptic-scale data are used with the methods presented.

One of the first studies of a tornado outbreak (TO) was conducted by Carr (1952), who considered a TO that affected the lower Mississippi Valley and the Tennessee Valley on 21–22 March 1952. Other studies of individual TOs included Fujita (1974), who analyzed the well-known 3 April 1974 ‘‘super outbreak,’’ and more recently, Roebber et al. (2002), who examined the famous 3 May 1999 outbreak. TO classification was initially conducted by Pautz (1969), who defined outbreaks

Corresponding author address: Andrew E. Mercer, 120 David L. Boren Blvd. #5632, Norman, OK 73072.
E-mail: amercer@rossby.metr.ou.edu

DOI: 10.1175/2009MWR2897.1

as small, medium, or large. Galway (1975) used the Pautz (1969) TO classes as a baseline for classifying outbreak type based on tornado deaths by state. Galway (1977) classified TOs into three main categories: a local outbreak (radius less than 1000 mi.), a progressive outbreak (advances from west to east with time), and a line outbreak (tornadic thunderstorms form along a narrow corridor). Grazulis (1993) categorized TOs as groups of 6 or more tornadoes within a single synoptic system. Nontornadic severe weather outbreaks (NTOs), per se, have not been studied, although Kelly et al. (1978) and Doswell et al. (2005) produced climatologies of nontornadic severe weather events.

The work by D06 is the most recent outbreak classification study analyzing TOs and NTOs. D06 avoided any arbitrary definition for what constitutes an outbreak but rather produced a ranking of the different outbreak cases based on the Glickman (2000) definition for a TO: namely, ''multiple tornado occurrences within a single synoptic-scale system.'' Outbreaks are limited to a single day (1200 through 1159 UTC), although several such days could occur in succession as a synoptic-scale system traverses the United States. The outbreak occurrence data are from the SPC database described in detail by Schaefer and Edwards (1999). Several variables are used for the ranking of TO types, including the destruction potential index (DPI; Thompson and Vescio 1998), the number of deaths, etc. A weighted combination of these variables yielded the *O* index, which is used to rank the TO events.

D06 ranked NTOs as well, based on a different set of variables. An event was classified as an NTO if it had 6 or fewer tornado reports. Variables selected included the number of significant wind reports ($\geq$65 kt or 33 m s^{-1}), the number of significant hail reports (diameters $\geq$ 2 in. or 5 cm), the number of tornadoes, the number of severe wind reports ($\geq$50 kt or 25 m s^{-1}), and the number of severe hail reports (diameters $>$ ¾ in. or 2 cm). A similar weighted combination of these variables, denoted in D06 as the *S* index, allowed for the ranking of the NTO events considered. The D06 outbreak study is the baseline for the present work, and the top 50 ranked TOs and NTOs from the D06 study are evaluated using the statistical methodology developed in this study.

These top 50 events are simulated with the Weather and Research Forecast model (WRF; Skamarock et al. 2005) initialized with synoptic-scale input. To determine the WRF's capability to classify outbreak type correctly, objective statistical and learning methods are employed on the WRF output. Statistical techniques are commonly utilized in meteorological studies (i.e., Reap and Foster 1979; Michaels and Gerzoff 1984; Billet et al. 1997; Marzban et al. 1999; Schmeits et al. 2005). Learning methods, such as support vector machines (SVMs; Haykin 1999), are not so widely used in meteorology but have been applied to previous severe weather studies. For example, Trafalis et al. (2005, hereinafter T05) applied SVMs to improve the mesoscale detection algorithm (MDA) on the Weather Surveillance Radar-1988 Doppler (WSR-88D). T05 expanded the previous work by Marzban and Stumpf (1996), who used an artificial neural network (ANN) to improve the algorithm. T05 considered roughly 800 samples for training the SVM model, with less than 2%–10% of the training set consisting of tornado cases. T05 tested the same number of tornado and nontornado events, and the Heidke skill score (Wilks 1995) and the probability of detection (Wilks 1995) were used to evaluate the classification performance. Bayesian neural networks (BNNs; MacKay 1992) produced the largest Heidke skill score values, although the BNN suffered from significant false-alarm ratios (Wilks 1995), which can be problematic for tornado forecasting. SVMs minimized this false-alarm ratio and only decreased the Heidke skill score slightly, so it was chosen as the best method. Other techniques were tested in T05, including an ANN and minimax probability machines (MPMs; Lanckriet et al. 2002), but these methods suffered from large bias and high false-alarm ratio.

The scope of the current project is to determine the extent to which the synoptic signal provides classification ability between TOs and NTOs. This goal will be accomplished through WRF simulations of synoptic-scale input data and statistical classification of outbreak type from the WRF output data. It is important to note that the present study is strictly diagnostic; no prognostic applications are considered. Although this topic has obvious forecast applications, without the ability to classify significant TOs and NTOs with some skill from a diagnostic point of view, further pursuit into a prognostic application of this work will be unproductive. Hence, the current work sets a baseline for further study into a prognostic application of the outbreak classification. Section 2 contains a description of the data and methods used. Section 3 shows results from each of the three temporal initializations. Section 4 contains conclusions and a summary of the results.

2. Data and methodology

a. Data and WRF model simulation

To assess the classification capability of the synoptic-scale signal, a synoptic-scale base dataset of the top 50 TOs and NTOs from D06 was required. As a result, the National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP–NCAR) reanalysis data (Kalnay et al. 1996), which reside on

TABLE 1. NCEP–NCAR reanalysis variables required for WRF simulations and their associated reliability grade from Kalnay et al. (1996).

Input variable	Upper air (U) or surface (S)	Grade
Ice Concentration (1 = ice/0 = no ice)	S	D
Land–Sea mask (1 = land/0 = sea)	S	D
Geopotential Height	U/S	A
Temperature	U/S	A
Relative humidity	U/S	B
"Best" 4-layer lifted index	U	B
Lifted index	S	B
U-wind component	U/S	A
V-wind component	U/S	A
Absolute vorticity	U/S	A
Mean sea level pressure	S	A
Tropopause pressure	U	A
Precipitable water	U/S	B
Vertical speed shear at the tropopause	U	A
Vertical velocity	U/S	B
Surface pressure	S	B
Volumetric soil moisture content	S	C
Specific humidity	S	B
Temperature between two layers below surface	S	C
Temperature at depth below surface	S	C
2-m temperature	S	B
10-m *U* wind	S	B
10-m *V* wind	S	B
Water equivalent of accumulated snow depth	S	C

TABLE 2. Model physics schemes used in the simulation of the 100 outbreak cases (from Shafer et al. 2009).

