THE UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

DISCRIMINATION OF TORNADIC AND NON-TORNADIC SEVERE WEATHER OUTBREAKS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

BY

ANDREW EDWARD MERCER
Norman, Oklahoma
2008

DISCRIMINATION OF TORNADIC AND NON-TORNADIC SEVERE WEATHER OUTBREAKS

A DISSERTATION APPROVED FOR THE
SCHOOL OF METEOROLOGY

By

Michael B. Richman, Chair

Charles A. Doswell III

Kelvin K. Droegemeier

Lance M. Leslie

Theodore B. Trafalis

ACKNOWLEDGEMENTS

I wish to thank my advisor, Dr. Michael Richman, for constantly supporting me through this process and for allowing me to work with him on this project. I would also like to thank my colleague, Chad Shafer, for his time spent on providing the WRF simulations which were used in the objective discrimination in this study, as well as for his insight into many other aspects of the project. I wish to thank Dr. Charles Doswell III for his help in the writing of this document, as well as his assistance in understanding the physical processes that are underlying the results below. I wish to thank Dr. Lance Leslie for his constant support and positive reinforcement through this process, as well as his numerical modeling insight into the project. I would like to acknowledge my other two committee members, Dr. Theodore Trafalis and Dr. Kelvin Droegemeier, for their time spent reading my documents and their assistance in my completing my degree. Thank you for CIMMS for assistance in my financial issues. Thank you to OSCER for providing the supercomputing facilities for which the WRF simulations were conducted. Finally, I wish to thank my wife Jaimie for her constant support as I went back to school to get my doctorate, and my extended family for their constant support and encouragement through this process. Jaimie, you are a blessing to my life and I could not have accomplished what I have without your support. I also want to acknowledge my daughter Albany for providing distractions when I needed a break from this work and for being such a wonderful girl. I am truly blessed to have you in my life.

TABLE OF CONTENTS

ABSTRACT

Outbreaks of severe weather affect the majority of the conterminous United States. An outbreak is characterized by multiple severe weather occurrences within a single synoptic system. Outbreaks can be categorized by whether or not they produce tornadoes. It is hypothesized that the antecedent synoptic signal contains important information about outbreak type. Accordingly, the scope of this research is to determine the extent that the synoptic signal can be utilized to classify outbreak type at various lead times.

Outbreak types are classified using the NCEP/NCAR reanalysis data, which are arranged on a global 2.5º latitude-longitude grid, include 17 vertical pressure levels, and span from 1948 to the present (2008). Fifty major tornado outbreak (TO) cases and fifty major non-tornadic severe weather outbreak (NTO) cases are selected for this work. Two types of analyses are performed on these cases to assess discrimination ability. One analysis involves outbreak classification using the Weather Research and Forecasting (WRF) model initialized with the NCEP/NCAR reanalysis dataset. Meteorological covariates are computed from the WRF output and used in training and testing of statistical classification models. The covariate fields are depicted on a 21 X 21 gridpoint field with an 18 km grid spacing centered on the outbreak. Covariates with large discrimination potential are determined using permutation testing. A P-mode principal component analysis (PCA) is used on the subset of covariates determined by permutation testing to reduce data dimensionality, since numerous redundancies exist in the initial covariate set. Three statistical classification models are trained and tested with the resulting PC scores: a support vector machine (SVM), a logistic regression model (LogR), and a multiple linear regression model (LR). Promising results emerge from these methods, as a probability of detection (POD) of 0.89 and a false alarm ratio (FAR) of 0.13 are obtained from the best discriminating statistical technique (SVM) at 24-hours lead time. Results degrade only slightly by 72-hours lead time (maximum POD of 0.833 and minimum FAR of 0.276).

Synoptic composites of the outbreak types are the second analysis considered. Composites are used to reveal synoptic features of outbreak types, which can be utilized to diagnose the differences between classes (in this case, TOs and NTOs). The composites are created using PCA. Five raw variables, height, temperature, relative humidity, and *u* and *v* wind components, are extracted from the NCEP/NCAR reanalysis data for North America. Converging longitude lines with increasing latitude on the reanalysis grid introduce bias into correlation calculations in higher latitudes; hence, the data are mapped onto both a latitudinal density grid and a Fibonacci grid. The resulting PCA produces two significant principal components (PCs), and a cluster analysis on these PCs for each outbreak type results in two types of TOs and NTOs. TO composites are characterized by a trough of low pressure over the central United States and major quasigeostrophic forcing features such as an upper level jet streak, cyclonic vorticity advection increasing with height, and warm air advection. These dynamics result in a strong surface cyclone in most tornado outbreaks. These features are considerably less pronounced in NTOs. The statistical analyses presented herein were successful in classifying outbreak types at various lead times, using synoptic scale data as input.

1. INTRODUCTION

a) Motivation

According to the American Meteorological Society (AMS), tornadic severe weather occurs with a highest frequency over the United States (Glickman 2000). Tornadic severe thunderstorms are characterized by large, damaging hail, strong wind gusts, and tornadoes. While over 1000 tornadoes affect the United States per year (Glickman 2000), groups of these events, known as outbreaks, are comparatively uncommon events (only 20-30 outbreak days per year, Schneider et al. 2004). These outbreaks are considerably more dangerous than individual tornadoes, since they can result in multiple, significant tornado occurrences that affect a relatively large geographic region.

In addition to tornado outbreaks (hereafter TOs), numerous primarily nontornadic outbreaks of severe weather (hereafter NTOs) impact the United States annually. NTOs are more common (50 or more per year) than TOs (20-30 per year, Glickman 2000). However, NTOs are generally less threatening to life than TOs. Advance knowledge of outbreak type would aid forecasters and emergency management teams in anticipation of these dangerous events.

Many studies (section 1.2) classify different types of TOs and NTOs, but no work has appraised the potential to distinguish between these two main classes at various lead times. The scope of this project is to assess the disparities between TOs and NTOs through statistical objective methods. These goals will be accomplished through statistical outbreak classification and synoptic storm typing of TOs and NTOs.

b) Literature Review

1) SEVERE WEATHER OUTBREAK REVIEW

The AMS glossary (Glickman 2000) defines a TO as “multiple tornado occurrences within a single synoptic-scale system.” An early study of a TO (Carr, 1952) examined the significant 21-22 March 1952 event which encompassed the lower Mississippi Valley and the Tennessee Valley. This study analyzed surface features contributing to the event (low pressure system with associated cold front over Louisiana) and described significant weather occurrences produced by the event. Many classes of TOs have been defined in previous studies, including Pautz (1969), who defined TOs based on their size (small, medium, and large), and Galway (1975), who considered the number of tornado deaths by state and compared that with the Pautz (1969) outbreak definitions. Galway (1977) classified three different types of TOs: a local outbreak (those confined to radii not exceeding 10 000 square miles), a progressive outbreak (an outbreak that advances from west to east with time in which the distance between the first and last tornado report generally exceeds 350 miles), and a line outbreak (one in which the tornadic thunderstorms form along a narrow corridor). Grazulis et al. (1993) categorized TOs as groups of 6 or more tornadoes within a single synoptic system.

While many studies had grouped TOs into different categories, Doswell et al. (2006), [hereafter called D06] presented the first objective ranking of TOs based on the AMS glossary (Glickman 2000) definition of a TO. The TO database used in D06 was documented by Schafer and Edwards (1999) and included data related to individual TOs which occurred on a single day (1200 UTC through 1159 UTC the following day). D06 formulated an index (*O* - index) based on weighting different TO parameters, including total path length of all tornadoes, the destructive potential index (DPI, Thompson and Vescio 1998), the number of killer tornadoes, the number of deaths, etc. D06 found that small permutations in the weights led to significant differences in the rankings, revealing the highly subjective nature of the definition of a TO that was manifest in the numerous types of TOs in previous research.

D06 ranked NTOs in the same manner as TOs, although a different set of weighting parameters was selected. They defined an NTO as a severe weather outbreak with 6 or fewer tornadoes. The NTO ranking index (*S* – index) was formulated from a weighted sum of the total number of severe weather reports, the number of significant wind reports, the number of significant hail reports, the number of tornadoes, the number of wind reports, and the number of hail reports. Some NTO events consisted of individual smaller outbreaks from independent synoptic systems that occurred on the same day. These NTOs were geographically widespread, and a purely objective ranking of NTOs classified these events as significant, despite multiple independent synoptic systems triggering the storms. To account for this geographic distribution of the NTOs, D06 sorted the individual severe weather reports based on latitude and longitude and retained the middle 50% of the latitude-longitude distribution. They scaled the resulting area to the order of the *S*- index and subtracted this scaled variable as a new *S*- index, which thereby included information of the severe weather report spatial distribution. The D06 top 50 ranked cases of TOs and NTOs were used in the present research for the statistical classification and compositing.

Many outbreak studies, including D06, have ranked the 3 April 1974 "superoutbreak" as the most important outbreak of tornadoes in recorded history, with over 100 long path significant tornadoes observed (Fig. 1). Fujita (1974) noted many synoptic precursors that led to the April 1974 TO. However, some TOs were less synoptically evident, such as the 3 May 1999 outbreak (ranked 20 in D06). Numerous investigators have investigated this TO (Roebber et al. 2002, Edwards et al. 2002, Stensrud and Weiss 2002, Thompson and Edwards 2000, others), noting that it had atmospheric features that did not suggest convection would initiate (weak dryline convergence, cirrus deck reducing instability, etc.). Accordingly, forecasters were unable to determine if convection would occur a few hours prior to initiation.

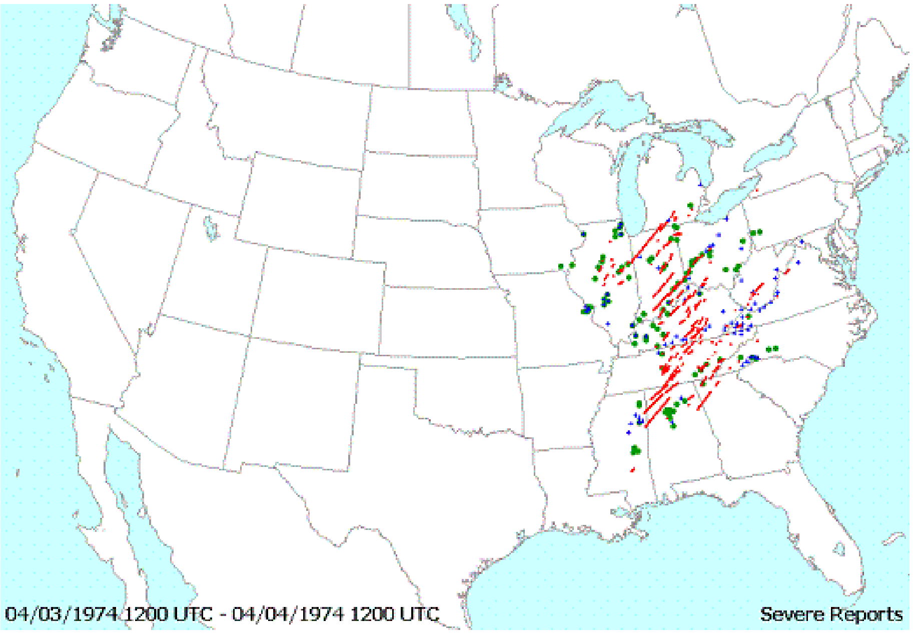


Fig. 1. Storm reports from 3 April 1974, courtesy of the Storm Prediction Center Severe Plot software (Hart 1993). Red lines represent tornado tracks; blue crosses represent severe wind reports, and green points represent severe hail reports.

Marginal outbreaks, such as 3 May 1999 TO, helped motivate the current work, since they are more poorly understood than the classic outbreaks (i.e. 3 April 1974). Whereas the physics of convective development in forecast TOs and NTOs (e.g. 3 May 1999) lies outside the scope of this project, the application of statistical methods to classify TOs and NTOs can provides a baseline to motivate such research.

2) STATISTICAL METHODS REVIEW

The classification of outbreak classes was investigated in the present study through the use of artificial intelligence (AI) techniques and statistical classification methods, all of which have been used in previous meteorological research. Some AI studies have considered severe weather problems, including Trafalis et al. (2005) [hereafter T05], who used AI to revise the mesoscale detection algorithm (MDA) on the WSR-88D Doppler radar system. The MDA, designed to detect storm-scale circulations within radar echoes, is currently used by the National Weather Service as a tool for issuing tornado warnings. The goal of T05 was to implement AI techniques to improve the ability of radar to detect a tornado signature, increasing the likelihood of a correctly issued tornado warning. Marzban and Stumpf (1996) provided the motivation for this idea through applying an artificial neural network (ANN) to the MDA.

In T05, several learning techniques were compared within the context of the MDA, ANN, support vector machines (SVM), Bayesian neural networks (BNN – MacKay 1992), and minimax probability machines (MPM - Lanckriet et al. 2002). A set of roughly 800 training samples was used for training and testing of these learning algorithms. In order to determine the sensitivity of the algorithms to tornadic events, the amount of tornado data in the test datasets varied from 2% to 10%. Multiple experiments with different statistical model parameters (i.e. cost, kernel function, etc.) were conducted, and the best model parameters were selected based on a series of forecast evaluation indices that produced the most accurate forecasts. The methodology employed herein follows closely with that of T05. The framework employed by T05 is similar to the current study.

In addition to statistical classification methods, this study presented composite fields of TOs and NTOs, which showed the physical features of each outbreak type. A commonly used compositing methodology, which was applied herein, is rooted in principal component analysis (PCA – Wilks 1995). Jones et al. (2004) used a PCA and a binary classification on 100 000 MDA instances to observe the MDA’s tornado detection capability. In their study, many aspects of the MDA were shown to be useful for tornado detection, including the neural network tornado detection algorithm (Marzban and Stumpf 1996), the mesocyclone strength index, maximum gate-to-gate velocity difference, the mesocyclone depth, and the mesocyclone rank. Lanicci and Warner (1991) performed a mean composite analysis on severe weather soundings to search for the type 1 tornado sounding (Fawbush and Miller 1952). Their work used mean severe weather parameters to analyze the temporal development of the type 1 sounding. They found a relationship between the intensity of the type 1 sounding and the intensity of the associated severe weather. Schaefer and Doswell (1984) employed empirical orthogonal functions (EOFs – Wilks 1995) in the creation of synoptic storm types of TOs. The EOFs revealed different synoptic features of different TO storm types. An updated synoptic storm typing approach was applied herein to determine NTO and TO types, and these types will help accomplish the main goals of the present research.

c) Objectives

The scope of this investigation is to assess the ability to discriminate between TOs and NTOs using primarily objective methods. It is hypothesized that the synoptic-scale signal contains pertinent information of the impending outbreak type, but the details of this relationship are not well understood (Doswell and Bosart 2001). To specify the details of this relationship, synoptic-scale data were used for initial input into the statistical and numerical methods. A set of 50 significant TOs and NTOs were classified by statistical methods and synoptic storm typing in order to determine if the capability to distinguish between the two exists. It is important to emphasize that *all* cases selected included an outbreak, and the study determined the ability of the synoptic scale input data to classify the outbreak type. Null cases (no outbreak) or weakly severe outbreaks were not tested. If the outbreak classification of the distinct events is successful, it is of interest to know how far in advance of the outbreak the classification performs well. If the methods used in this study cannot distinguish between these extremely distinct TO and NTO outbreak types, further investigation into marginal TOs and NTOs or null cases (outbreak versus no outbreak) likely would not be warranted. Accordingly, the present work sets a baseline for future research on this topic.

The objective statistical classification of outbreak type involved a binary decision, since two only outbreak types were considered. Three statistical methods, SVM, logistic regression (LogR), and linear regression (LR), were tested to document the method which classifies with the most success. These three methods were selected since they include a linear technique, a non-linear technique working in a low dimensional space, and a non-linear technique working in a high dimensional space. To facilitate description of the severe weather atmosphere, which is typically done in a mesoscale framework, 17 covariates (Brown and Murphy 1996) computed using output from the Weather and Research Forecast (WRF) model (Skamrock et al. 2005) were considered for the statistical classification. These covariates are severe weather parameters which are often used to describe the severe weather environment . Numerous combinations of covariates were analyzed to determine those with the highest classification capability.