Model Physics	References
WRF Single Moment 6-class (WSM6) microphysics	Lin et al. (1983); Dudhia (1989); Hong et al. (1998); Skamarock et al. (2005)
Grell–Devenyi convective scheme	Grell and Devenyi (2002)
Yonsei University planetary boundary layer scheme	Hong and Pan (1996)
MM5-derived surface layer scheme	Skamarock et al. (2005)
5-layer thermal diffusion land surface model	Skamarock et al. (2005)
Rapid radiative transfer model for longwave radiation	Mlawer et al. (1997)
Dudhia shortwave radiation scheme	Dudhia (1989)

a 2.5° × 2.5° global grid with 17 vertical levels, were selected. The NCEP–NCAR reanalysis data are derived from an assimilation of model-derived data, climatological data, and observational data. This assimilation results in varied reliability of the reanalysis variables (observational data is generally more reliable than model-derived or climatological data). In Kalnay et al. (1996), reanalysis variables that consist primarily of observational data are graded higher ("A" and "B"), whereas those relying on model-derived quantities are graded as "C" variables and those purely based on climatology are rated "D" variables. Because the synoptic signal's capability to classify outbreak type is based on output from the WRF, those variables required for the WRF simulation (Table 1) were scrutinized. Most variables (75%) were graded as A or B, but a few surface variables (i.e., water-equivalent snow depth and below-surface temperatures and moisture contents) were graded as C variables. Thus, some error may be introduced into the WRF simulations from the variables with poor reliability.

Simulations of the top 50 TOs and NTOs from D06 were conducted at 24-, 48-, and 72-h lead times (see Shafer et al. 2009 for a detailed review of the simulation process). One NTO event, 5 July 1980, had a 1200 UTC valid time (as opposed to the 0000 UTC valid time for the remaining 99 cases), so it was rejected. The resulting 99 WRF simulations were conducted using nested grids. The WRF "mother" domain was fixed at 162-km grid spacing over a 70 × 35 gridpoint domain centered over North America, and four nested domains were placed inside this mother domain (54-, 18-, 6-, and 2-km grid spacings for each). A decrease by a multiple of 3 in grid spacing was required for the WRF simulations, because any other decrease led to model instabilities. The model physics schemes are given in Table 2.

Because the WRF model is incapable of resolving tornadoes, even at 2-km grid spacing, commonly studied severe weather parameters, known as covariates (Brown and Murphy 1996), were computed from the WRF output. A total of 15 different covariates (Table 3) were considered (some at multiple levels) owing to their common usage within the meteorological literature (Table 3, column 3). Because these covariates typically are studied on the mesoscale, the WRF calculated covariates on domain 3 (18-km grid spacing) were retained for the statistical classification. A 21 × 21 gridpoint portion of domain 3 centered on the TO or NTO was extracted from the WRF output, and this domain was used as input into the statistical methods. The outbreak centers (Fig. 1) were chosen subjectively based in the Storm Prediction Center's storm reports (SeverePlot; Hart 1993).

b. Covariate selection

Many of the covariates exhibited large correlations (i.e., Table 4). These high correlations implied some redundancies in the data, so creation of a smaller base

TABLE 3. Initial 15 covariates tested for the classification of outbreak type. References for each covariate as well as levels that are considered for each covariate are indicated.

Covariate	Level(s)	Reference
Surface-based CAPE	Surface	Stensrud et al. (1997)
Surface-based CIN	Surface	Markowski (2002)
LCL		Rasmussen and Blanchard (1998)
Level of free convection		Davies (2004)
Bulk shear	0–1, 0–3, and 0–6 km	Weisman and Klemp (1984)
EHI	0–1 and 0–3 km	McNulty(1995)
SREH	0–1 and 0–3 km	Colquhoun and Riley (1996)
Product of CAPE and bulk shear	0–1, 0–3, and 0–6 km	
Bulk Richardson number shear		Droegemeier et al. (1993)

set of covariates was desirable. Permutation testing was performed on fields of each of the 15 covariates considered, following the methodology of Mercer and Richman (2007). The permutation test (Efron and Tibshirani 1993) is a resampling technique that determines if the means of two distributions are statistically different, making no assumptions of the distributions of the data. This test is superior to the *t* test for the present study, because the distribution of the covariates is unknown and the *t* test requires a normal distribution.

Permutation tests were conducted on the candidate covariates for all 50 TOs and all 49 NTOs at each grid point, and the resulting *p* values were calculated. In the *p* value fields, values of 0.1, 0.05, and 0.01 were contoured (corresponding to the 90%, 95%, and 99% confidence limits). If *p* values are larger than these contoured values, the means of the TO and NTO covariates are not statistically different to the given significance level, so the null hypothesis, which states the two distributions are the same, cannot be rejected.

A sample plot of *p* values for convective available potential energy (CAPE; Fig. 2) provides an example of a covariate with marginal outbreak discrimination capability. In Fig. 2, some areas of statistical significance, up to the 99% confidence level, appear in the southern third of the domain. This result may be due to a dependence of CAPE on latitude and may not be meaningful. In this example, CAPE was rejected. Similar analyses were performed for all covariates at each lead time (Table 5). Many covariates were significant throughout the entire domain [i.e., 0–1-km storm relative environmental helicity (SREH) at 24 h was significant to the 99% confidence level for every grid point in the domain].

c. Statistical methodology

Once a reduced set of covariates for each lead time was determined from permutation testing, a principal component analysis (PCA; Richman 1986; Wilks 1995) was performed to reduce the gridded covariate fields to individual values. These PC scores contained information about the spatial structure of the data, accounting for the physical features of the outbreaks despite a smaller input dataset. The resulting PC scores from the PCA were used as inputs into the statistical models.

Two methods, logistic regression (Log*R*) and SVMs, were evaluated in the discrimination of the outbreak types. Log*R* is a method that is linear with respect to the logit of the binary outcome. The logit, for the present study, is the probability of a TO versus an NTO. This probability lends itself for a forecast application; however, because the scope of this project is to classify outbreak type, a threshold of 0.5 is used to discriminate between the two types.

SVM (Haykin 1999; Cristianini and Shawe-Taylor 2000; appendix A) is a learning method that defines a decision hyperplane for classification. This nonlinear technique has been used in previous meteorological studies (i.e., T05; Mercer et al. 2008), but its appearance in the literature is limited. The SVM method requires several parameters (the cost coefficient *C* and the kernel function with its associated parameters), which are tuned through cross validation. This cross validation was conducted by withholding 80% of the dataset for training and using the remaining 20% for subsequent testing. Numerous kernel functions and cost coefficients were tested using this cross-validation dataset to determine the optimal values of these SVM parameters for our dataset. This method determined that the radial basis function,

$$k(\mathbf{x}, y) = e^{-\gamma\|\mathbf{x}-y\|^2}, \tag{1}$$

was the optimal kernel function, and a cost value of 25 000 produced the best classification results on the testing data. The optimal γ value in the RBF was found to be 0.1.

To measure the performance of a classification scheme, contingency statistics were computed on the results from

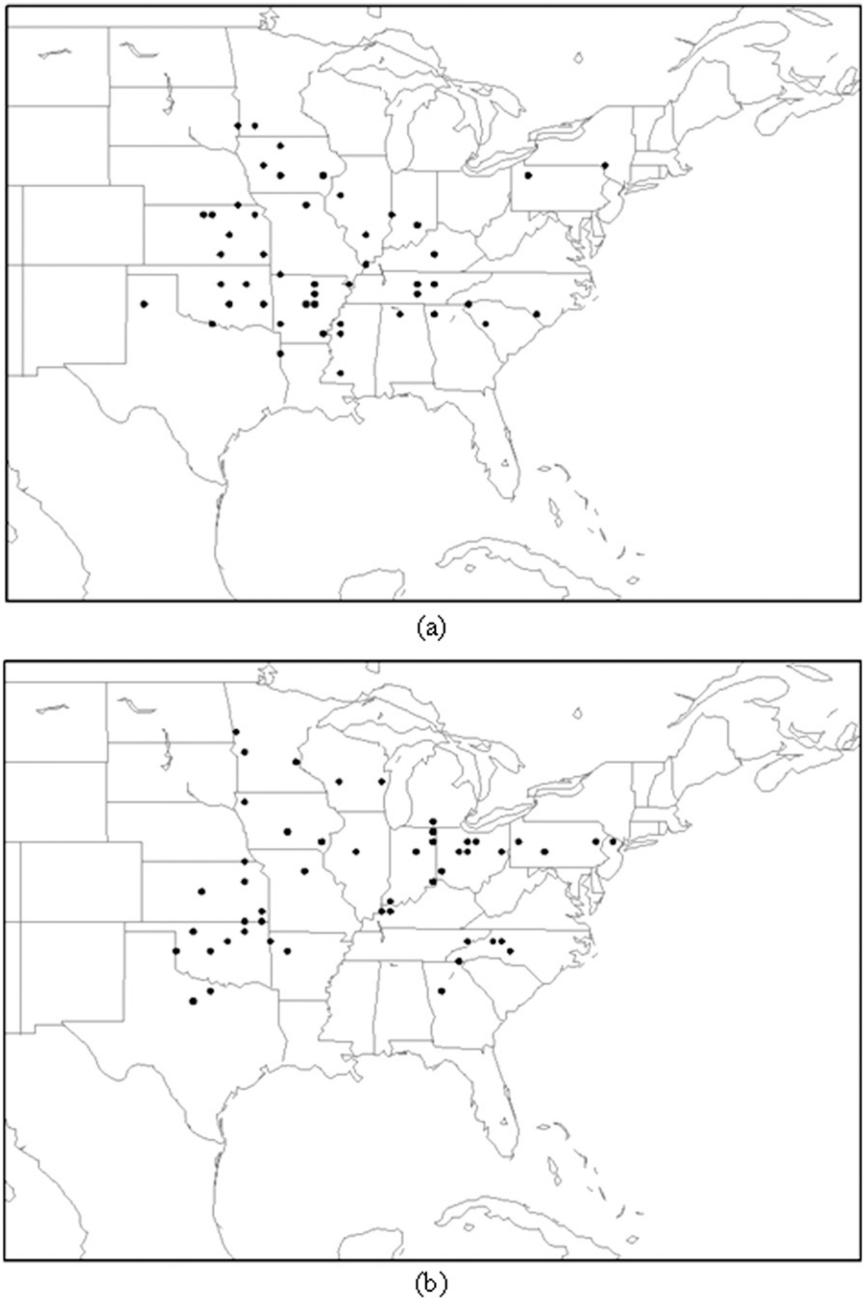

FIG. 1. Outbreak centers for the (a) 50 tornado outbreaks and (b) 50 nontornadic severe weather outbreaks. Some overlap exists between centers, so less than 50 points are on each figure.

the statistical techniques. The contingency statistics require the use of a contingency table (Wilks 1995; appendix B). Four contingency statistics, hit rate (HR), probability of detection (POD), false-alarm ratio (FAR), and Heidke skill score (HSS), were used to determine the classification capabilities of both statistical methods. These contingency statistics appear throughout the meteorological literature (i.e., Doswell et al. 1990; McGinley et al. 1991; Schaefer 1990; and others) and are defined in Wilks (1995).

A method derived from the jackknife (Efron and Tibshirani 1993) was used in cross-validation of the PC score data. The jackknife technique samples without replacement, so that each case was trained and tested

TABLE 4. Lower triangle of the correlation matrix of the 15 covariates from the 3 May 1999 TO. A large percentage of the correlations exceed 0.5 (over 40%), implying data redundancies and suggesting a need for data reduction.

	CAPE	CIN	LCL	CFC	0–1-km shear	0–3-km shear	0–6-km shear	0–1-km EHI	0–3-km EHI	0–1-km SREH	0–3-km SREH	0–1-km CAPE shear product	0–3-km CAPE shear product	0–6-km CAPE shear product	BRN shear
CAPE	1.00														
CIN	0.22	1.00													
LCL	−0.55	0.09	1.00												
CFC	−0.06	0.51	0.05	1.00											
0–1-km shear	0.55	0.26	−0.05	−0.42	1.00										
0–3-km shear	0.24	0.28	0.32	−0.46	0.64	1.00									
0–6-km shear	0.28	0.39	0.34	−0.19	0.41	0.90	1.00								
0–1-km EHI	0.94	0.46	−0.40	0.02	0.64	0.45	0.51	1.00							
0–3-km EHI	0.93	0.11	−0.54	−0.32	0.73	0.42	0.34	0.89	1.00						
0–1-km SREH	0.71	0.56	−0.04	−0.10	0.79	0.75	0.75	0.88	0.75	1.00					
0–3-km SREH	−0.10	−0.34	0.05	−0.84	0.54	0.58	0.25	−0.07	0.25	0.18	1.00				
0–1-km CAPE shear product	0.94	0.28	−0.43	−0.18	0.79	0.40	0.35	0.94	0.96	0.82	0.12	1.00			
0–3-km CAPE shear product	0.90	0.33	−0.28	−0.21	0.69	0.63	0.64	0.96	0.92	0.90	0.13	0.92	1.00		
0–6-km CAPE shear product	0.94	0.35	−0.34	−0.08	0.58	0.50	0.59	0.97	0.89	0.85	−0.04	0.90	0.98	1.00	
BRN shear	0.13	0.33	0.38	−0.21	0.31	0.89	0.97	0.37	0.21	0.64	0.30	0.20	0.51	0.45	1.00