Synoptic storm types were computed from the NCEP/NCAR reanalysis dataset (Kalnay et al. 1996) at 17 vertical levels over the continental United States. These storm types were developed to provide insight into the synoptic precursors of TOs and NTOs. Five raw variables were included in the composites, including temperature, relative humidity, height, *u*-component wind, and *v*-component wind. These statistical methods provided an excellent capability to discriminate these distinct TOs and NTOs, which set the baseline for additional research on outbreak classification.

The methods used in creation of the synoptic storm types and in the statistical classification are given in Chapter 2. Chapter 3 provides results from the statistical classification, and Chapter 4 shows the storm type results. Chapter 5 summarizes the classification capabilities of the methods presented herein.

2. METHODOLOGY

a. Data

One goal of this work is outbreak type discrimination; therefore, the top 50 ranked NTOs and TOs (Appendix A) from D06 were retained to provide the strongest (most robust) contrast for statistical analyses. All cases, except for 8 July 1980, had an outbreak valid time near 0000 UTC, (the valid time for 8 July 1980 was 1200 UTC). Therefore, that case was eliminated from the NTO set, leaving 49 NTOs.

Once a robust set of TOs and NTOs was obtained, meteorological data, from the event days, were required. One of the primary goals of this study was to determine the role of synoptic scale influences on outbreak classification based on model forecasts. To help assess these effects, an input dataset with a synoptic-scale grid spacing over the United States was needed. As a result, the NCEP/NCAR reanalysis data (Kalnay et al. 1996), which reside on a 2.5º longitude by 2.5º latitude global grid and 17 vertical levels (synoptic-scale grid spacing), were selected as source data for this study.

The NCEP/NCAR reanalysis data are based on the assimilation of model-derived quantities and observations, which results in varied reliability of the reanalysis variables. Kalnay et al. (1996) ranked the reliability of all of the NCEP/NCAR reanalysis variables based on their observational and model-derived input. Variables ranked “A” were based primarily on observations and were considered the most reliable variables in the dataset. As model-derived input was introduced into the calculation of other variables, the reliability grade was lowered to a “B” or a “C”.

Parameters based almost entirely on climatology and model-derived input were graded as “D” variables.

The NCEP/NCAR reanalysis data were used for both the objective statistical classification and the synoptic storm typing. The objective classification methodology required WRF simulations (Section 2.2.1) of the 100 cases, and several reanalysis variables were required for WRF initialization. The synoptic storm typing methodology used five reanalysis variables (temperature, relative humidity, *u*-wind, *v*-wind, and height). Since the dependability of the reanalysis variables varied, the reliability of each variable used was needed. Table 1 lists the NCEP/NCAR reanalysis variables used herein, as well as their reliability grade.

Table 1. List of variables used in WRF simulations and synoptic storm types, their level (upper air or surface) and reliability grade described by Kalnay et al. (1996). Note that some variables considered “surface” variables are near-surface (lowest sigma layer or within 30 hPa of surface pressure).

Input Variable	**(U)pper air or (S)urface**	**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 meter temperature	S	B
10 meter U-wind	S	B
10 meter V-wind	S	B
Water equivalent of accumulated snow depth	S	C

b. Objective Statistical Classification

The first analysis conducted with the NCEP/NCAR reanalysis data was a classification study, which used statistical techniques to determine the ability to discriminate outbreak type from the WRF simulations. A summary of the methods used in the objective statistical classification follows.

1) WRF SIMULATIONS

For optimal classification of outbreak type, the statistical models required comprehensive local information about the severe weather environment for each case. Raw synoptic-scale variables (height, *u*-wind, *v*-wind, temperature, etc.) do not provide this information (only a few gridpoints per case exist in the outbreak region). As a result, numerical model simulations (the WRF in the present study) were needed to obtain detailed mesoscale knowledge of each outbreak. This mesoscale output from the WRF simulations was used in the statistical classification techniques.

The WRF simulations used model physics summarized in Table 2 and employed five two-way nested domains (Fig. 3). The first ("mother") domain was fixed and had a grid spacing of 162 km. Domain 2 was positioned surrounding the contiguous United States, and had a grid spacing of 54 km. Domain 3, used in the objective statistical classification, was positioned according to the general location of the simulated outbreak and had a grid spacing of 18 km. Domains 4 and 5 had grid spacing of 6 km and 2 km, respectively, and were outbreak-relative. All domains had 31 vertical levels (Table 3), defined by the η coordinate, which was the default output for WRF. WRF required the grid spacing to decrease by a factor of 3 (other factors resulted in large model instability on all domains) with additional nested domains, so the grid spacing values were selected by increasing the spacing by a factor of 3 from storm scale (2 km, domain 5). The mother domain grid spacing of 162 km is comparable to the 2.5º native grid spacing on the NCEP/NCAR reanalysis (about 250 km).

Table 2. WRF physical schemes used by Shafer (2007) for simulation of the 100 cases. Adapted from Shafer (2007).

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)

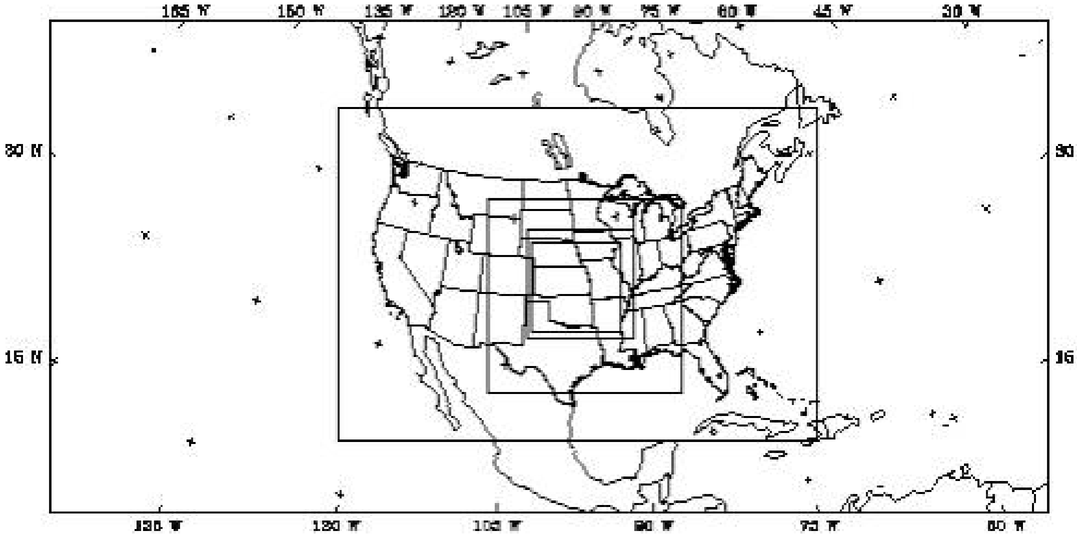


Fig. 3. A sample of the five domains used in WRF simulations by Shafer (2007) valid for 3 May 1999 (see Fig. 2 for outbreak on this day). Output from domain three centered on the outbreak was used in the objective discrimination of outbreak type. Taken from Shafer et al. (2008).

Table 3. The 31 eta levels and their corresponding pressure level using standard pressure (i.e. 1013.25 mb) as the surface pressure and 10 mb as the top of the atmosphere.

Eta Level	Pressure Level (mb)
1.000	1013.25
0.993	1006.23
0.880	993.19
0.966	979.14
0.950	963.09
0.933	946.03
0.913	925.97
0.892	904.90
0.869	881.82
0.844	856.74
0.816	828.65
0.786	798.56
0.753	765.45
0.718	730.33
0.680	692.21
0.639	651.08
0.596	607.94
0.550	561.79
0.501	512.63
0.451	462.47
0.398	409.29
0.345	356.12
0.290	300.94
0.236	246.77
0.188	198.61
0.145	155.47
0.108	188.35
0.075	85.24
0.046	56.15
0.021	31.07
0.000	10.00

2) COVARIATES

Given that the WRF simulations do not make explicit predictions of the occurrence of tornadoes, some way to distinguish between outbreak types in the simulations is necessary. To diagnose outbreak type in the statistical classification techniques, fields of meteorological *covariates* were computed from the domain 3 WRF output. Domain 3 was chosen since most of the selected covariates were commonly defined in the mesoscale (i.e. synoptic-scale and storm-scale CAPE was not desirable). Since domain 3 provided thousands of gridpoints, a smaller subdomain of domain 3 was used to narrow the analysis region for the statistical classification techniques. To accomplish this, a subjective center of each TO and NTO (Fig. 4) was determined through inspection of the storm reports as provided in the Storm Prediction Center's SeverePlot software (Hart 1993), and a subdomain of 21 X 21 gridpoints, centered on the subjective outbreak center, was preserved from the domain 3 output for each covariate. This subdomain size encompassed most TOs. Additionally, the top 50 NTOs as defined by D06 encompassed a small domain (this was a criteria in D06 for ranking the NTOs).

The covariates included 17 commonly used severe weather parameters that measured thermodynamics, shear, and vorticity (Appendix B describes each covariate in detail). The product of CAPE and bulk shear (Appendix B.10) is a covariate which has not been considered in the literature previously, but is included to provide another measure combining instability and shear.

One issue in the covariate computation was noted. The WRF computation of surface based CIN was suspect, since CIN typically forms in the lowest 1-2 km. This

1-2 km depth includes 8-10 WRF vertical levels (Table 3) which is incapable of resolving CIN accurately. This might have an impact on how effectively CIN would serve as a useful covariate.

The 17 covariates selected were the base set of parameters for the statistical classification. Many of these covariates are highly correlated (e.g., the different EHI values had a Pearson correlation higher than 0.98 for many TOs), suggesting redundancies in the variables, that could cause instability in subsequent statistical analyses. Many redundant covariates (such as those that were considered over multiple layers), were removed by permutation testing.

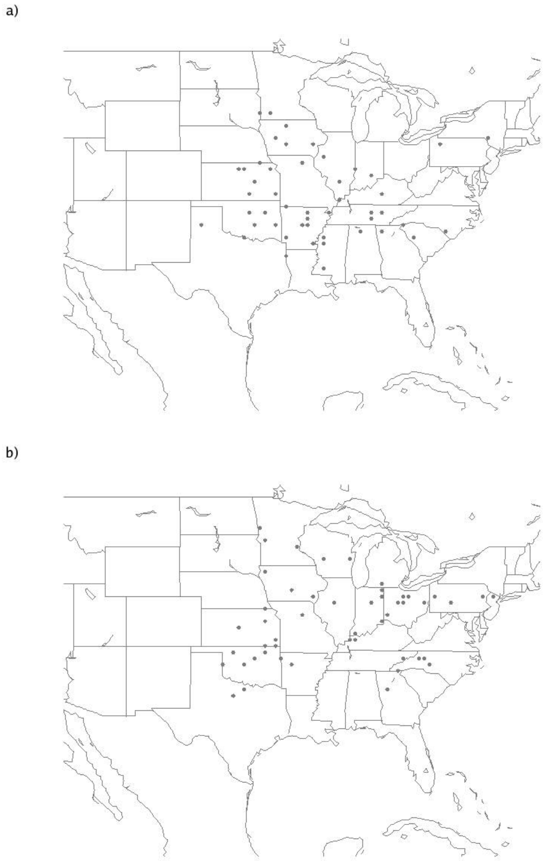


Fig. 4. Outbreak centers determined subjectively using SeverePlot. Panel (a) represents TOs, while panel (b) represents NTOs. Some overlap in points exists, leading to fewer than 50 points per panel.

3) PERMUTATION TESTING

According to Efron and Tibshirani (1993), a permutation test determines if the means of two data samples are the same at a statistical significance level of the user's choosing. The null hypothesis (H_0) for the permutation test is that the mean of the two samples is the same, or that their mean difference is zero. Figure 5illustrates the permutation testing method. Initially, two separate data distributions (the dark gray pools) are tested. The mean of each pool is computed first, and the difference between the two means is stored. The two data distributions are then combined into a single pool (the white pool), and two random permutations are sampled with replacement from the pool. The mean difference between these two permutations is stored and compared to the initial mean difference. If the difference of the permutation means is larger than the initial mean difference, the permutation is counted toward the p-value. This process is repeated many times (1000 times), and the percentage of permutations that are counted is the corresponding probability that the two distributions are the same, known as a *p*-value. *P*-values closer to zero represent a higher probability that one can reject H_0. The permutation test, unlike other commonly used hypothesis tests, such as the *t*-test, does not make assumptions of the data distribution. Since the distribution of each covariate is unknown, this property of the permutation test makes it ideal for the present work.

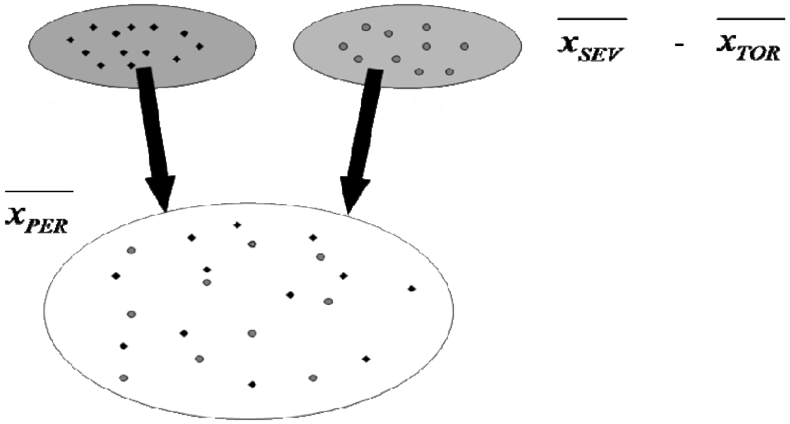


Fig. 5. Illustration of the permutation methodology presented above.

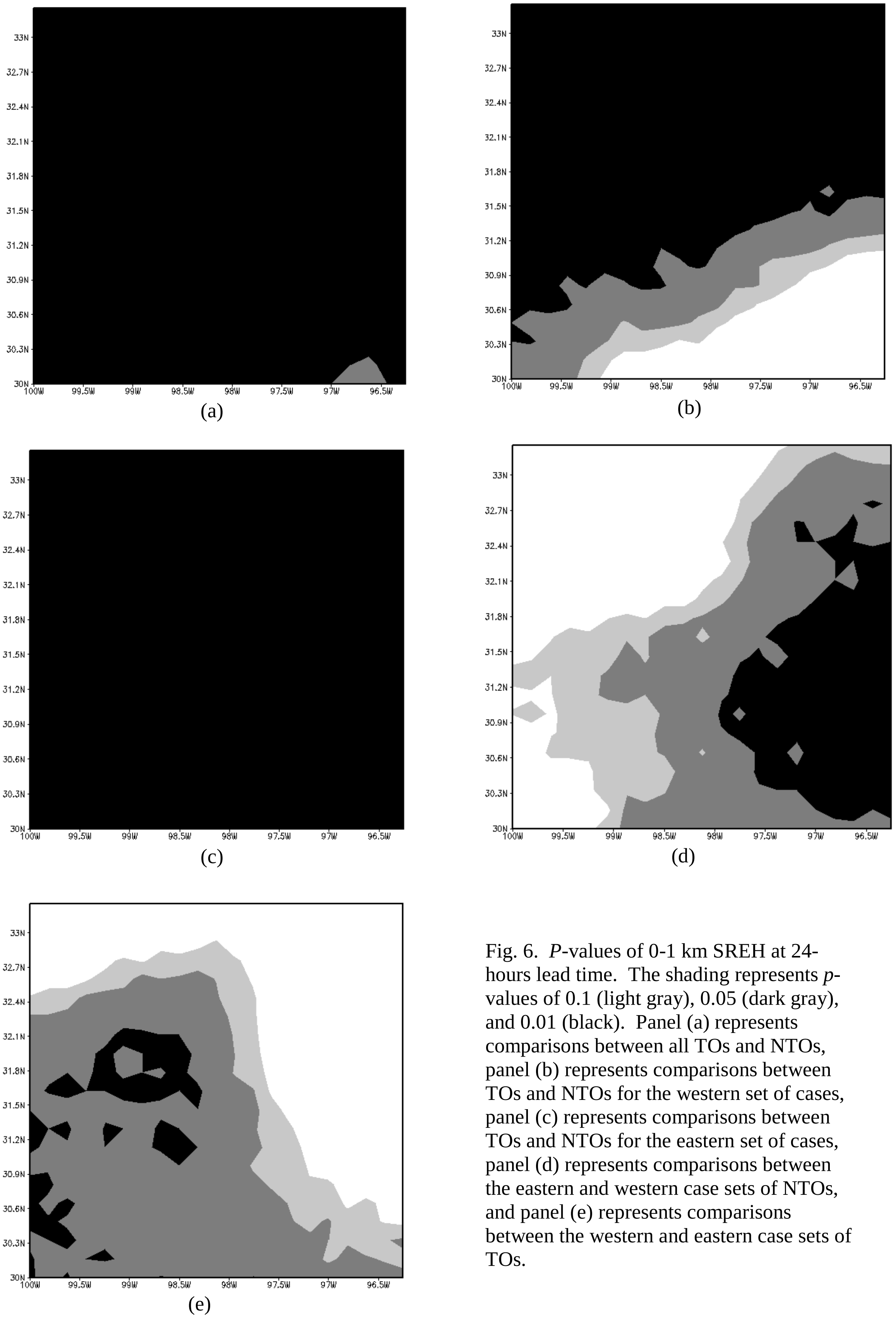
Permutation testing was used to determine each covariate's ability to distinguish between TOs and NTOs. Those covariates which showed large regions of small p-values were described as proficient outbreak discriminators. All NTOs and TOs were tested initially. In addition to the entire case set, cases west and east of the Mississippi River (Fig. 4) were tested separately, to show which covariates discriminate best in each geographic region. Finally, any regional dependence within each outbreak type was tested (west cases versus east cases for each outbreak type). The intra-outbreak regional tests revealed covariates whose magnitudes strongly depended on geographic region (an undesirable property for this study).

Since fields of the covariates resulted from the WRF simulations, permutation testing on a gridpoint by gridpoint basis was performed, with each gridpoint assigned a *p*-value based on the results of the test. *P*-values were plotted on the 21 X 21 covariate grid, with values of 0.1, 0.05, and 0.01 displayed. These three *p*-values are consistently used to show statistical significance throughout the literature (Daley and Chervin 1985, Wilks 1996).

Figure 6 shows an example of a covariate that exhibits statistically significant differences (0-1 km SREH) at 24-hours lead time. The gray and black colors in panels a – c of Fig. 6 represent regions of statistical significance, which indicate areas where the covariate discriminates outbreak type successfully. Panels d and e, which represent the regional dependence of each covariate, show multiple p-values, indicating that 0-1 km SREH has modest regional dependence. However, the modest regional dependence of 0-1 km SREH was not significant enough to ignore its good discrimination capabilities, resulting in retaining this covariate.

To contrast a covariate capable of outbreak classification, Fig. 7 illustrates permutation test results from a poor classifier, surface based CAPE. The southern portion of the domain in panel a shows some differences (darker colors), but most of the region is not statistically significant. These results were observed when considering eastern and western outbreaks as well (panels b and c). Little regional dependence of CAPE was observed (panels d and e), but the limited discrimination capabilities of CAPE, as revealed by the permutation testing, led to rejection of this covariate from use in the objective statistical classification.

Similar analyses to those presented above were conducted for all covariates at 24-hour, 48-hour, and 72-hour lead times, in order to determine the best covariate set for the objective discrimination. Percentages of the fields which were significant to a particular *p*-value were tabulated and used to reduce the base set of covariates to those best suited for outbreak classification at each lead time.



(a)

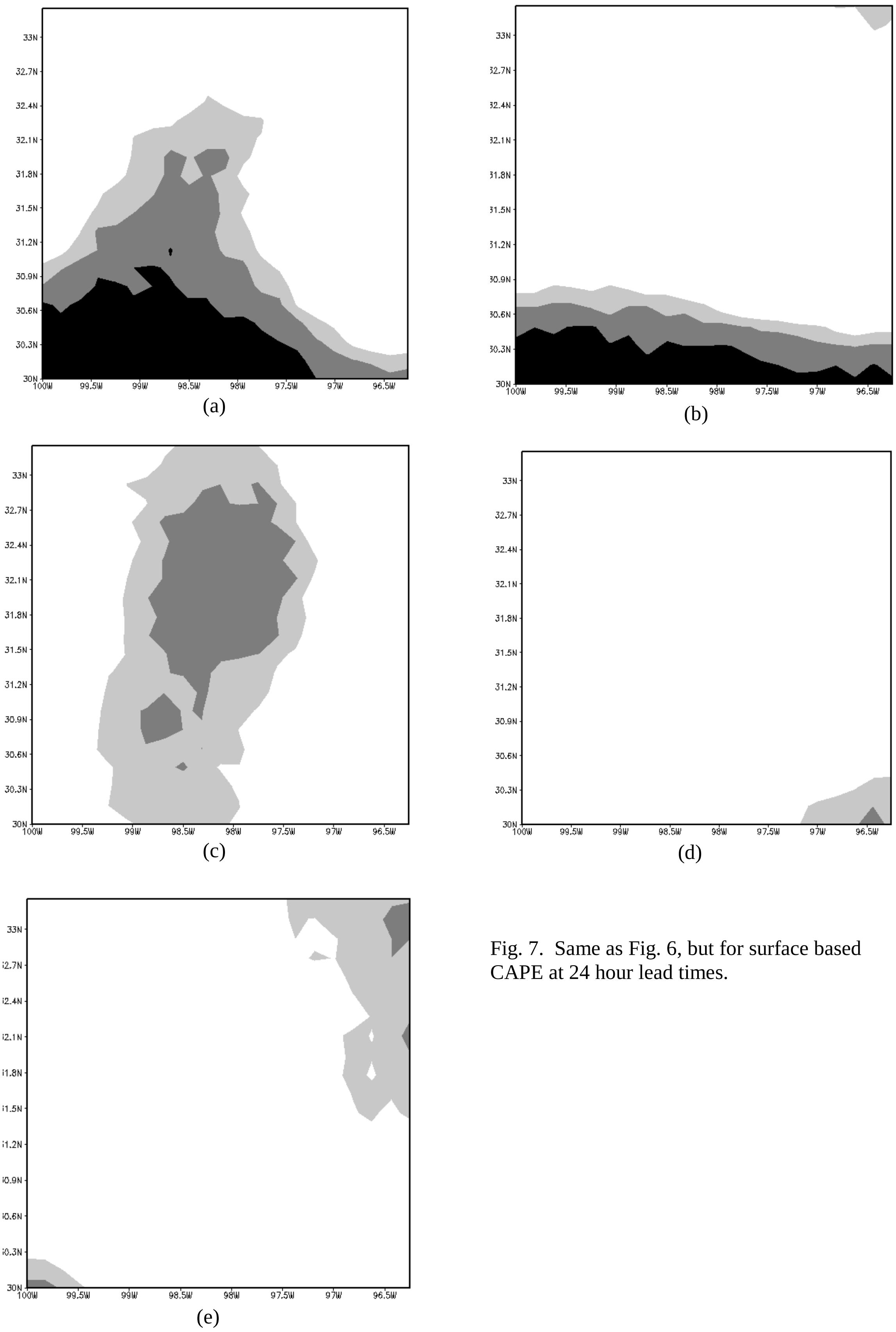
(b)

(c)

(d)

(e)

Fig. 6. *P*-values of 0-1 km SREH at 24-hours lead time. The shading represents *p*-values of 0.1 (light gray), 0.05 (dark gray), and 0.01 (black). Panel (a) represents comparisons between all TOs and NTOs, panel (b) represents comparisons between TOs and NTOs for the western set of cases, panel (c) represents comparisons between TOs and NTOs for the eastern set of cases, panel (d) represents comparisons between the eastern and western case sets of NTOs, and panel (e) represents comparisons between the western and eastern case sets of TOs.



(a)

(b)

(c)

(d)

(e)

Fig. 7. Same as Fig. 6, but for surface based CAPE at 24 hour lead times.

A summary of a percentage of gridpoints significant to each *p*-value (0.1, 0.05, and 0.01) at 24 hours lead time is given in Table 4. As previously discussed, SREH at 0-1 km at 24-hours lead time was retained for the statistical outbreak classification owing to its low *p*-values in the discrimination fields, whereas surface-based CAPE showed little discrimination capability. Surface based CIN, SREH at 0-3 km, and LCL exhibited good discrimination ability with modest regional dependence, so these were preserved. Additional covariates show good discrimination ability with little regional biases, including 0-1 km bulk shear, 0-1 km EHI, and the product of 0-1 km bulk shear and CAPE. This smaller covariate set consists mostly of shear or vorticity parameters, which are widely considered to be good indicators of tornadic development (see Appendix B).

Table 4. Percent of gridpoints that are significant to α=0.1, α=0.05, and α=0.01. The first column uses all cases in the permutation testing, while the second uses cases east of the Mississippi River. The third column considers cases west of the Mississippi River. The fourth and fifth columns compare the western and eastern region for each outbreak type to test for regional dependence. Values near 100% in columns 1-3 and near 0% in columns 4 and 5 are best.

p ≤ 0.1					
Covariate	All	East	West	Tornado East vs West	Severe East vs West
Surface Based CAPE	39.46	37.87	23.36	13.83	2.95
Surface Based CIN	99.77	35.60	84.35	3.85	2.95
LCL	98.19	100.00	86.39	39.00	38.32
LFC	26.08	4.31	34.92	8.39	18.59
0-1 km Bulk Shear	100.00	100.00	100.00	48.30	18.82
0-3 km Bulk Shear	100.00	100.00	70.98	96.83	0.23
0-6 km Bulk Shear	100.00	100.00	61.90	98.41	40.36
0-1 km SREH	100.00	100.00	85.26	61.22	65.99
0-3 km SREH	100.00	100.00	56.92	98.87	79.82
BRN Shear	74.38	100.00	20.63	99.32	58.50
Storm Relative Flow	99.55	100.00	43.76	89.80	100.00
0-1 km EHI	88.66	100.00	58.05	17.69	54.42
0-3 km EHI	79.37	100.00	40.59	23.81	64.17
Vorticity Generation Potential	39.00	67.80	44.67	16.33	57.14
Product of 0-1 km shear and CAPE	79.59	45.35	55.10	9.52	10.43
Product of 0-3 km shear and CAPE	64.17	37.19	39.91	11.34	15.87
Product of 0-6 km shear and CAPE	61.45	46.26	28.80	8.39	35.60
p ≤ 0.05					
Covariate	All	East	West	Tornado East vs West	Severe East vs West
Surface Based CAPE	29.48	15.87	18.37	2.72	1.13
Surface Based CIN	91.84	22.22	69.84	0.00	1.13
LCL	94.56	100.00	82.31	23.13	22.22
LFC	17.46	2.49	25.17	2.72	9.07
0-1 km Bulk Shear	100.00	100.00	100.00	25.62	10.66
0-3 km Bulk Shear	100.00	100.00	63.72	93.42	0.00
0-6 km Bulk Shear	100.00	100.00	52.38	96.37	9.98
0-1 km SREH	100.00	100.00	80.27	50.11	48.98
0-3 km SREH	98.64	100.00	50.34	93.88	66.89
BRN Shear	67.35	100.00	11.79	98.64	32.20
Storm Relative Flow	97.96	100.00	37.19	78.68	100.00
0-1 km EHI	83.90	100.00	49.89	8.62	43.99
0-3 km EHI	71.66	100.00	31.07	14.06	56.69
Vorticity Generation Potential	29.93	48.53	35.15	11.34	42.86
Product of 0-1 km shear and CAPE	71.43	27.44	45.35	3.40	2.27
Product of 0-3 km shear and CAPE	48.53	25.17	25.17	5.22	7.71
Product of 0-6 km shear and CAPE	47.85	35.37	15.42	4.54	23.36
p ≤ 0.01					
Covariate	All	East	West	Tornado East vs West	Severe East vs West
Surface Based CAPE	16.10	0.23	11.34	0.00	0.00
Surface Based CIN	50.79	4.54	26.98	0.00	0.00
LCL	84.35	93.65	70.29	9.52	2.72
LFC	4.08	0.00	9.98	0.23	0.00
0-1 km Bulk Shear	100.00	100.00	100.00	1.36	0.45
0-3 km Bulk Shear	98.19	100.00	50.34	77.10	0.00
0-6 km Bulk Shear	98.19	100.00	34.69	86.17	0.00
0-1 km SREH	99.32	100.00	65.99	13.38	21.09
0-3 km SREH	94.10	100.00	35.60	77.55	28.12
BRN Shear	54.20	100.00	2.95	92.52	0.68
Storm Relative Flow	93.20	100.00	24.72	50.34	98.41
0-1 km EHI	73.47	98.64	35.83	0.91	11.34
0-3 km EHI	53.51	94.56	16.55	2.95	30.84
Vorticity Generation Potential	16.10	29.25	16.33	3.63	12.70
Product of 0-1 km shear and CAPE	44.67	10.20	25.85	0.23	0.00
Product of 0-3 km shear and CAPE	22.00	15.19	3.17	0.91	0.68
Product of 0-6 km shear and CAPE	22.68	21.32	1.81	0.68	5.90

Forty-eight hours prior to outbreak initiation, the permutation testing (Table 5) results varied slightly from those at 24-hours lead time. Surface-based CIN, which was selected at 24-hours lead time, was not chosen at 48-hours lead time due to poor discrimination capability (*p*-values > 0.1 throughout the fields). In contrast to 24-hours lead time, all three layers of bulk shear (0-1 km, 0-3 km, and 0-6 km) showed large discrimination capability and little regional bias, so all were retained. The LCL and the two layers of SREH (0-1 km and 0-3 km) continued to exhibit large discrimination capability and modest regional biases, so these were included in the statistical classification analysis. Most of the 48-hour reduced covariate set consists of shear or vorticity parameters, which was a result consistent with 24-hours lead time. Additionally the limited number of instability variables retained demonstrated the inability of these covariates to discriminate outbreak type.