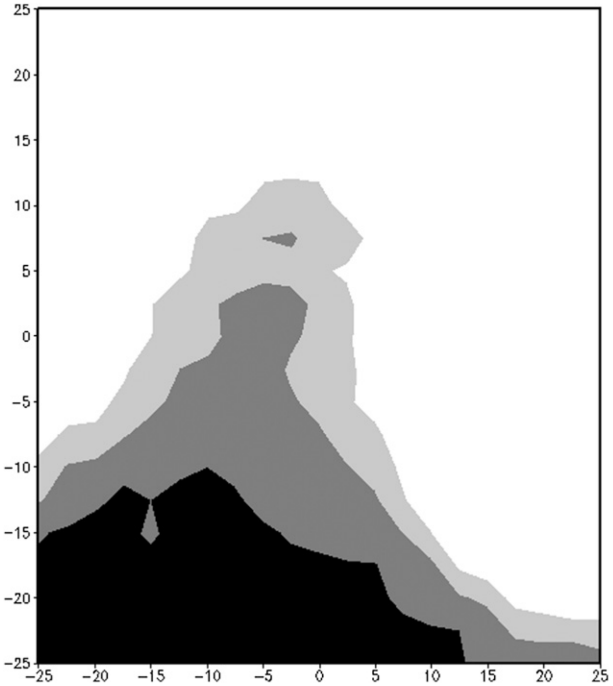

FIG. 2. The p values from permutation testing for surface-based CAPE at 24 h. White colors represent p values larger than 0.1, whereas the lightest gray represents p values less than 0.1, the darker gray represents p values less than 0.05, and black represents p values less than 0.01. The axes are the latitudinal and longitudinal deviations from the outbreak center in degrees.

TABLE 5. Optimal covariate sets determined using permutation testing for the results at 24, 48, and 72 h. These covariates are the base sets used for each statistical technique prior to backward covariate selection.

24-h selected covariates
0–1-km SREH
0–3-km SREH
Surface-based CIN
0–1-km bulk shear
Product of 0–1-km bulk shear and surface-based CAPE
Lifting condensation level
0–1-km EHI
48-h selected covariates
0–1-km SREH
0–3-km SREH
0–1-km bulk shear
0–3-km bulk shear
0–6-km bulk shear
Lifting condensation level
Bulk Richardson number shear
72-h selected covariates
0–1-km SREH
0–3-km SREH
0–3-km bulk shear
0–6-km bulk shear
Lifting condensation level
0–1-km EHI

upon. However, in the present work, a small percentage of the data (15%) was withheld for testing, whereas the remaining 85% was used for training. Once the results for the initial training and testing set were computed, a new testing and training set was obtained through removing the first test case and adding the first training case to the testing set, while adding the removed testing case to the training set. For example, our first iteration used cases 1 through 84 for training and cases 85 through 99 for testing. The second jackknife iteration used cases 2 through 85 for training, then cases 86 through 99 and case 1 for testing. This method allows for each case to be tested 15 times with different training sets. This method provided a set of 99 statistical models for each lead time, as well as 99 contingency statistics from each model. Note that this method could not be used in a forecasting application, because a determination of which of the 99 models is superior would be required. The present study is diagnostic, so this additional step was not done.

Once the 99 contingency statistics were obtained, bootstrap samples (Efron and Tibshirani 1993) of the contingency statistics were computed. The bootstrap samples with replacement the 99 contingency statistics a user-defined number of times (for the present study, 1000 times). The same bootstrap sample was used for each covariate combination and each statistical technique to allow for comparison between the different methods. Subsequently, confidence limits based on the tilted bootstrap (Efron and Tibshirani 1993) were obtained to determine which statistical method performed best. The optimal contingency statistics and their associated confidence limits for each method and each lead time are provided in section 3.

3. Results

The base set of covariates for each lead time (Table 5) was used in a backward elimination methodology. Individual covariates were removed from the base set to determine if results could be improved further, resulting in over 20 combinations of covariates for each lead time (i.e., Table 6). Tilted bootstrap confidence limits were plotted (i.e., Fig. 3) to determine which covariate combinations provided the optimal contingency statistics. Median contingency statistics that are smaller (or larger for FAR) than the lower confidence limit of the combination that produces the largest (or smallest for FAR) median are statistically inferior to the 95% confidence limit and can be rejected. This rejection produced

TABLE 6. A sample of the backward elimination conducted on the base covariate sets. This table represents 24-h lead time. Rows 1–19 show the combinations based on the optimal combination (Table 5), and rows 20–26 show the combinations after leaving off the product of CAPE and bulk shear, which gave the best results among rows 1–19.

Model no.	Variable(s)
1	All
2	No LCL
3	No 0–1-km CAPE shear
4	No 0–1-km bulk shear
5	No surface CIN
6	No SREH (0–1 km)
7	No SREH (0–1 km)
8	No EHI (0–1 km)
9	No shear
10	No SREH
11	Only LCL
12	Only surface CIN
13	Only 0–1-km bulk shear
14	Only 0–1-km CAPE shear
15	Only 0–1-km SREH
16	Only 0–3-km SREH
17	Only 0–1-km EHI
18	Only SREH
19	Only shear
20	No 0–1-km EHI
21	No 0–1-km bulk shear
22	No 0–1-km SREH
23	No 0–3-km SREH
24	No LCL
25	No surface-based CIN
26	No SREH (all)

a smaller set of optimal covariate combinations; in some cases, it resulted in a single optimal combination.