Table 5. Same as Table 4, but for 48-hours lead time.

p ≤ 0.1					
Covariate	All	East	West	Tornado East vs West	Severe East vs West
Surface Based CAPE	49.66	2.04	57.14	58.28	0.00
Surface Based CIN	68.25	39.68	44.44	23.81	8.62
LCL	87.98	92.29	61.45	87.30	27.66
LFC	6.58	41.27	0.00	69.39	21.54
0-1 km Bulk Shear	100.00	100.00	100.00	21.09	73.02
0-3 km Bulk Shear	100.00	100.00	93.88	47.17	0.00
0-6 km Bulk Shear	100.00	100.00	86.85	49.66	3.40
0-1 km SREH	100.00	100.00	89.57	36.05	12.93
0-3 km SREH	100.00	100.00	78.23	58.73	10.66
BRN Shear	97.05	100.00	44.67	58.05	6.35
Storm Relative Flow	100.00	100.00	41.04	68.03	100.00
0-1 km EHI	88.89	100.00	26.76	60.32	13.61
0-3 km EHI	63.04	100.00	19.95	59.64	39.68
Vorticity Generation Potential	21.32	91.84	27.66	80.27	0.91
Product of 0-1 km shear and CAPE	67.80	71.88	14.06	60.77	3.17
Product of 0-3 km shear and CAPE	46.03	83.45	12.47	62.59	0.68
Product of 0-6 km shear and CAPE	42.40	99.09	19.50	63.72	4.99
p ≤ 0.05					
Covariate	All	East	West	Tornado East vs West	Severe East vs West
Surface Based CAPE	40.14	0.00	46.71	45.35	0.00
Surface Based CIN	58.50	31.75	25.85	15.42	1.13
LCL	81.18	85.49	48.07	66.89	7.94
LFC	1.13	36.28	0.00	59.64	0.68
0-1 km Bulk Shear	100.00	100.00	100.00	14.51	52.61
0-3 km Bulk Shear	100.00	100.00	90.93	39.00	0.00
0-6 km Bulk Shear	100.00	100.00	76.87	40.36	0.00
0-1 km SREH	100.00	100.00	83.67	27.89	5.90
0-3 km SREH	100.00	100.00	70.29	51.93	2.04
BRN Shear	90.02	100.00	36.05	46.49	0.91
Storm Relative Flow	100.00	100.00	35.15	57.60	100.00
0-1 km EHI	77.55	100.00	20.63	52.61	0.00
0-3 km EHI	54.88	100.00	10.66	53.29	17.23
Vorticity Generation Potential	13.83	80.95	12.02	71.20	0.00
Product of 0-1 km shear and CAPE	54.88	50.57	7.03	53.06	1.36
Product of 0-3 km shear and CAPE	30.39	64.85	4.54	54.20	0.00
Product of 0-6 km shear and CAPE	28.57	87.98	7.71	53.51	0.68
p ≤ 0.01					
Covariate	All	East	West	Tornado East vs West	Severe East vs West
Surface Based CAPE	25.85	0.00	26.08	11.34	0.00
Surface Based CIN	34.24	18.37	0.91	0.68	0.00
LCL	63.72	73.02	20.18	29.25	0.00
LFC	0.00	30.61	0.00	37.19	0.00
0-1 km Bulk Shear	100.00	99.55	100.00	5.22	35.83
0-3 km Bulk Shear	100.00	100.00	82.77	24.26	0.00
0-6 km Bulk Shear	99.55	100.00	58.96	19.27	0.00
0-1 km SREH	100.00	100.00	71.88	12.02	0.00
0-3 km SREH	100.00	100.00	52.38	39.23	0.00
BRN Shear	63.49	97.05	24.04	19.95	0.00
Storm Relative Flow	99.09	100.00	25.17	36.28	95.24
0-1 km EHI	59.18	100.00	7.03	22.68	0.00
0-3 km EHI	35.60	100.00	2.95	38.78	0.00
Vorticity Generation Potential	4.54	53.06	0.91	38.55	0.00
Product of 0-1 km shear and CAPE	20.41	16.78	0.23	8.84	0.00
Product of 0-3 km shear and CAPE	6.12	16.55	0.00	22.45	0.00
Product of 0-6 km shear and CAPE	3.40	32.88	0.23	19.05	0.00

Most of the covariates used at 48-hours and 24-hours prior to the outbreak showed the highest discrimination capability at 72-hours as well (LCL, bulk shear, SREH – Table 6). However, at 72-hours, 0-1 km bulk shear showed no discrimination ability when considering all outbreaks, and was rejected from the final 72-hour set. The 0-1 EHI, which showed good discrimination ability at 24-hours lead time, was retained at 72-hours as well, as a large percentage of the domain (over 90%) was significant at $p < 0.1$. The final covariate set at 72-hours lead time included 0-3 and 0-6 km bulk shear, 0-1 and 0-3 km SREH, the LCL, and 0-1 km EHI. Primarily, these covariates consist of shear and vorticity measures, which is consistent with the previous two lead times. Clearly, the permutation testing selects covariates that correspond well with the literature (Appendix B) and reinforce the ideas presented in Rasmussen and Blanchard (1998) that CAPE cannot distinguish between tornadic and non-tornadic supercells to any statistical significance but shear parameters can distinguish with up to a 99% confidence.

Table 6. Same as Table 4, but for 72-hours lead time.

p ≤ 0.1					
Covariate	All	East	West	Tornado East vs West	Severe East vs West
Surface Based CAPE	21.77	4.76	21.32	16.55	50.11
Surface Based CIN	74.83	84.13	12.47	0.00	46.94
LCL	81.63	94.56	79.37	66.67	90.02
LFC	11.11	0.00	36.96	27.89	82.99
0-1 km Bulk Shear	0.00	100.00	0.00	44.44	0.00
0-3 km Bulk Shear	100.00	100.00	92.74	62.13	8.62
0-6 km Bulk Shear	100.00	100.00	59.86	44.44	0.00
0-1 km SREH	100.00	100.00	71.88	85.03	1.81
0-3 km SREH	100.00	100.00	60.77	80.95	0.00
BRN Shear	90.25	100.00	13.15	35.15	13.83
Storm Relative Flow	84.58	100.00	8.62	80.73	96.15
0-1 km EHI	97.05	97.05	48.98	54.20	1.13
0-3 km EHI	0.00	90.25	0.00	51.93	0.00
Vorticity Generation Potential	1.81	18.59	5.90	47.39	19.95
Product of 0-1 km shear and CAPE	54.88	1.36	36.05	43.08	64.40
Product of 0-3 km shear and CAPE	52.61	4.08	36.96	35.15	44.22
Product of 0-6 km shear and CAPE	54.42	9.52	34.01	28.80	16.78
p ≤ 0.05					
Covariate	All	East	West	Tornado East vs West	Severe East vs West
Surface Based CAPE	16.55	0.68	2.95	4.76	25.62
Surface Based CIN	53.74	65.08	5.44	0.00	35.15
LCL	72.11	61.68	71.43	50.57	82.77
LFC	6.12	0.00	20.63	17.46	60.77
0-1 km Bulk Shear	0.00	100.00	0.00	31.52	0.00
0-3 km Bulk Shear	100.00	100.00	76.87	52.61	2.72
0-6 km Bulk Shear	100.00	100.00	44.22	29.02	0.00
0-1 km SREH	100.00	100.00	61.68	78.23	0.00
0-3 km SREH	100.00	100.00	51.93	73.70	0.00
BRN Shear	72.79	96.83	3.17	14.97	0.00
Storm Relative Flow	73.70	100.00	4.08	65.76	78.00
0-1 km EHI	89.34	91.38	40.82	47.62	0.00
0-3 km EHI	0.00	83.45	0.00	44.67	0.00
Vorticity Generation Potential	0.00	9.75	1.59	39.00	2.27
Product of 0-1 km shear and CAPE	31.52	0.00	26.30	35.15	51.02
Product of 0-3 km shear and CAPE	13.61	0.68	28.57	27.89	28.57
Product of 0-6 km shear and CAPE	36.73	4.99	24.49	20.63	3.40
p ≤ 0.01					
Covariate	All	East	West	Tornado East vs West	Severe East vs West
Surface Based CAPE	6.80	0.00	0.00	0.00	0.91
Surface Based CIN	23.58	24.04	0.00	0.00	7.71
LCL	44.67	23.58	26.98	27.66	43.08
LFC	0.45	0.00	9.30	0.45	24.49
0-1 km Bulk Shear	0.00	99.32	0.00	3.17	0.00
0-3 km Bulk Shear	99.32	99.77	37.19	29.71	0.00
0-6 km Bulk Shear	97.51	99.77	20.41	7.48	0.00
0-1 km SREH	98.41	100.00	48.75	60.32	0.00
0-3 km SREH	93.42	99.77	38.78	53.97	0.00
BRN Shear	35.37	53.97	0.00	0.00	0.00
Storm Relative Flow	45.35	100.00	1.59	36.28	39.68
0-1 km EHI	65.08	71.66	18.37	31.52	0.00
0-3 km EHI	0.00	57.37	0.00	30.84	0.00
Vorticity Generation Potential	0.00	0.45	0.00	24.26	0.00
Product of 0-1 km shear and CAPE	0.45	0.00	13.38	14.29	0.68
Product of 0-3 km shear and CAPE	0.00	0.00	4.31	9.07	0.91
Product of 0-6 km shear and CAPE	0.00	0.23	0.45	2.95	0.00

4) STATISTICAL CLASSIFICATION MODELS

Once a robust set of covariates was found for the three lead times considered, objective statistical classification was performed using the reduced covariate sets. Three statistical methods were chosen: LR, LogR, and SVMs.

Statistical models using input from numerous closely-spaced gridpoints may suffer from problems with multiplicity. However, the covariate fields selected from the permutation testing reside on 21 X 21 point spatial grids. In order to reduce the dimensionality of these covariate grids to individual variables for each case, a P-mode PCA (detailed description in section 2.3.1) was conducted on the data. The PCA resulted in a reduced number (less than 7) of statistically independent variables (PC scores) for each case, which was more desirable, since the number of variables was reduced but the scores implicitly contained spatial structure of the covariates. The three statistical classification methods which use these PC scores for outbreak discrimination are summarized below.

(i) Linear Regression (LR)

The LR model was included to determine the discriminatory capability of a traditional method with a long history of meteorological applications (e.g., Marzban et al. 1999, Reap and Foster 1979, and Michaels and Gerzoff 1984), and to incorporate several covariates as predictors, simultaneously. The prediction equation for multiple LR (Wilks 1995) is given as:

$$\hat{Y} = \beta_o + \sum_{i=1}^{k} \beta_i x_i \qquad (1)$$

where β_i represents the coefficients analogous to the slope of the regression line, β_o represents the y-intercept, x_i are the covariates, and $\hat{Y}$ are the predictions. The β_i coefficients are computed by:

$$\beta_i = \frac{n\sum_{j=1}^{n} x_{ij} y_j - \sum_{j=1}^{n} x_{ij} \sum_{j=1}^{n} y_j}{n\sum_{j=1}^{n} (x_{ij})^2 - (\sum_{j=1}^{n} x_{ij})^2} \tag{2}$$

In (2), n represents the number of cases being analyzed, x_{ij} represents the covariates, and y_j represents the observation, in this case, coded with a 1 (or a 0) tag for a TO (or a NTO). The predictions obtained from (1) ranged from near 0 to near 1, as opposed to individual classes. Therefore, a threshold of 0.5 was set as the limit between classifying outbreak types. Values larger than this threshold were classified as a TO, while those less than the value were called a NTO. Other threshold values were tested, but no significant improvement in the classification was achieved.

(ii) Logistic Regression (LogR)

The LogR method is suited by its design for classification (e.g. Billet et al. 1997, Schmeits et al. 2005). Wilks (1995) defines LogR by the prediction equation:

$$\hat{Y} = \frac{1}{1 + \exp[-(\boldsymbol{\beta_o} + \sum_{\mathbf{k}=1}^{\mathbf{n}} \beta_{\mathrm{k}} \mathbf{x_k})]} \tag{3}$$

LogR will assign a probability to $\hat{Y}$, based on the ratio of the probability of a TO versus a NTO (known as a logit). The logistic regression equation is derived by considering the natural log of the logit as the dependent variable of a multiple LR:

$$\ln\left(\frac{\mathrm{P(Y=1)}}{\mathrm{P(Y=0)}}\right) = \beta_{\mathrm{o}} + \sum_{\mathrm{i}=1}^{\mathrm{k}} \beta_{\mathrm{i}} \mathbf{x}_{\mathrm{i}} \tag{4}$$

Assigning the logit from the regression equation to $\hat{Y}$ and solving for $\hat{Y}$, gives:

$$\hat{Y} = \frac{\exp[\beta_{\mathrm{o}} + \sum_{\mathrm{k}=1}^{\mathrm{n}} \beta_{\mathrm{k}} \mathbf{x}_{\mathrm{k}}]}{1 + \exp[\beta_{\mathrm{o}} + \sum_{\mathrm{k}=1}^{\mathrm{n}} \beta_{\mathrm{k}} \mathbf{x}_{\mathrm{k}}]} \tag{5}$$

Dividing through by the exponential term in the numerator yields the final form for the LogR prediction equation, given in (3). This regression type only applies to binary classification problems (such as the current study since there are two outbreak types) that allow for computation of the logit.

(iii) Support Vector Machines

In addition to two statistical classification techniques, an artificial intelligence (AI) technique known as SVM was used for outbreak type classification. This non-linear learning method fits a decision hyperplane to a linearly separable dataset (e.g. Fig. 8). From Haykin (1999), a decision hyperplane is first determined, given by:

$$\mathbf{w}^{T}\mathbf{x} + b = 0 \tag{6}$$

which can then be divided into two parts to be used for classification, namely:

$$\begin{aligned} \mathbf{w}^{T}\mathbf{x} + b &> 0 \; for \; y = 1 \\ \mathbf{w}^{T}\mathbf{x} + b &< 0 \; for \; y = 0 \end{aligned} \tag{7}$$

where **w** is the vector of weights for the decision hyperplane and **x** is the input data vector. When a separating hyperplane is applied to a set of positive and negative classifiers, the distance (margin) between the closest points to the separating hyperplane of each class (support vectors) should be maximized. In other words, the quadratic optimization problem for support vector machines is given as:

$$\begin{aligned} &\min \phi(\mathbf{x}) = \tfrac{1}{2}\mathbf{w}^T\mathbf{w} \\ &subject\ \ to \\ &y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \ \ i = 1....l \end{aligned} \tag{8}$$

In order to find the minimum, the Lagrangian for the quadratic optimization function is computed:

$$L(w,b,\Lambda) = \tfrac{1}{2}\left\|w^2\right\| - \sum_{i=1}^{l}\lambda_i[y_i(wx_i + b) - 1] \tag{9}$$

where λ_i represents Lagrange multipliers. The optimality conditions for *L* are:

$$\begin{aligned} \frac{\partial L(\mathbf{w},b,\Lambda)}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{i=1}^{l}\lambda_i y_i \mathbf{x}_i \ = 0 \\ \frac{\partial L(\mathbf{w},b,\Lambda)}{\partial b} &= -\sum_{i=1}^{l}\lambda_i y_i \ = 0 \end{aligned} \tag{10}$$

These partial derivatives are solved using a nonlinear optimization method such as steepest descent, which ensures a local or global minimum results from the differentiation. After differentiation, the optimal value for **w** is given as:

$$w^* = \sum_{i=1}^{l}\lambda_i y_i x_i \tag{11}$$

Substituting (10) and (11) into (9) gives the dual formulation (so called as it is a second formulation that solves the same optimization problem) of the quadratic optimization function:

$$\begin{aligned} &\max F(\Lambda) = \sum_{i=1}^{l}\lambda_i - \tfrac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\lambda_i\lambda_j y_i y_j \mathbf{x}_i \mathbf{x}_j \\ &subject\ \ to \\ &\sum_{i=1}^{l}\lambda_i y_i \ = 0 \\ &\lambda_i \geq 0 \end{aligned} \tag{12}$$

In this problem, data points which correspond to λ values greater than zero are called support vectors. Solving this quadratic optimization problem will yield values of λ, which in turn can be used to determine the optimal values for **w** and b, and thus give a classification algorithm.

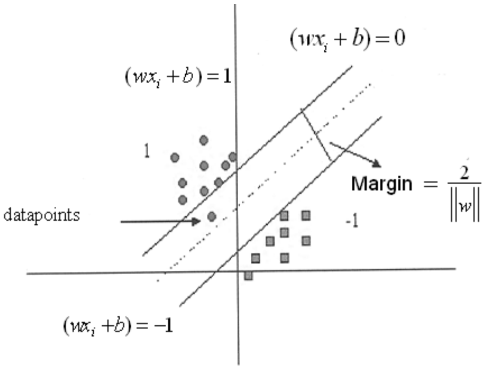


Fig. 8. Idealized SVM application for two linearly separable classes. The support vectors touch the solid boundaries of the margin. The norm of the **w** vector is minimized in the primal optimization solution of SVMs (adapted from T05).

Most binary classification datasets, including the current dataset, cannot be linearly separated initially. In these scenarios, the use of a kernel function will map the dataset into a higher dimensional space in which it is linearly separable (similar to Fig. 8, but with higher dimensionality). The kernel function does not compute the explicit coordinates of the data point in the higher dimensional (feature) space, which is computationally expensive, but instead is comprised of the product of the image $\varphi(\mathbf{x})$ of the input vectors in the feature space (i.e. $\varphi(\mathbf{x}_i)^T \varphi(\mathbf{x}_j) = k(\mathbf{x}_i,\mathbf{x}_j)$ where k is the kernel function – Cristianini and Shawe-Taylor 2000). Hence, the exact dimensionality of the feature space can be unknown and is not necessary to determine for SVMs, since the kernel function maps to this higher dimensionality directly. This method does not guarantee linear separability (since 100% accuracy is not achieved by mapping with a kernel function), but significant improvement for non-linearly separable datasets is seen when using a kernel function with SVMs. Some families of kernel functions include:

1. polynomial $$k(\mathbf{x},\mathbf{y}) = (\mathbf{x}^{\mathbf{T}}\mathbf{y}+1)^{p} \tag{13}$$

2. radial basis function $$k(\mathbf{x},\mathbf{y}) = \exp^{\left[\frac{-1}{2\sigma^{2}}\|\mathbf{x}-\mathbf{y}\|^{2}\right]} \tag{14}$$

3. tangent hyperbolic $$k(\mathbf{x},\mathbf{y}) = \tanh(\beta_{\mathbf{0}}\mathbf{x}^{\mathbf{T}}\mathbf{y}+\beta_{1}) \tag{15}$$

where *x* and *y* are the data (inputs) and output vectors, respectively. The non-linear map function $\varphi(\mathbf{x})$ can replace **x** in (12), and since the dot product of **x** is given in (12), one is replacing this dot product with the kernel matrix. Due to this inclusion of the kernel matrix, SVMs are also known as kernel methods. Multiple SVM experiments are required to determine the kernel function which provides the best classification. For the current study, the radial basis function was most successful [not shown] in classifying outbreak type based on contingency statistics.

5) CONTINGENCY STATISTICS

The three classification methods each produced a binary output, either a 0 for a NTO or a 1 for a TO. Binary output is often verified by using a contingency table. This 2x2 table (Table 7) organizes the classification results into four categories. The upper left value of the contingency table represents correctly classified 1 (TO) outputs, the upper right represents a forecast 1 value when a 0 (NTO) value is observed, the lower left represents a forecast value of 0 when 1 is observed, and the lower right represents a correctly forecast 0 value. Contingency statistics are formulated from the results in the contingency table. Several contingency statistics are used in the present study, and are described below.

Table 7. Sample contingency table. In the 2x2 contingency table, *a* represents the correctly classified 1 value, *b* represents the incorrectly classified 1 value , *c* represents the incorrectly classified 0 value, and *d* represents the correctly classified 0 value.

Forecast		Observations	
		TO (1)	NTO (0)
	TO (1)	a	b
	NTO (0)	c	d

The most basic contingency statistic measures the number of correctly classified outbreaks versus the total number of classifications. This statistic, known as the hit rate (HR), is defined in Wilks (1995) as:

$$HR = \frac{a + d}{n} \qquad (16)$$

where *n* is the total number of cases for the entire set and *a* and *d* represent the number of correctly classified TOs and NTOs (Table 7). A HR of 1 represents a perfect classification, so values closer to 1 are desirable. This statistic credits correctly classified TOs and NTOs equally. However, it provides no information on the two error types (variables *b* and *c*), which are treated differently in most meteorological applications. Hence, additional contingency statistics with information on misclassifications is needed.

A commonly applied contingency statistic which measures the likelihood that a “yes” (in this study, a TO) is correctly classified is known as the probability of detection (POD). According to Wilks (1995), the POD is given as:

$$POD = \frac{a}{a+c} \tag{17}$$

where *a* represents correctly classified TOs and *c* represents the number of TOs that are observed when an NTO was predicted. A perfect TO classification has a POD of 1, so values closer to 1 are desirable. The POD represents the fraction of TOs that were correctly predicted by the classification scheme. Forecasters often are most concerned with high POD values to ensure that no "yes" (for the present study, TOs) events are missed. However, the POD does not provide a measure of the number of incorrectly classified TOs, which also is of interest to forecasters wishing to reduce the rate of false alarms.

To account for the number of incorrectly classified TOs, the false alarm ratio (FAR) is computed from the classification results. The FAR is represented in Wilks (1995) as:

$$FAR = \frac{b}{a+b} \tag{18}$$

The FAR represents the ratio of forecast TO events (*a* + *b*) that fail to become TOs (*b*). A perfect classification will have a FAR of 0, so smaller FAR values are desirable.

Although these contingency statistics provide different properties of the classification results, a summary performance measure for each classification method is helpful. A commonly used skill statistic that provides a performance measure is the Heidke skill score (HSS). Wilks (1995) defines the HSS as:

$$HSS = \frac{2(ad-bc)}{(a+c)(c+d)+(a+b)(b+d)} \tag{19}$$

The HSS provides a measure of the likelihood that the HR for the given statistical method is obtained by random chance instead of by skill. Values nearer to 1 indicate a higher skill, hence a lower probability that the HR results are from random chance.

The HSS has been used in numerous classification studies as a measure of classification performance (Doswell et al. 1990, McGinley et al. 1991, Schaefer 1990, others). The HSS provides a skill measure using all members of the contingency table, which is desirable.

A final contingency statistic, bias, has been calculated. According to Wilks (1995), the bias is computed as:

$$B = \frac{a+b}{a+c} \tag{20}$$

The bias represents the ratio of the number of TO forecasts ($a + b$) to the number of TO observations ($a + c$). An unbiased result will have a value of B = 1. When B > 1, TOs are overforecast, whereas B < 1 indicates that NTOs are overforecast. This measure allows the user to adjust individual model parameters (such as the classification threshold in LR, section 2.5.1) to produce a bias value closer to 1. The bias reveals any artificially inflated POD or FAR values that are due to overforecasting of a particular outbreak type, as well.

6) OBJECTIVE CLASSIFICATION METHODOLOGY

Proper statistical classification methods employ a training and testing phase for their development. The training phase is implemented by taking a subset of the total input dataset and computing the model coefficients (β terms in LR and LogR, **w** in SVM) for that subset of data. The data withheld from the training phase then are input into the resulting statistical models to determine their performance through the contingency statistics (the testing phase). This training and testing methodology for statistical modeling is called cross-validation. Many cross-validation methods exist, including the "leave one out" approach, which uses all data but one point for training and tests on the point that has been “left out”, and simply dividing the data in half, using half for training and half for testing. The current study used a cross-validation method known as a “jackknife”. The jackknife is a resampling technique that samples without replacement, and is often considered a “leave one out” approach. However, the jackknife cross-validation method employed in this study used a large percentage of the data for training (85%) and withheld a smaller subset for testing (15%) as a first iteration. After the first iteration was complete, the first point of the testing set was used for training and the first point of the training set was used for testing in the second iteration. That is, for the first jackknife iteration, cases 1-84 were used for training and 85-99 are used for testing. Once results were compiled for the first iteration, a second iteration, which used cases 2 – 85 for training and 86 – 99 and case 1 for testing, was conducted. This was applied for all data, so all cases were used 15 times for testing and 84 times for training. This method provided a more robust solution for the contingency statistics, as many combinations were considered. This technique had several disadvantages though, since jackknifing resulted in an overestimation of variability in the results versus sampling with replacement (the bootstrap) which can result in a less representative result for the data distribution, and this method generated multiple (99) models. With this multiplicity of models, another objective method would seem to be required to determine the best model of the 99 produced by the jackknifing. However, since this was a purely diagnostic study (as opposed to a forecasting study addressing the development of a forecast application, which would require the best model of the 99), the contingency statistics were able to be computed on the 99 jackknife model outputs simultaneously.

To improve the jackknife contingency results, backward elimination of covariates was conducted on the input datasets. This technique yielded the opportunity to improve the contingency statistic results by removing covariates which were worsening the results and simplified the datasets being input into the statistical models. Often, results were improved by removing further covariates from the sets obtained from the permutation testing.

To facilitate finding the optimal combinations of covariates which accomplish superior classification ability, bootstrap confidence intervals were computed on the contingency statistics. The bootstrap sample, according to Efron and Tibshirani (1993), is a sample of *n* size, where *n* represents the length of the data vector being considered, that is randomly drawn from the initial dataset. Numerous iterations (e.g., 1000 in the present study) of this bootstrap sampling show the uncertainty of the statistic being estimated. Knowledge of this uncertainty allows for decisions to be made about the statistic that are not possible without bootstrapping.

The mean contingency statistics from the 99 jackknife iterations were bootstrapped, providing 1000 sample mean values of each contingency statistic. These sample means are presented using boxplots, which show the median (central line in the boxplot), the first and third quartiles (bottom and top of the box, respectively), and 1.5 * the interquartile range (IQR – the range between the first and third quartiles, shown by the whiskers).

While the aforementioned statistical analyses described overall performance of the classification schemes, these methods do not evaluate individual cases. In order to assess the statistical classification performance on individual cases, the number of correct and incorrect classifications of each case was retained. This additional step identified specific cases that were classified poorly, which then were studied to determine possible causes for statistical model failure on these events. Chapter 3 summarizes the results from the objective classification methods discussed previously.

c. Storm Typing Methodology

A second statistical analysis, synoptic storm typing, was undertaken as well, which provided physical fields associated with each outbreak class. The storm types were created through a statistical compositing method, and were determined from 72 hours prior to the outbreak to 6 hours prior at 6 hour intervals. Several methods were considered, including mean fields (Mercer and Richman 2007), canonical correlation analysis (CCA - Barnston and Ropelewski, 1992), and PCA (Jones et al. 2004). Mean fields are not robust enough for the current compositing since the sample size of each outbreak type was small enough for outliers to affect the results. Additionally, when datasets exhibit high variability between events, the mean likely will not capture the true composite storm type. As a simple way of seeing this, if the sample includes an equal number of cases with northwesterly and southwesterly airflow, the mean would be pure westerly, which would represent none of the actual cases. CCA is not appropriate for this study either, since it requires pairs of input data vectors from unique datasets to be transformed to single fields and assumes that predictors (the input data vectors) are used to find predictands (the output fields). PCA does not presume the data are predictors or predictands, but instead assumes the data are interrelated and can be projected onto a new set of orthogonal basis vectors. As a result, PCA, which provides individual fields of

TOs and NTOs and accounts for data variability, is selected as the compositing method for this study.

PCA is a technique that projects a large dataset onto a set of independent basis vectors. The basic model equation for PCA is given as:

$$\mathbf{Z} = \mathbf{F}\mathbf{A}^{T} \tag{21}$$

where **Z** represents a standardized (mean removed from the original data) input data matrix, **F** represents a matrix of principal component (PC) scores (defined below), and **A** represents PC loadings, which are the independent basis vectors. The PC score matrix **F** represents the relationship between the loading matrix **A** and the original standardized data. Larger PC scores for a particular input data point indicate a stronger relationship between the loading matrix and the standardized data (i.e. a large score on PC1 for a given case means that the high magnitude absolute loadings on PC1 are important for that event). The PC scores also contain spatial structure of the data, as they can be multiplied by the loading matrix to recreate the original standardized data Z. As mentioned previously, PC scores were used as input into the objective classification schemes.

In order to obtain the PC loading and score matrices, several calculations are required. First, a correlation or covariance matrix must be computed from the original standardized data. An eigenanalysis is performed to diagonalize the correlation or covariance matrix, and the resulting eigenvalue and eigenvector matrices are used for computation of the PC loading matrix **A**. Finally, a least squares approach to inverting the PC loading matrix is used in combination with (21) to compute the PC score matrix. A detailed description of these methods follows.

1) Correlation matrix calculation

A series of matrix calculations is required to obtain the PC loading matrix **A** and the PC score matrix **F**. First, a correlation matrix on the standardized input matrix **Z** is computed by:

$$\mathbf{R} = \frac{\mathbf{Z}^T\mathbf{Z}}{(n-1)} \tag{22}$$

The correlation matrix, which is computed on TOs and NTOs separately, represents the correlations between the individual TO and NTO cases.

The distance between the gridpoints in **Z** can affect the calculation of **R**, since gridpoints which are geographically closer likely will be more highly correlated, possibly leading to more highly correlated cases. Since the NCEP/NCAR reanalysis data reside on a latitude-longitude grid (Fig. 9a), the distance between gridpoints decreases with increasing latitude (longitude lines converge with increasing latitude). This longitudinal convergence artificially inflates correlation calculations in northern latitudes.

Several methods exist to remove biases from converging longitude lines, and two are tested herein. The first technique, proposed by Araneo and Compagnucci (2004), uses a latitudinal density correction to obtain an equally spaced grid, such as the one seen in Fig. 9b. The latitudinal density λ is calculated using:

$$\lambda(\varphi_0) = \frac{n(\varphi_0)}{L(\varphi_0)} \tag{23}$$

where n is the number of gridpoints on a reference latitude φ_0 and L represents the approximate length of a longitude circle calculated by:

$$L(\varphi_0) = 2\pi R\cos(\varphi_0) \tag{24}$$

R is the radius of the Earth at the equator. Once λ is determined, the number of gridpoints *N* on the reference latitude φ_0 is used to determine the number of gridpoints $n(\varphi)$ for a given latitude by (25).

$$n(\varphi) = \text{int}[\frac{N\cos\varphi}{\cos\varphi_0}] \tag{25}$$

Once $n(\varphi)$ is computed, the grid spacing is given by:

$$\Delta(\varphi) = \frac{360}{n(\varphi)} \tag{26}$$

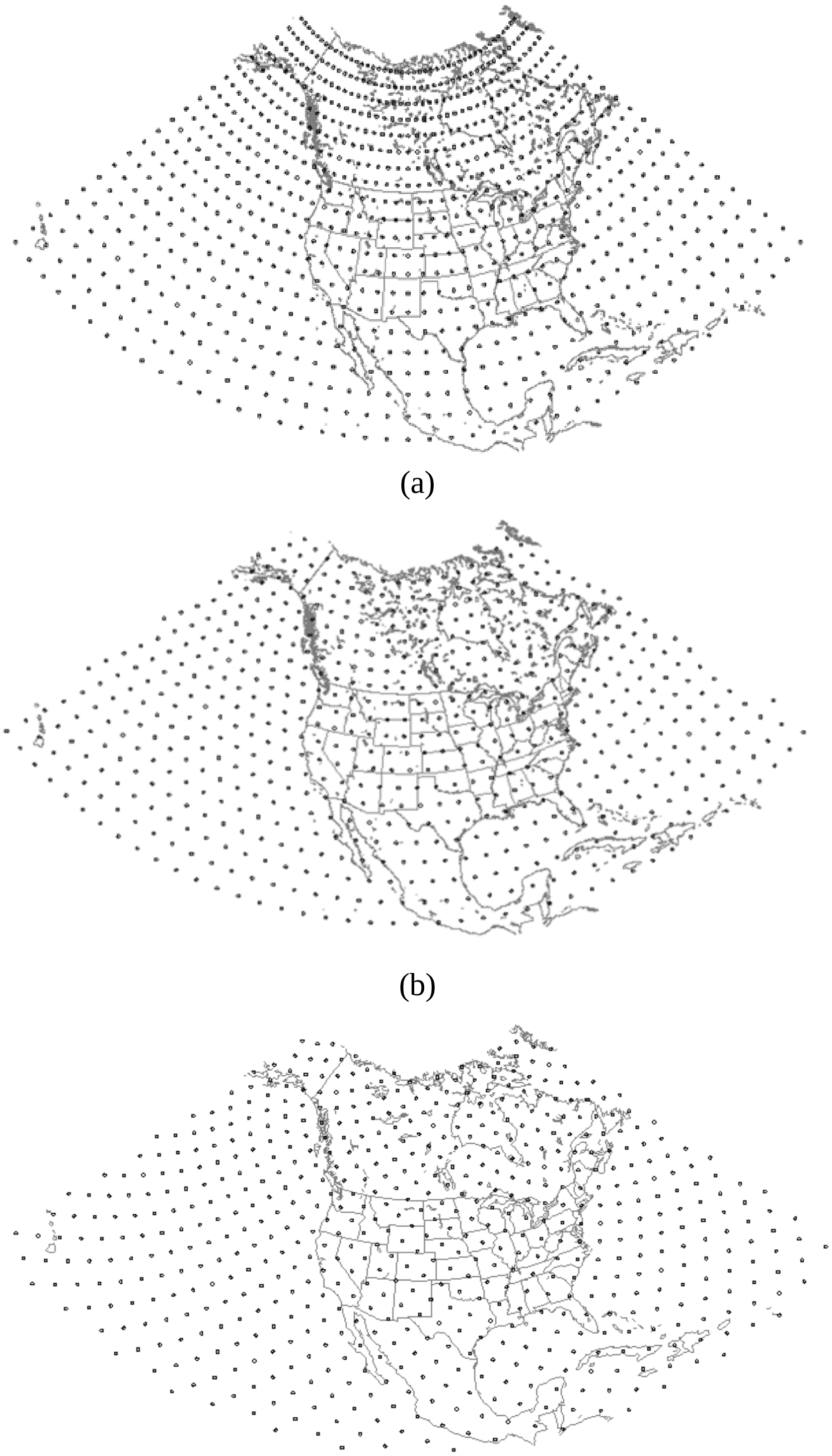
For the current study, the reference latitude φ_0 selected was the equator, since this led to equal grid spacing in the latitudes and longitudes.