As an example, consider the tilted bootstrap confidence intervals for 24-h lead time SVM mean contingency statistics (Fig. 3). The HR plot (Fig. 3a) has the largest median HR value with model 1, which included all seven of the base set covariates (Table 5, top section). Visual inspection of the figure reveals that only four combinations are within the 95% confidence limit of the best group (group 1) for HR. For POD (Fig. 3b), all groupings except combination 1 are outside the 95% limit, allowing for further rejection of covariate combinations. The FAR results must be interpreted differently; the upper limit of the grouping with the lowest median FAR (in this example group 17) must be compared to the remaining results. FAR medians lying above this upper limit are statistically inferior and should be rejected. Group 1, which is the best for POD and HR, is not within the 95% confidence level of group 17 and should be rejected. Consequently, for SVM at 24-h, the optimal combination depends on the contingency statistic being considered. The HSS results (Fig. 3d) reveal a combination of the results from the other three statistics, suggesting that models 1, 9, 21, and 25 have the same skill to a 95% confidence level. These analyses were conducted for each statistical technique at each lead time.

a. 24-h results

As indicated previously, several combinations, including one that contained the entire base set of covariates, one that rejected both shear variables, and one that rejected the product of 0–1-km bulk shear and CAPE and surface-based CIN, produced the optimal contingency statistics for SVM. For Log*R*, eight covariate combinations had contingency statistic values that were within the 95% confidence limit of the group with the highest median HR and POD (the group that rejected the product of 0–1-km bulk shear and CAPE and surface-based CIN). Hence, numerous covariate groupings were ideal for 24-h classification with Log*R*, although the set is reduced considerably from the initial set of 26 groupings.

One covariate group that was common between Log*R* and SVM rejected surface-based CIN and the CAPE shear product at 0–1 km. This grouping only included one covariate that contained information on the thermodynamic instability in the atmosphere [0–1-km energy–helicity index (EHI)]. This result is expected (Johns et al. 1993), because thermodynamic instability magnitude varies considerably between outbreak type (i.e., CAPE is a necessary but not sufficient condition for a tornado outbreak). This result confirms previous conclusions (Stensrud et al. 1997; Johns and Hart 1993; Johns and Doswell 1992) that thermodynamic variables are not crucial for distinguishing storm type or outbreak type, although they can be useful for differentiating storm cases from those without storms.

To assess the best statistical technique for each lead time, the two statistical methods were compared using one of the best covariate combinations from each method. The confidence intervals for these combinations were tabulated to determine if one method was superior to the other. For 24-h (Table 7) lead time, the HR results for SVM are statistically superior (to a 95% confidence) to the Log*R* results. The Log*R* POD, FAR, and HSS results are within the 95% limit of the SVM results, so it is not possible to determine which is superior.

Because most contingency statistics were indecipherable, it is not possible to say with certainty that one method is superior to another, although the HR results are statistically superior with SVM.

b. 48-h results

Both statistical methods produced smaller magnitudes of HR and POD at 48-h lead time, which was anticipated because of increased WRF error with

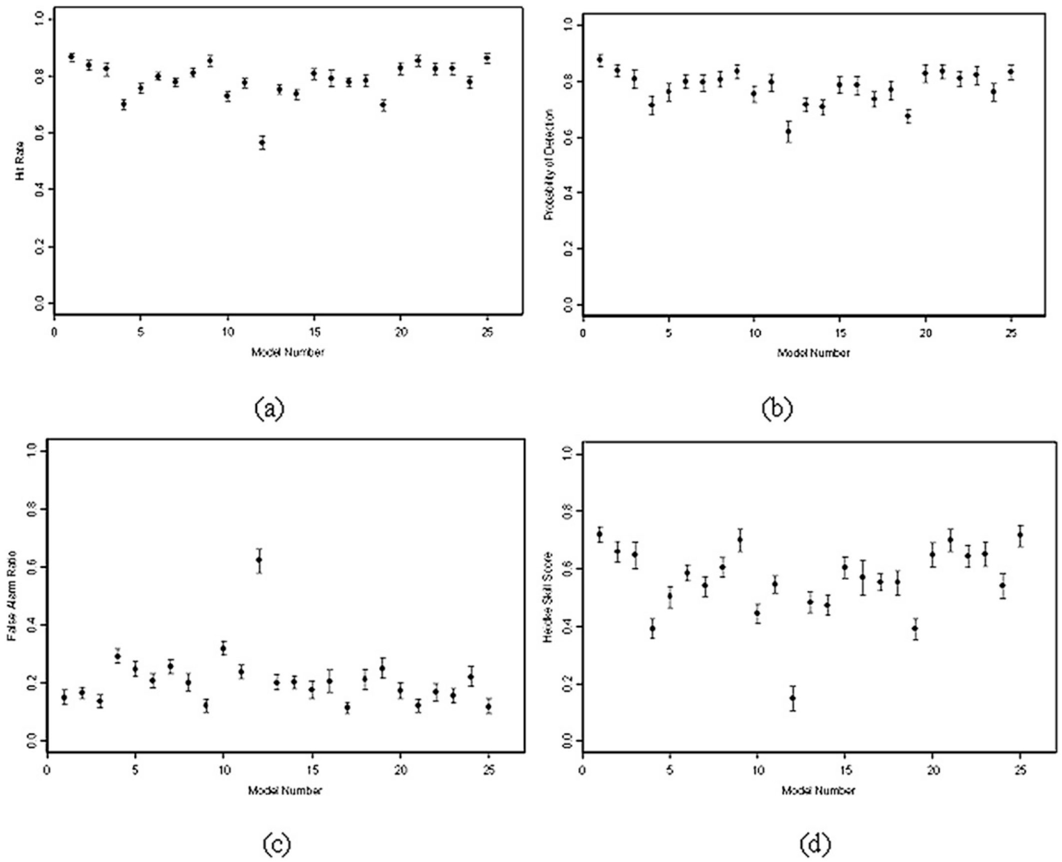

FIG. 3. Tilted bootstrap confidence intervals on the four contingency statistics—(a) HR, (b) POD, (c) FAR, and (d) HSS—for 24-h lead time and SVM. Numbers along the horizontal axis correspond to the model numbers given in Table 6.

increased lead time. Reductions were typically less than 15% (i.e., Table 8). When considering SVMs, numerous covariate combinations (more than 5) were within the 95% confidence limit of the combination with the largest median POD and HR [0–1-km SREH, 0–1- and 0–6-km bulk shear, bulk Richardson number shear, and lifted condensation level (LCL)]. Log*R* produced two covariate combinations that were within the 95% limit of the top combination (one that included all covariates from Table 5 and one that only culled LCL). The 48-h FAR results were optimal for one covariate combination using SVMs (which was not the same combination as for POD and HR), whereas Log*R* FAR results were optimal for numerous combinations. Numerous SVM covariate combinations were statistically the same as the best combination when considering HSS, whereas only two were statistically similar with Log*R*. Interestingly, SVMs performed better when more covariates were culled (usually 0–3-km SREH and another covariate), whereas Log*R* results were best when only one covariate (0–3-km SREH) was withheld. This result is attributed to the tendency toward linearity of the data with increased lead time and the inclusion of multiple highly correlated covariates that contain redundant information.

When comparing two of the top covariate combinations for SVM and Log*R* (Table 8), the magnitudes of the contingency statistics for Log*R* were larger (or smaller

TABLE 7. Intertechnique comparison of the three methods employed for classification. The boldfaced contingency statistics represent the best technique for the given contingency statistic. The numbers are the jackknife contingency results for the statistics.

	2.5% limit	Median	97.5% limit
		HR	
SVM	0.845	0.862	0.880
Log*R*	0.825	0.843	0.859
		POD	
SVM	0.806	0.832	0.857
Log*R*	0.811	0.840	0.867
		FAR	
SVM	0.093	0.116	0.145
Log*R*	0.110	0.141	0.163
		HSS	
SVM	0.678	0.728	0.752
Log*R*	0.646	0.681	0.713

TABLE 8. As in Table 7, but for 48-h lead time.

	2.5% limit	Median	97.5% limit
		HR	
SVM	0.779	0.800	0.821
Log*R*	0.818	0.837	0.856
		POD	
SVM	0.768	0.801	0.832
Log*R*	0.816	0.840	0.865
		FAR	
SVM	0.167	0.191	0.216
Log*R*	0.122	0.153	0.186
		HSS	
SVM	0.551	0.593	0.634
Log*R*	0.632	0.670	0.708

TABLE 9. As in Table 7, but for 72-h lead time.

	2.5% limit	Median	97.5% limit
		HR	
SVM	0.686	0.706	0.726
Log*R*	0.710	0.734	0.758
		POD	
SVM	0.659	0.689	0.716
Log*R*	0.671	0.703	0.732
		FAR	
SVM	0.238	0.269	0.302
Log*R*	0.169	0.199	0.230
		HSS	
SVM	0.372	0.410	0.448
Log*R*	0.423	0.470	0.515

with FAR) than the results from SVM. Additionally, all Log*R* results were statistically superior to the results obtained from SVM at 48-h lead time. Hence, Log*R* is conclusively the superior method of outbreak classification at 48-h lead time.

c. 72-h results

Further degradation (up to 15%) of the contingency statistic results was observed at 72-h lead time from both methods. By 72 h, single covariates were better at classification using both methods than combinations of covariates. However, large variance in the bootstrap replicates (not shown) was observed, so these results were artificially inflated and not ideal for this work. Thus, any results that involved employing individual covariates for classification were rejected. Both SVMs and Log*R* produced the best POD and HR results when culling 0–1-km EHI. This covariate combination was statistically similar to the best combination for FAR; when considering the HSS, rejection of 0–1-km EHI produced the best results. Hence, it was possible at 72 h to obtain a single covariate combination that was superior for the given dataset. This result supports previous conclusions about the capability of thermodynamic parameters to classify outbreak type.

The two methods were compared using the covariate combination that rejected 0–1-km EHI (Table 9). The confidence intervals given in Table 9 reveal that Log*R* is superior to SVMs in producing the lowest HR, FAR, and HSS. The two are indistinguishable at a 95% confidence level when comparing POD. However, because Log*R* is superior in the other three contingency statistics and has the greatest skill, it was deemed the superior method at 72 h for outbreak classification.

To assess the skill of the methods with increased lead time, the confidence limits of the HSS for the two statistical methods were considered. These were plotted against lead time (Fig. 4), and the results revealed a statistically significant drop-off of HSS at each lead time with SVMs. The Log*R* results did not show a statistically significant drop of HSS between 24 and 48 h, but by 72 h the results had degraded considerably. The skill degradation, although significant, is still less than hypothesized, so additional lead times will be tested in future work to assess any further loss in skill.

4. Summary and conclusions

The scope of this project was to determine the ability to discriminate between significant TOs and NTOs using model output when the model is initialized with the synoptic-scale signal. To test the influence of the synoptic scale, the WRF was initialized with the NCEP–NCAR reanalysis data, which lie on a 2.5° latitude–longitude grid. A total of 15 severe weather parameters, referred to herein as covariates, were computed from WRF model output on an 18-km grid. Because many of the covariates were highly correlated, a smaller set of covariates (6 or 7, depending on the lead time considered) was determined using permutation testing. Two statistical methods, SVM and Log*R*, were employed to distinguish between outbreak types using the reduced covariate set as input. Jackknife cross validation was performed to determine contingency statistics, and bootstrap resampling of the jackknife cross-validation results provided confidence limits on the four statistics (HR, POD, FAR, and HSS). Analyses were performed for 24-, 48-, and 72-h forecasts.

Neither SVM nor Log*R* could be singled out as the best technique at 24 h for all contingency statistics, but the optimal covariate sets suggested by both methods rejected the product of 0–1-km bulk shear and CAPE and surface-based CIN. This result supports previous conclusions that thermodynamic quantities such as CAPE and CIN are unable to distinguish outbreak type (e.g., Doswell and Evans 2003; Johns et al. 1993). The

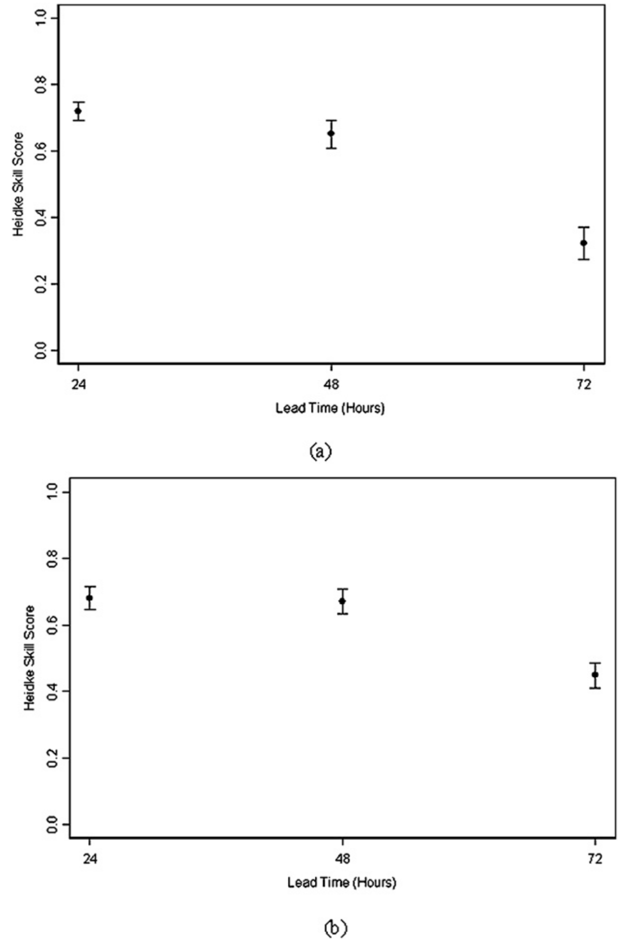

FIG. 4. Median and confidence intervals of HSS with lead time for (a) SVMs and (b) Log*R*.

results at 48 h did not reveal individual covariate combinations as ideal but instead suggested many different statistically superior combinations. The 48-h confidence limits of Log*R* were statistically better to 95% confidence than SVM, revealing Log*R* as the superior 48-h classification method for this dataset. At 72 h, Log*R* was statistically superior to SVM, and a single covariate combination (which culled 0–1-km EHI from the base set in Table 5, bottom section) was optimal for both methods. Thus, Log*R* was the superior method at both 48- and 72-h lead times.