In addition to the latitudinal density grid, a Fibonacci grid (Swinbank and Purser 2006, Figure 9c) was tested remove biases from converging longitude lines. The Fibonacci grid uses the Golden ratio, $\Phi = (1 + \sqrt{5})/2$, a pre-determined number of gridpoints *N*, and a latitude equation (27) and longitude equation (28).

$$\sin\theta_i = \frac{2i}{2N+1} \tag{27}$$

$$\lambda_i = 2\pi i\Phi^{-1} \tag{28}$$

In (27) and (28), *i* represents the *i*th gridpoint of the *N* chosen. The Fibonacci grid places gridpoints which are slightly offset from the poles and does not require an arbitrary reference latitude, although the user must select a predetermined number of gridpoints *N*. For the current study, *N* was selected to be 3000, as this gave the closest grid spacing to that seen from the latitudinal density grid. Since the grid spacing between these two methods is similar, a comparison to determine the superior method is possible.



(a)

(b)

(c)

Fig. 9. Plots of the reanalysis grid (a), the latitudinal grid (b), and the Fibonacci grid (c). Convergence of gridpoints in (a) requires the need for additional grid types, such as (b) and (c). The subtle convergence of gridpoints with increasing latitude in panels (b) and (c) is caused of the map projection (polar stereographic).

Before each grid type could be applied to the PCA, an interpolation technique was required to convert the NCEP/NCAR reanalysis grid to the latitudinal density grid or the Fibonacci grid. A one-pass Barnes analysis (Barnes 1964) was used for this interpolation. The scale length selected for the Barnes analysis was:

$$\kappa = \frac{\kappa_o}{2\Delta n} \tag{29}$$

where κ_0 is the reference scale length (5.052) and Δn represents the average grid spacing (taken to be 250 km for the reanalysis grid). The Great Circle Distance formula was used to compute distances between the longitudes and latitudes, which were needed for the Barnes analysis. The Great Circle Distance formula is given as:

$$d(\lambda_1, \theta_1; \lambda_2, \theta_2) = 2R\sin^{-1}(\sqrt{\sin\left(\frac{\lambda_2 - \lambda_1}{2}\right)^2 + \cos(\lambda_1)\cos(\lambda_2)\sin\left(\frac{\theta_2 - \theta_1}{2}\right)^2} \tag{30}$$

where λ_i are the longitudes of the two points, θ_i are the latitudes of the two points, and R is the Earth radius in meters. An error analysis of the one-pass Barnes technique was performed to assess any interpolation errors, interpolating the reanalysis 500 hPa height field to the Fibonacci grid and back to the reanalysis. Root mean square errors (RMSE) of the heights were computed, and RMSE values are less than 20 m (less than 1% of the mean) were noted, validating the interpolation technique.

Since two methods were tested, an analysis of each method’s performance was required to select the best one. Both methods were compared by plotting fields of 24-hours 500-hPa height anomalies after a PCA was performed with each grid type (Fig. 10). A single case, 26 April 1994, corresponded to the highest PC loading from both methods, so this field was plotted as well to provide a comparison to the anomaly fields. An anomaly ridge over western Canada using the latitudinal density PCA (Fig. 10a) did not correspond with the 500 mb height field on 26 April 1994 (Fig. 10c). This anomaly ridge in Fig. 10a displaced the anomaly trough over eastern Canada present in both grid type PCA fields (Figs. 10a and 10b). Since the latitudinal density grid PCA presented anomaly features in the northern latitudes which were not consistent with the case set, the Fibonacci grid was selected for the calculation of the correlation matrix.

Once the proper grid type was selected, two methods of correlation matrix computation were possible. An O-mode analysis involves computation of the correlation matrix along the input (cases) dimension of the correlation matrix and requires the other dimension to be parameters for the cases (as is the case in this study). When the correlation matrix is computed along the parameter dimension, the method is called a P-mode analysis. The O-mode analysis was chosen for the present study since knowledge of the correlation between cases was needed for the synoptic storm types. Additionally, the OU Supercomputing Center for Education and Research (OSCER) did not allow for the solution of a large eigenproblem (53000 X 53000 correlation matrix) which was needed for a P-mode PCA in this study.

One computational complication resulting from the O-mode analysis was the combination of numerous variables for each case which have extremely different magnitudes (i.e. 100 mb height magnitudes are on the order of 10000, while relative humidity magnitudes range from 0 to 100). Each raw variable at each of the 17 vertical levels was standardized individually to account for disparate means. This standardization subtracted the mean and divided by the standard deviation, so that the variables for each case had a mean of zero and a standard deviation of one. This standardization would not have been necessary with a P-mode analysis, since vectors of the same variable for different cases were used in the calculation of **R** in (22).

2) EIGENANALYSIS

Once the **R** matrix, which is 50 X 50 (or 49 X 49 for NTOs), was computed, an eigenanalysis was performed on **R** to obtain an eigenvector matrix **V** and an eigenvalue diagonal matrix **D**. The eigenvector and eigenvalue matrices are calculated from:

$$\mathbf{R} = \mathbf{V}\ \mathbf{D}\ \mathbf{V}^T \tag{31}$$

Typically, an eigensolver such as S-Plus (Insightful 2007 – used in the present study) is used to obtain **V** and **D**. These eigenvectors define a new coordinate system which has the same number of variables as the smaller of the number of columns or number of rows minus one in **Z**. Geometrically, the first eigenvector will point in the direction of the largest variability in the dataset, and will be associated with the largest eigenvalue. Subsequent eigenvectors will describe monotonically lower variability and are associated with monotonically smaller eigenvalues (e.g., $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \lambda_n$).

Since real datasets include both signal and noise, if too many eigenvectors are retained, noise in the data dominates the signal in the latter eigenmodes. However, if too few eigenvectors are retained, some of the physical signal will be discarded. The scree test is one method to determine the number of eigenvectors to retain so that the important signal information is kept without excess noise. For a scree test, the eigenvalues are plotted (Y-axis) sequentially for each root number (X-axis), and when the eigenvalues subjectively level off (a scree), eigenvectors prior to this point should be retained. To obtain more sensitivity, it is customary to plot a subset of the largest eigenvalues. Figure 9 shows a sample scree test plot indicating two possible truncation locations (either at 2 or 5 eigenvectors). This subjectivity can result in removal of important signal information or inclusion of noise data, so an objective truncation method would be helpful.

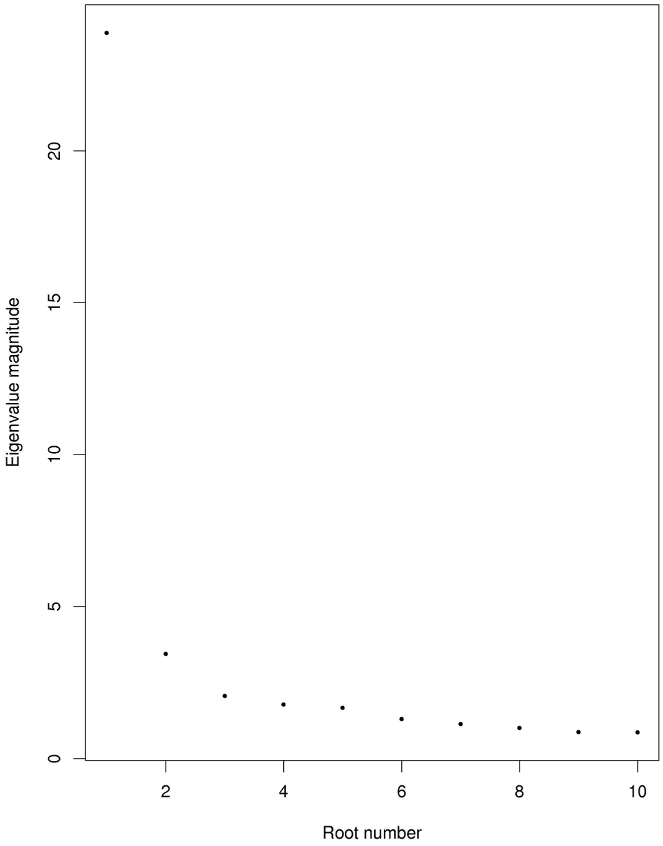


Fig. 9. Scree test plot of TO data from 24-hour lead times prior to the outbreak. The y-axis is the magnitude of the eigenvalue, ordered in descending variance explained, while the x-axis represents the eigenvalue number. Note that only the variance associated with the first 10 eigenvectors is shown

A purely objective test introduced in Richman (1986) is based on the so-called congruence coefficient η. This method requires the computation of the loading matrix A from an initial guess set of eigenvectors. The loading matrix is computed by:

$$\mathbf{A} = \mathbf{V}\,\mathbf{D}^{1/2} \qquad (32)$$

In the current study, the congruence coefficient was computed on the first two eigenvectors initially. The congruence coefficient is given by:

$$\eta = \frac{\sum \mathbf{XY}}{(\sum \mathbf{X}^2 \sum \mathbf{Y}^2)^{1/2}} \tag{33}$$

where **X** represents the vector of the original correlation matrix associated with the largest absolute loading magnitude in the loading vector **Y**. As an example from the current study, the computation of η for the first principal component (PC1, first column of the loading matrix A) involves ranking the loading vector associated with PC1 based on absolute magnitude. The largest magnitude in PC1 corresponds to a vector in the correlation matrix (i.e., if the 12th loading were largest, the 12th column of the correlation matrix is **X**). The magnitude of η is then computed from **X** and **Y**. If the value of the congruence coefficient for the first PC is larger than 0.81 (deemed by Richman 1986 as a reasonable match), this PC is retained. If both PCs have η values larger than 0.81, three PCs are tested. This process continues until a value of 0.81 or less for η is discovered. The congruence coefficient approach is superior to the scree test, as it provides a single, well-defined best answer, and its computation is based on the embedded signal in the correlation matrix, which guarantees the physical structure is part of the decision process.

The congruence coefficient test in the current study yielded two main PCs for NTOs and TOs for each lead time. Once the optimal number of PCs to retain was determined for the different outbreak types and different lead times, the PC loading and PC score matrices were computed. The PC loading matrix **A** was computed from the truncated eigenvector matrix **V** through (32). Once **A** was determined, the base model equation was solved for **F**, the PC score matrix. However, since **A** was not symmetric, the inverse of **A** required a least-squares solution, so that the PC score matrix results from:

$$\mathbf{F} = \mathbf{Z}(\mathbf{A}) * (\mathbf{A}^T \mathbf{A})^{-1} \tag{34}$$

Since an O-mode analysis was used in the computation of **R**, the PC score matrix represents the relationship between the individual gridpoints and the PC loading matrix **A**. The PC score matrix had a dimensionality of 53000 X 2, where the 2 represented the 2 PCs which were retained and the 53000 indicated the number of gridpoints. The PC scores from this O-mode analysis represented standardized anomalies of the gridpoint values. Since vectors (columns) of **F** represented gridpoint fields, weighted sums of the columns of **F** were used to create the synoptic storm types.

3) CREATION OF STORM TYPES

The PC score matrix **F** provides anomaly patterns of two PCs of each outbreak type. However, these anomaly patterns do not represent the synoptic storm types, which are computed using weighted sums of the PC scores. In order to determine the number of TO and NTO types, a cluster analysis (Wilks 1995) was employed. The cluster analysis conducted herein considered the Euclidian distance of a set of PC loadings for the 50 cases and used an average linkage method (Wilks 1995). These Euclidian distances are graphically grouped with others whose distance is within a certain threshold via a dendrogram. A typical example of a dendrogram (Fig. 10) shows groupings of storms at any given Euclidean distance (Y-axis), and these groups can be combined to determine storm types. The groups share similar physical properties and/or values of the raw variables that are used in the PCA.

To provide an example of the cluster analysis, the entire 100 case set of TOs and NTOs was analyzed in an O-mode PCA. PC loadings from the PCA were input into the cluster analysis, and the resulting dendrogram (Fig. 10) showed two main groups which largely corresponded with the TOs and NTOs. The point at which two cases merge on the dendrogram represents their Euclidian distance, and the largest Euclidian distances on Fig. 10 are near 0.2. At this level on the dendrogram, two clusters are visually apparent. Areas below this do not reveal significantly distance groupings of cases, so two storm types are noted.   However, there was some overlap between the two outbreak types in the cluster analysis (Fig. 11), which underscores the need for statistical methods to aid in classification of TOs and NTOs.

Once the cluster analysis provided storm types (two for each outbreak type), the mean loading of the events within a particular cluster was computed and squared, since the mean provided an explained-variance measure which was used to weight the PC scores. The weighted PC scores were summed to obtain anomaly patterns which represented the storm types. Chapter 4 provides the results from performing this methodology from 72 hours prior to the outbreak to 6 hours prior to the outbreak.

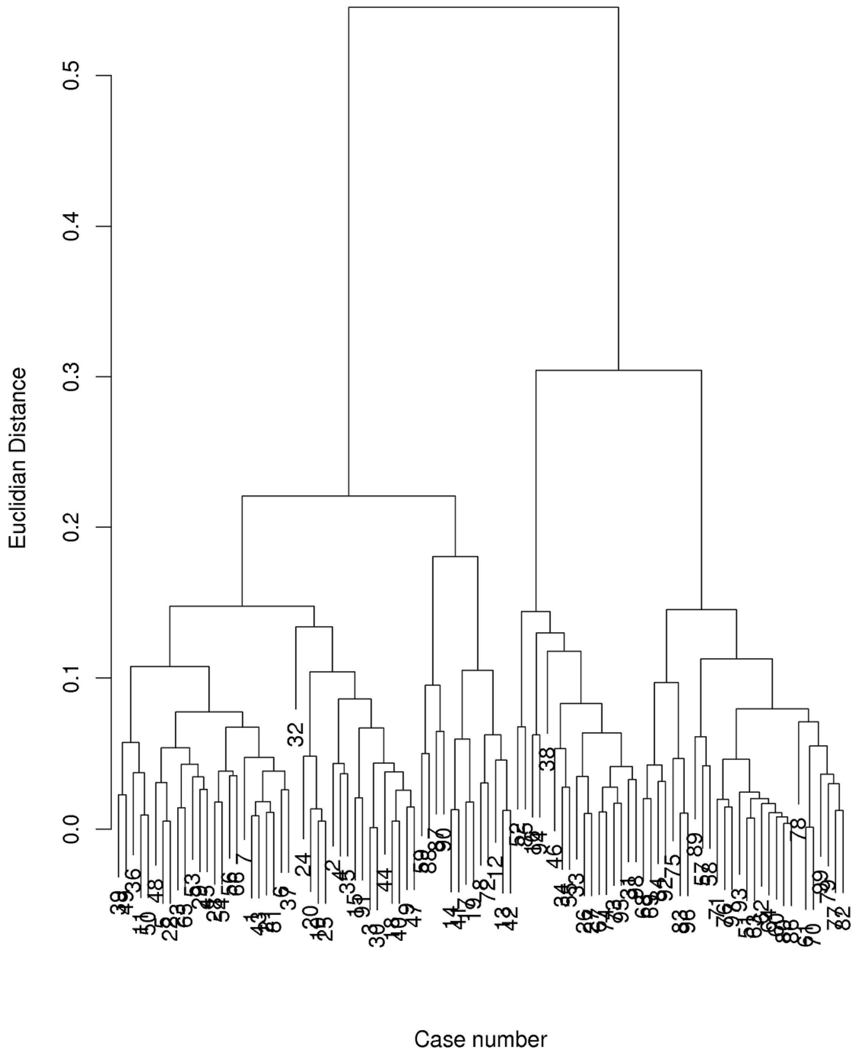


Fig. 10. Dendrogram from a cluster analysis of the PC loadings from a PCA involving all TO and NTO cases. Cases 1-50 along the bottom represent TOs, and 51-99 represent NTOs. Two main groups are apparent, although these do not correspond directly with the TO and NTO case numbers.