Some degradation of the HSS (5%–10%) was noted between 24- and 48-h lead times, which is expected owing to forecast uncertainty increasing with increasing lead time. This degradation was not statistically significant to a 95% confidence level with Log*R*, but the drop-off was significant to the 95% level for SVM. By 72 h, a significant (to a 95% confidence) drop-off of the HSS was observed from both methods. The authors hypothesize that the skill will decrease further as lead time increases, and additional lead times may be considered in future work. Overall, these results suggest that model-predictable covariates on the synoptic scale seem to play a substantial role in the occurrence and type of severe weather outbreak that occurs. Outbreak discrimination using model forecasts initialized with synoptic-scale data is potentially reliable several days before the outbreak, and further consideration into a prognostic application of this work should be undertaken.

Future work in this research will include modifying the case list to comprise only those cases that occur in the same season(s), owing to seasonal dependence of some of the covariates considered (CAPE, CIN, etc.). This problem likely has artificially inflated the initial results (Shafer et al. 2009), so it is important to account for any seasonal dependence by ensuring a seasonally uniform dataset. Adding cases to the training and testing phase allows for more robust statistical models, which likely would improve results. However, consideration of additional null cases (those that do not produce an outbreak) or marginal cases (those that could be classified as a TO or an NTO) will significantly increase the challenge of classifying outbreak type. One such challenge associated with the addition of these cases is the determination of new optimal covariate sets for these marginal and null cases. Forecast applications eventually will be considered, because the probability of a TO or NTO can be obtained from Log*R* and SVMs, although the increased difficulty of classifying null cases from TOs and NTOs may require significant modification of the current methodology.

Acknowledgments. We thank Dr. Peter Lamb for support in this research. We also thank Dr. Theodore Trafalis for assistance with tuning support vector machines and Dr. Kelvin Droegemeier for providing additional insight as to the best methods to approach this problem. Dr. Henry Neeman and OSCER provided computing resources for running the WRF simulations. This work is funded by Grant ATM-0527934. Finally, we wish to thank the reviewers for their help in improving this work.

APPENDIX A

Support Vector Machine Description

An SVM is a learning method that defines a decision hyperplane for classification. A decision hyperplane (analogous to a decision line in linear regression) for the higher dimensional problem is obtained, and the classification is performed based on this plane. According to Haykin (1999), the decision hyperplane can be given as

$$\mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0, \tag{A1}$$

where **w** is a vector of weights, **x** represents the covariates, and b is an intercept. For classification, this decision hyperplane can be formulated as

$$\begin{aligned}\mathbf{w}^{\mathrm{T}}\mathbf{x} + b &> 0 \quad \text{for} \quad y = 1\\ \mathbf{w}^{\mathrm{T}}\mathbf{x} + b &< 0 \quad \text{for} \quad y = -1,\end{aligned} \tag{A2}$$

where 1 and -1 represent the two classes. To discriminate best between the two classes, the separation between the points nearest the separating hyperplane must be maximized. This leads to a quadratic programming optimization problem given by

$$\min\varphi(\mathbf{x}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$
$$\text{subject to} \quad y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 \quad i = 1 \ldots l. \tag{A3}$$

This problem can be solved by first determining the Lagrangian, which is given by

$$L(w, b, \Lambda) = \frac{1}{2}\|w^2\| - \sum_{i=1}^{l}\lambda_i[y_i(wx_i + b) - 1], \tag{A4}$$

where the values of λ_i are Lagrange multipliers. The optimality conditions of (A4) are given by

$$\begin{aligned}\frac{\partial L(\mathbf{w}, b, \Lambda)}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{i=1}^{l}\lambda_i y_i \mathbf{x}_i = 0\\ \frac{\partial L(\mathbf{w}, b, \Lambda)}{\partial b} &= -\sum_{i=1}^{l}\lambda_i y_i = 0,\end{aligned} \tag{A5}$$

so that the optimal weights **w*** are

$$\mathbf{w}^* = \sum_{i=1}^{l}\lambda_i y_i \mathbf{x}_i. \tag{A6}$$

Substituting (A6) into (A4) gives the dual formulation of this quadratic optimization problem:

$$\max F(\Lambda) = \sum_{i=1}^{l}\lambda_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\lambda_i\lambda_j y_i y_j \mathbf{x}_i \mathbf{x}_j$$
$$\text{subject to} \quad \sum_{i=1}^{l}\lambda_i y_i = 0 \quad \lambda_i \geq 0, \tag{A7}$$

which is the SVM dual problem that is solved in this study.

Many datasets that use SVMs are not linearly separable (i.e., a separating hyperplane cannot be found). In such cases, the data are input into a kernel function, which increases the dimensionality of the data so that a separating hyperplane can be found (Cristianini and Shawe-Taylor 2000; Schölkopf and Smola 2002). This is an appealing characteristic of kernel methods and a powerful way to handle nonlinearity in the data. Multiple SVM experiments are conducted to determine the kernel function that provides the largest discrimination ability. Some examples of kernel functions include

1) polynomial

$$k(\mathbf{x}, y) = (\mathbf{x}^{\mathrm{T}}y + 1)^p, \tag{A8}$$

2) radial basis function

$$k(\mathbf{x}, y) = e^{(-1/2\sigma^2)\|\mathbf{x}-y\|^2}, \quad \text{and} \tag{A9}$$

3) tangent hyperbolic

$$k(\mathbf{x}, y) = \tanh(\beta_o \mathbf{x}^{\mathrm{T}}y + \beta_1). \tag{A10}$$

APPENDIX B

Contingency Statistic Description

To measure the performance of a classification scheme, contingency statistics (Wilks 1995) are computed on the results from the statistical techniques. The contingency statistics require the creation of a contingency table (Table B1). Four contingency statistics are computed from the contingency table (Table B1) in the present study to determine classification performance. The hit rate is given as

$$\mathrm{HR} = \frac{a + d}{n}. \tag{B1}$$

TABLE B1. A sample contingency table. The top row of the contingency table showed the number of correctly classified TOs (a) and the number of predicted TOs when an NTO was observed (b). The bottom row gives the number of TOs observed when an NTO resulted from the algorithm (c) and the number of correctly classified NTOs (d).