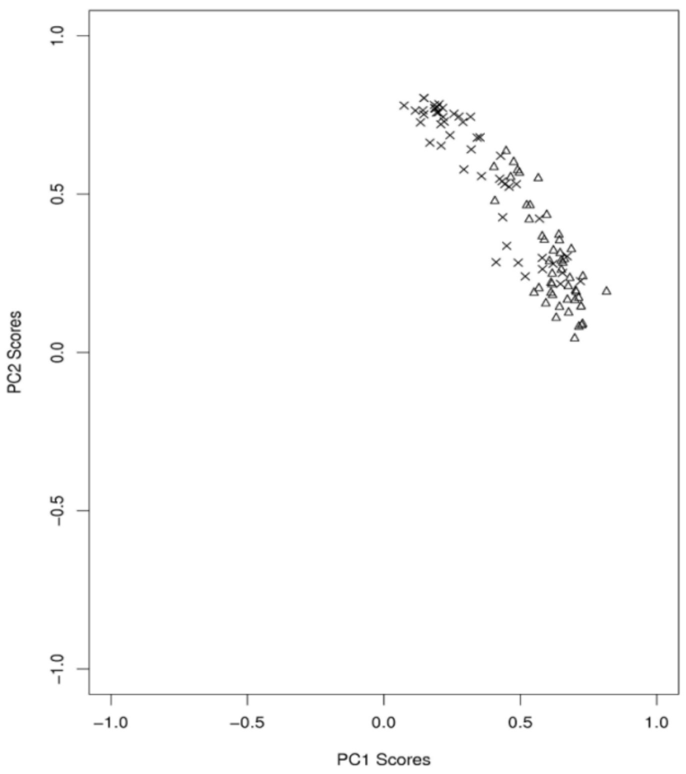


Fig. 11. Scatterplot of PC1 loadings and PC2 loadings from the PCA which considered all TOs and NTOs. Triangles represent TOs and crosses represent NTOs. This diagram shows some overlap, indicating the difficulty in separating these classes without some other statistical methods.

3. OBJECTIVE DISCRIMINATION RESULTS

The objective statistical classification of outbreaks was accomplished through three statistical methods (defined in Chapter 2) for 24-, 48-, and 72- hours prior to the outbreak. Tables of the contingency statistics are presented for each model type and each lead time, and boxplots of the contingency statistics' distributions are shown. The performance of the statistical methods with each case is assessed as well.

a. 24 Hour Results

At 24-hours prior to the outbreak, the contingency results generally were nearest to the ideal values. From the permutation testing, the optimal covariates included 0-1 km EHI, surface based CIN, 0-1 km bulk shear, the product of surface based CAPE and 0-1 km bulk shear, LCL, 0-1 km SREH, and 0-3 km SREH. Initially, the statistical models were trained and tested using these 7 covariates. Subsequent model tests were conducted, removing covariates individually to attempt to achieve better results than the initial analysis. If all of the contingency statistics improved from removing a covariate, the other covariates were individually removed in a second analysis. Single covariates were tested as well, to provide some insight as to each covariate's classification ability. In total, 26 covariate combinations were tested. The jackknife cross-validation results from these covariate combinations were bootstrapped, and boxplots of their distributions were created in order to determine the best covariate set. These boxplots show the median, the first and third quartiles, and the last data point prior to the 1.5 * IQR location. Models with a higher median and a smaller IQR are better (since they have less classification variability), and the best of these are deemed the best covariate combination for the particular statistical technique.

1) SVM CONTINGENCY AND CONFIDENCE LIMIT RESULTS

SVMs provided the best contingency results when both shear variables were culled (Table 8). While removing individual covariates during the initial analysis (models 2-8, Table 8a), it was noted that removal of the product of CAPE and 0-1 km bulk shear resulted in the highest POD (0.864). As a result, a second analysis (Table 8b) was conducted which removed this product and the other 6 covariates individually. This additional testing was conducted to determine if the results of the initial analysis (Table 8a) could be improved. The best SVM results (POD of 0.894, FAR of 0.161) were obtained by removing surface based CIN and the product of CAPE and 0-1 km bulk shear. The HSS values for this combination of covariates were high as well (0.725). Thus, for SVMs, the best parameter combination for classification was combination 25. The covariates rejected included a measure of stability or instability, which was not thought to vary significantly between outbreak type (see Appendix B).

The boxplots of the 26 covariate combinations for HR (Fig. 12) and POD (Fig. 13) reveal interesting features, and the most noticeable feature is the large IQR associated with models 9-19. These models only considered one or two covariates (Table 8), and such small covariate sets resulted in large variability in the contingency statistics. This large variability appeared consistently with all three statistical methods at all three lead times tested. Five models showed a large median value of POD and HR (1, 9, 20, 21, and 25). The FAR (Fig. 14) and HSS (Fig. 15) results corresponded well with the POD and HR results, as results from models 9 and 25 maintained the lowest FAR and highest HSS medians of all 26 combinations. However, model 9 consistently showed large IQR in the boxplots, which led to rejecting it as the best covariate set. Model 25 contained few outliers as well. Model 25 rejected the measure of CAPE and bulk shear and surface based CIN, which was consistent with the results in Table 8.

Table 8. 24 hour contingency table results for SVMs. Table (a) represents covariate combinations as stated and Table (b) leaves off the poorest covariate from Table (a) (in this case, the product of CAPE and bulk shear) and combines other parameters.

Model #	Variable(s)	HR	POD	FAR	HSS	BIAS
1	All	0.865	0.859	0.132	0.731	0.989
2	No LCL	0.836	0.838	0.168	0.673	1.007
3	No 0-1 km capeshear	0.824	0.864	0.202	0.649	1.083
4	No 0-1 km bulk shear	0.700	0.713	0.310	0.399	1.033
5	No surface CIN	0.758	0.762	0.248	0.515	1.014
6	No SREH (0-1 km)	0.799	0.800	0.205	0.597	1.007
7	No SREH (0-3 km)	0.777	0.754	0.213	0.554	0.958
8	No EHI (0-1km)	0.811	0.816	0.195	0.623	1.014
9	No Shear variables	0.853	0.880	0.167	0.707	1.057
10	No SREH variables	0.729	0.688	0.255	0.458	0.924
11	Only LCL	0.776	0.763	0.221	0.551	0.980
12	Only surface CIN	0.565	0.351	0.396	0.126	0.581
13	Only 0-1 km bulkshear	0.752	0.815	0.279	0.505	1.131
14	Only 0-1 km capeshear	0.737	0.799	0.293	0.474	1.129
15	Only 0-1 km SREH	0.807	0.833	0.211	0.614	1.056
16	Only 0-3 km SREH	0.793	0.803	0.216	0.585	1.024
17	Only 0-1 km EHI	0.778	0.887	0.274	0.558	1.222
18	Only SREH	0.782	0.801	0.231	0.565	1.042
19	Only shear	0.697	0.750	0.326	0.395	1.112

(a)

Model #	Variable(s)	HR	POD	FAR	HSS	BIAS
20	No 0-1 km EHI	0.827	0.831	0.179	0.654	1.012
21	No 0-1 km bulk shear	0.853	0.880	0.167	0.707	1.057
22	No 0-1 km SREH	0.826	0.839	0.186	0.651	1.031
23	No 0-3 km SREH	0.826	0.845	0.188	0.653	1.041
24	No LCL	0.778	0.790	0.232	0.556	1.030
25	No surface based CIN	0.863	0.894	0.161	0.725	1.065
26	No SREH variables	0.775	0.739	0.207	0.550	0.932

(b)

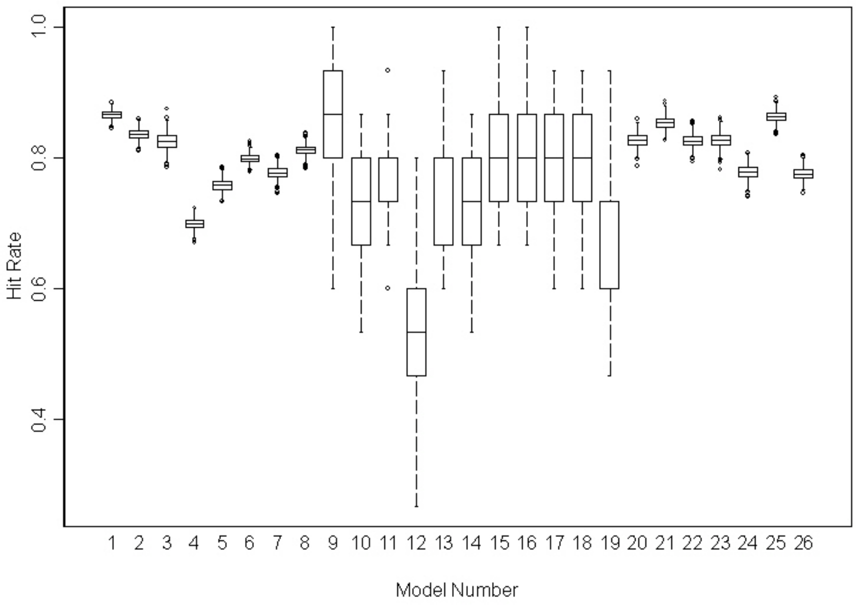


Fig. 12. Boxplots for SVM HR results. Model numbers correspond with the row number in Table 8.

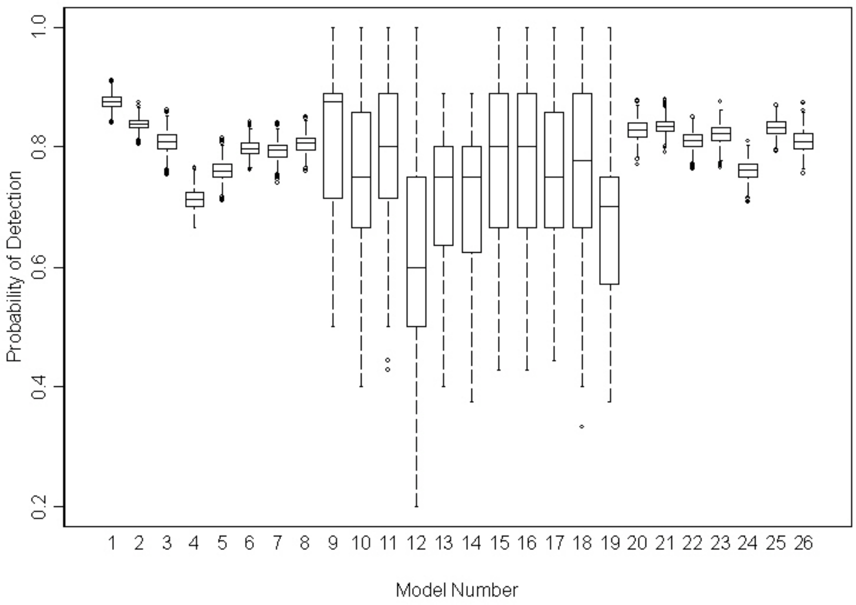


Fig. 13. Same as Fig. 12, but for POD.

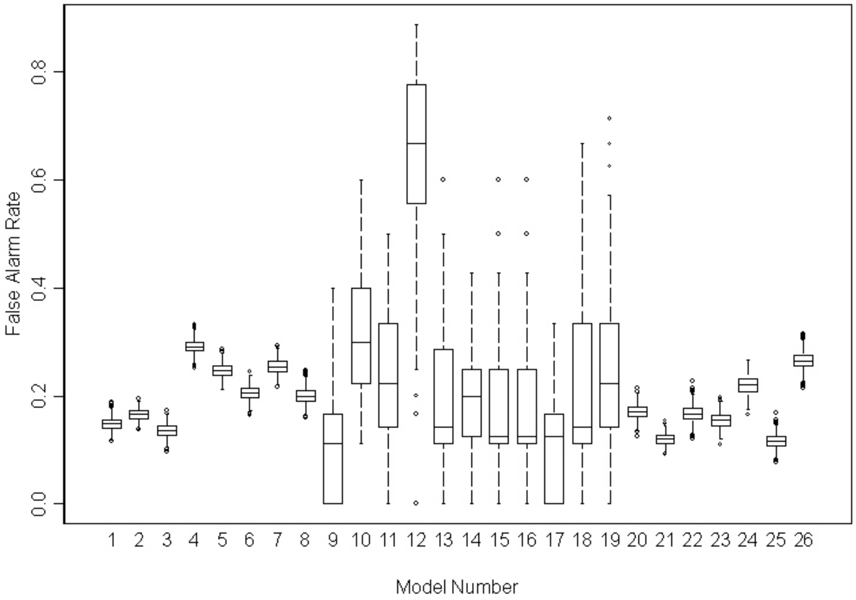


Fig. 14. Same as Fig. 12, but for FAR.

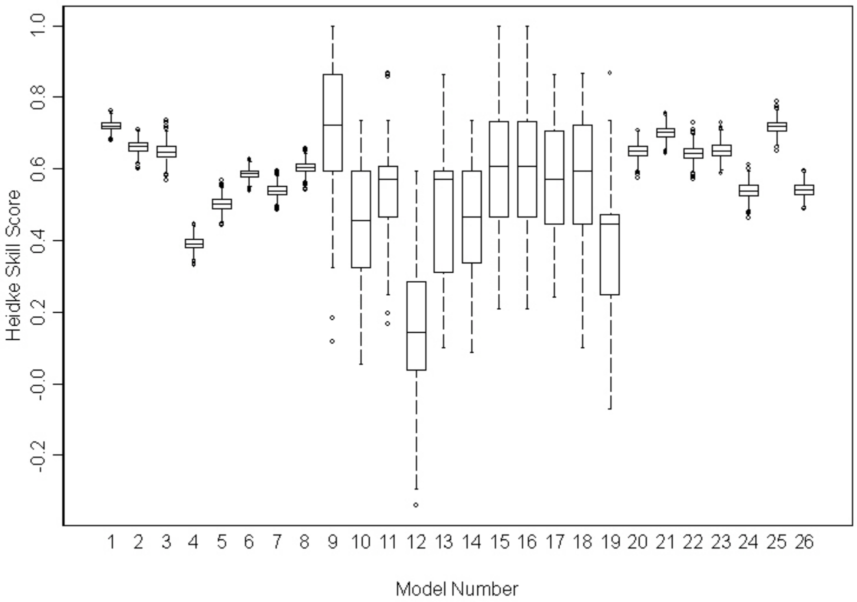


Fig. 15. Same as Fig. 12, but for HSS.

2) LOGR CONTINGENCY AND CONFIDENCE LIMIT RESULTS

Results in Table 9 suggest that LogR discriminates outbreak type successfully. . Rejection of the product of CAPE and bulk shear (the same as with SVMs), provided the highest HR, POD, and HSS results and the lowest FAR results of the initial analysis. To determine if more improvement was possible, the CAPE-bulk shear product was rejected in a second analysis (Table 9b), as was the case with SVMs. Overall, modest improvement was noted when considering combination 25.

The HR and POD boxplots (Figs. 15 and 16) portrayed models 9, 20, 21, and 25 as the best (highest median value). The FAR results (Fig. 17) and HSS results (Fig. 18) were consistent with the HR and POD results as well. All of these combinations suggested rejection of the product of 0-1 km bulk shear and CAPE, which supported the previous conclusion that instability was a poor discriminator of outbreak type. However, SVMs provided a more compact set of possible covariate combinations, which implied that SVMs may be the best classification method at 24-hours prior to outbreak initiation.

Table 9. Same as Table 8, but for LogR.