		Obs	
Forecast		Yes (1)	No (0)
	Yes (1)	a	b
	No (0)	c	d

This statistic measures the number of correct yes (tornado outbreak) classifications and no (nontornado outbreak) classifications, but it gives no insight into the errors associated with the techniques.

The probability of detection is

$$\mathrm{POD} = \frac{a}{a + c}. \tag{B2}$$

This statistic provides a measure of the number of correct tornado outbreak classifications versus the total number of tornado outbreak classifications. Higher POD values suggest better classification for the statistical technique.

The false-alarm ratio is given as

$$\mathrm{FAR} = \frac{b}{a + b}, \tag{B3}$$

and it provides a measure of the number of classifications of a tornado outbreak when one did not occur. A smaller value of the FAR indicates lower false alarms of tornado outbreaks, which is desirable.

The final contingency statistic considered is the Heidke skill score:

$$\mathrm{HSS} = \frac{2(ad - bc)}{(a + c)(c + d) + (a + b)(b + d)}. \tag{B4}$$

The HSS provides a skill measure to the discrimination methods employed herein. Values closer to 1 are desirable.

REFERENCES

Billet, J., M. DeLisi, B. G. Smith, and C. Gates, 1997: Use of regression techniques to predict hail size and the probability of large hail. *Wea. Forecasting,* **12,** 154–164.

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.

Carr, J. A., 1952: A preliminary report on the tornadoes of March 21–22, 1952. *Mon. Wea. Rev.,* **80,** 50–58.

Colquhoun, J. R., and P. A. Riley, 1996: Relationships between tornado intensity and various wind and thermodynamic variables. *Wea. Forecasting,* **11,** 360–371.

Cristianini, N., and J. Shawe-Taylor, 2000: *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, 189 pp.

Davies, J. M., 2004: Estimations of CIN and LFC associated with tornadic and nontornadic supercells. *Wea. Forecasting,* **19,** 714–726.

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

——, 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.

——, 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.

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

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

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

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

Galway, J. G., 1975: Relationship of tornado deaths to severe weather watch areas. *Mon. Wea. Rev.,* **103,** 737–741.

——, 1977: Some climatological aspects of tornado outbreaks. *Mon. Wea. Rev.,* **105,** 477–484.

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

Grazulis, T. P., 1993: *Significant Tornadoes 1680-1991*. Environmental Films, 1326 pp.

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

Hart, J. A., 1993: SVRPLOT: A new method of accessing and manipulating the NSSFC severe weather database. Preprints, *17th Conf. on Severe Local Storms,* St. Louis, MO, Amer. Meteor. Soc., 40–41.

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

Hong, S. Y., and H. L. Pan, 1996: Nonlocal boundary layer vertical diffusion in a medium-range forecast model. *Mon. Wea. Rev.,* **124,** 2322–2339.

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

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

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

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

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

Kelly, D. L., J. T. Schaefer, R. P. McNulty, C. A. Doswell, and R. F. Abbey, 1978: An augmented tornado climatology. *Mon. Wea. Rev.,* **106,** 1172–1183.

Lanckriet, G. R. G., L. El Ghaoui, C. Bhattacharyya, and M. I. Jordan, 2002: A robust minimax approach to classification. *J. Mach. Learn. Res.,* **3,** 555–582.

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

MacKay, D., 1992: The evidence framework applied to classification networks. *Neural Comput.,* **4,** 720–736.

Markowski, P. M., 2002: Mobile mesonet observations on 3 May 1999. *Wea. Forecasting,* **17,** 430–444.

Marzban, C., and G. J. Stumpf, 1996: A neural network for tornado prediction based on Doppler radar-derived attributes. *J. Appl. Meteor.,* **35,** 617–626.

——, E. D. Mitchell, and G. J. Stumpf, 1999: The notion of "best predictors": An application to tornado prediction. *Wea. Forecasting,* **14,** 1007–1016.

McGinley, J. A., S. C. Albers, and P. A. Stamus, 1991: Validation of a composite convective index as defined by a real-time local analysis system. *Wea. Forecasting,* **6,** 337–356.

McNulty, R. P., 1995: Severe and convective weather: A central region forecasting challenge. *Wea. Forecasting,* **10,** 187–202.

Mercer, A. E., and M. B. Richman, 2007: Statistical differences of quasigeostrophic variables, stability, and moisture profiles in North American storm tracks. *Mon. Wea. Rev.,* **135,** 2312–2338.

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

Michaels, P. J., and R. B. Gerzoff, 1984: Statistical relations between summer thunderstorm patterns and continental midtropospheric heights. *Mon. Wea. Rev.,* **112,** 778–789.

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

Pautz, M. E., 1969: Severe local storm occurrences, 1955-1967. ESSA Tech. Memo. WBTM FCST12, 3–4.

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

Reap, R. M., and D. S. Foster, 1979: Automated 12–36 hour probability forecasts of thunderstorms and severe local storms. *J. Appl. Meteor.,* **18,** 1304–1315.

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

Roebber, P. J., D. M. Schultz, and R. Romero, 2002: Synoptic regulation of the 3 May 1999 tornado outbreak. *Wea. Forecasting,* **17,** 399–429.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting,* **5,** 570–575.

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

Schmeits, M. J., K. J. Kok, and D. H. P. Vogelezang, 2005: Probabilistic forecasting of (severe) thunderstorms in the Netherlands using model output statistics. *Wea. Forecasting,* **20,** 134–148.

Schölkopf, B., and A. Smola, 2002: *Learning with Kernels*. MIT Press, 650 pp.

Shafer, C. M., 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.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers, 2005: A description of the Advanced Research WRF version 2. NCAR Tech. Note NCAR/TN-468+STR, 88 pp.

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 destructive potential index—A method for comparing tornado days. Preprints, *19th Conf. on Severe and Local Storms,* Minneapolis, MN, Amer. Meteor. Soc., 280–282.

Trafalis, T. B., B. Santosa, and M. B. Richman, 2005: Learning networks for tornado forecasting: A Bayesian perspective. *Proc. Sixth Int. Conf. on Data Mining,* Skiathos, Greece.

Weisman, M. L., and J. B. Klemp, 1984: The structure and classification of numerically simulated convective storms in directionally varying wind shears. *Mon. Wea. Rev.,* **112,** 2479–2498.

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