Model #	Variable(s)	HR	POD	FAR	HSS	BIAS
1	All	0.782	0.778	0.220	0.564	0.997
2	No LCL	0.751	0.756	0.256	0.502	1.016
3	No 0-1 km capeshear	0.836	0.853	0.178	0.671	1.038
4	No 0-1 km bulk shear	0.754	0.731	0.238	0.507	0.959
5	No surface CIN	0.745	0.716	0.243	0.491	0.946
6	No SREH (0-1 km)	0.755	0.750	0.246	0.510	0.995
7	No SREH (0-3 km)	0.748	0.716	0.239	0.496	0.940
8	No EHI (0-1km)	0.767	0.751	0.228	0.534	0.973
9	No Shear	0.831	0.848	0.182	0.662	1.037
10	No SREH	0.705	0.664	0.281	0.410	0.924
11	Only LCL	0.765	0.788	0.250	0.530	1.050
12	Only surface CIN	0.572	0.514	0.424	0.144	0.893
13	Only 0-1 km bulkshear	0.739	0.737	0.265	0.477	1.003
14	Only 0-1 km capeshear	0.639	0.705	0.381	0.279	1.139
15	Only 0-1 km SREH	0.800	0.816	0.213	0.600	1.037
16	Only 0-3 km SREH	0.793	0.830	0.231	0.585	1.079
17	Only 0-1 km EHI	0.698	0.752	0.326	0.396	1.116
18	Only SREH	0.782	0.812	0.237	0.565	1.064
19	Only shear	0.643	0.706	0.377	0.287	1.133

(a)

Model #	Variable(s)	HR	POD	FAR	HSS	BIAS
20	No 0-1 km EHI	0.837	0.864	0.183	0.674	1.057
21	No 0-1 km bulk shear	0.831	0.848	0.182	0.662	1.037
22	No 0-1 km SREH	0.808	0.816	0.200	0.616	1.020
23	No 0-3 km SREH	0.828	0.818	0.168	0.656	0.982
24	No LCL	0.824	0.838	0.188	0.647	1.033
25	No surface based CIN	0.842	0.856	0.169	0.685	1.030
26	No sreh (all)	0.751	0.735	0.245	0.501	0.973

(b)

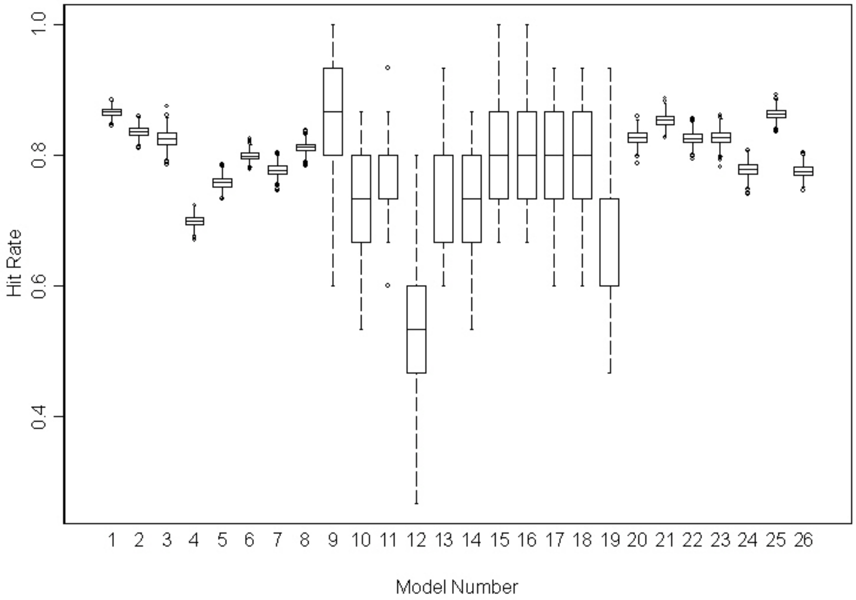


Fig. 16. Same as Fig. 12, but for LogR.

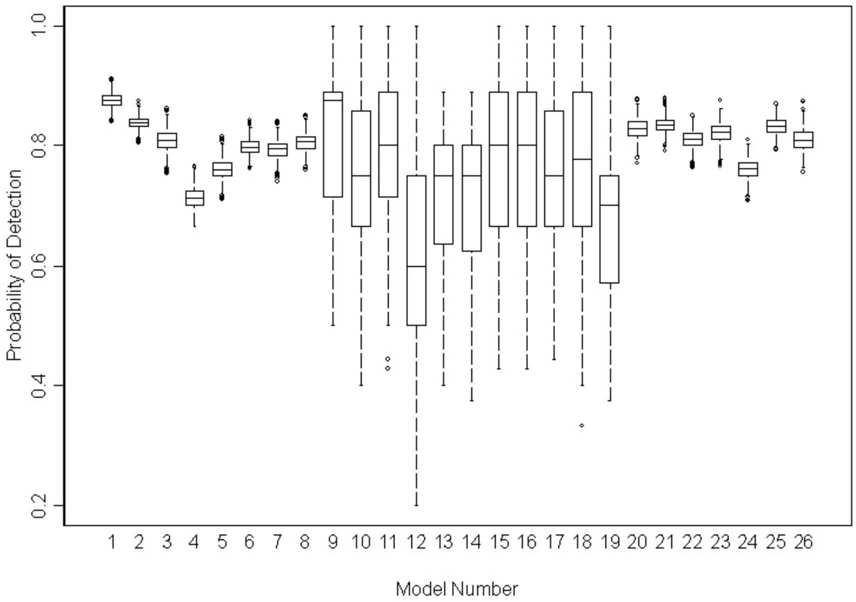


Fig. 17. Same as Fig. 16, but for POD.

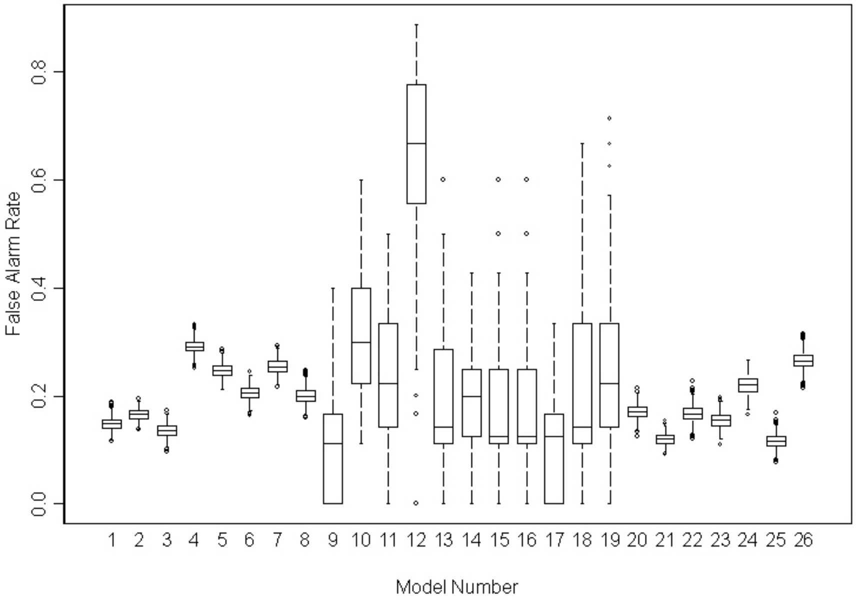


Fig. 18. Same as Fig. 16, but for FAR.

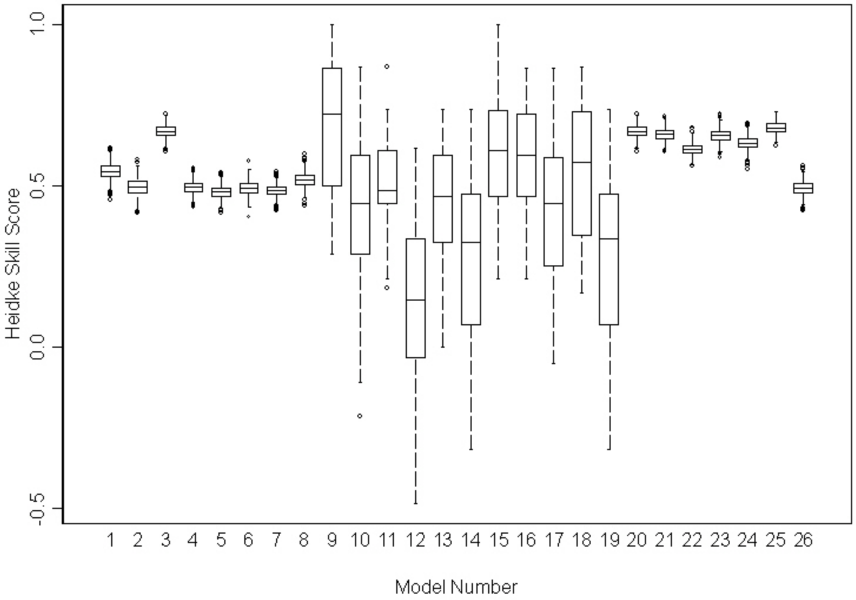


Fig. 19. Same as Fig. 16, but for HSS.

3) LR CONTINGENCY AND CONFIDENCE LIMIT RESULTS

The contingency statistics for LR from the initial analysis (Table 10a) showed rejection of the product of CAPE and shear at 0-1 km provided the best contingency statistic values. When rejecting this product and the other covariates individually (Table 10b), the best contingency results were seen when surface based CIN was rejected. This covariate set was identical to the two obtained by LogR and SVM.

The HR and POD distributions (Figs. 20 and 21) revealed five different models which provided the highest medians (models 9, 20, 21, 23, and 25). This set of combinations, along with numerous other combinations, produced low FAR results (Fig. 22). The HSS (Fig. 23) results showed these models as having the highest medians and smallest IQRs. In essence, it was not possible to determine which of these combinations was best. Since LR is a purely linear method, small adjustments to the threshold for classification (0.5 in the present study) can increase the prediction of TOs or NTOs. This can introduce artificial bias towards an outbreak type into the statistical model, which may be a cause of the low FAR results observed in the five models above. Overall, the simple linear method was successful at classifying outbreak type 24-hours prior to the event.

Table 10. Same as Table 8, but for LR.

Model #	Variable(s)	HR	POD	FAR	HSS	BIAS
1	All	0.776	0.793	0.237	0.552	1.039
2	No LCL	0.761	0.797	0.260	0.522	1.078
3	No 0-1 km capeshear	0.827	0.844	0.186	0.654	1.037
4	No 0-1 km bulk shear	0.748	0.717	0.240	0.496	0.943
5	No surface CIN	0.747	0.714	0.240	0.493	0.940
6	No SREH (0-1 km)	0.756	0.755	0.248	0.511	1.004
7	No SREH (0-3 km)	0.750	0.722	0.239	0.500	0.950
8	No EHI (0-1km)	0.780	0.771	0.219	0.559	0.988
9	No Shear	0.830	0.839	0.180	0.659	1.023
10	No SREH	0.729	0.668	0.245	0.457	0.884
11	Only LCL	0.771	0.792	0.243	0.542	1.046
12	Only surface CIN	0.595	0.479	0.382	0.189	0.776
13	Only 0-1 km bulkshear	0.738	0.756	0.274	0.476	1.042
14	Only 0-1 km capeshear	0.638	0.710	0.383	0.278	1.151
15	Only 0-1 km SREH	0.788	0.852	0.248	0.576	1.132
16	Only 0-3 km SREH	0.759	0.841	0.281	0.519	1.169
17	Only 0-1 km EHI	0.706	0.789	0.326	0.414	1.171
18	Only SREH	0.760	0.846	0.282	0.520	1.178
19	Only shear	0.641	0.707	0.379	0.283	1.140

(a)

Model #	Variable(s)	HR	POD	FAR	HSS	BIAS
20	No 0-1 km EHI	0.826	0.857	0.195	0.653	1.065
21	No 0-1 km bulk shear	0.830	0.839	0.180	0.659	1.023
22	No 0-1 km SREH	0.809	0.811	0.195	0.619	1.007
23	No 0-3 km SREH	0.830	0.823	0.168	0.661	0.989
24	No LCL	0.793	0.861	0.245	0.587	1.140
25	No surface based CIN	0.838	0.880	0.190	0.677	1.087
26	No sreh (all)	0.770	0.744	0.219	0.540	0.952

(b)

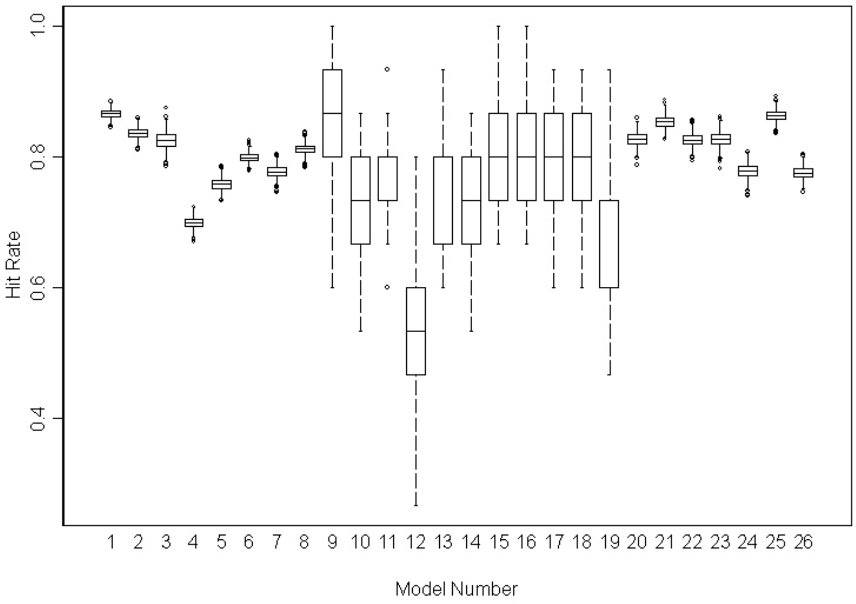


Fig. 20. Same as Fig. 12, but for LR.

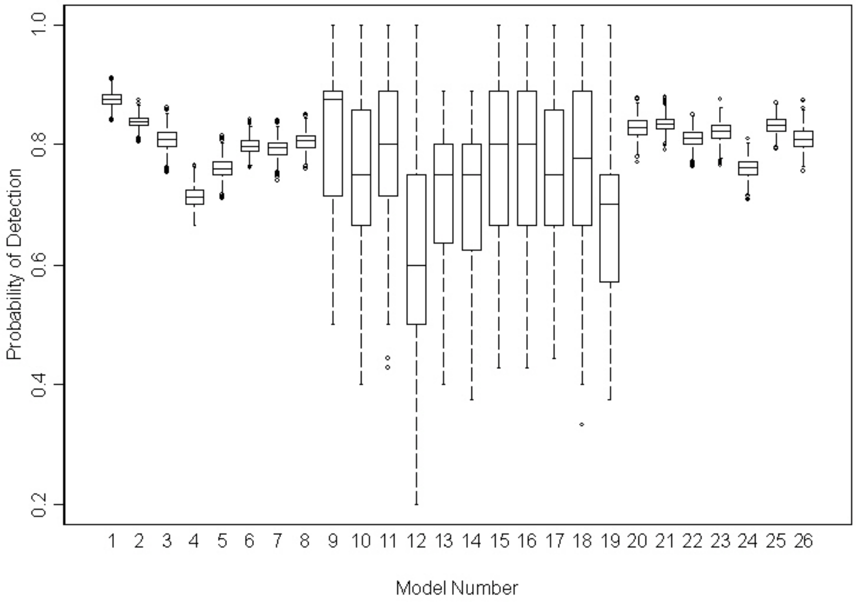


Fig. 21. Same as Fig. 20, but for POD.

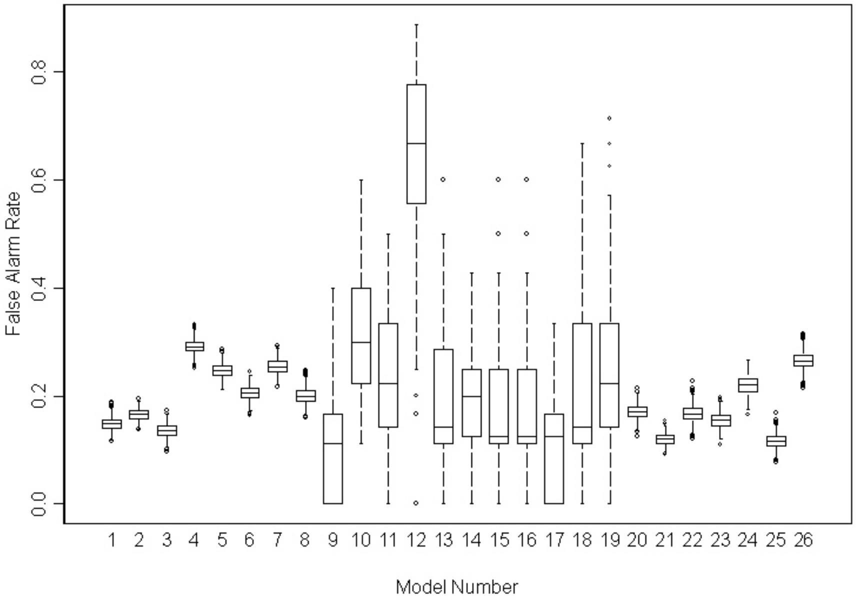


Fig. 22. Same as Fig. 20, but for FAR.

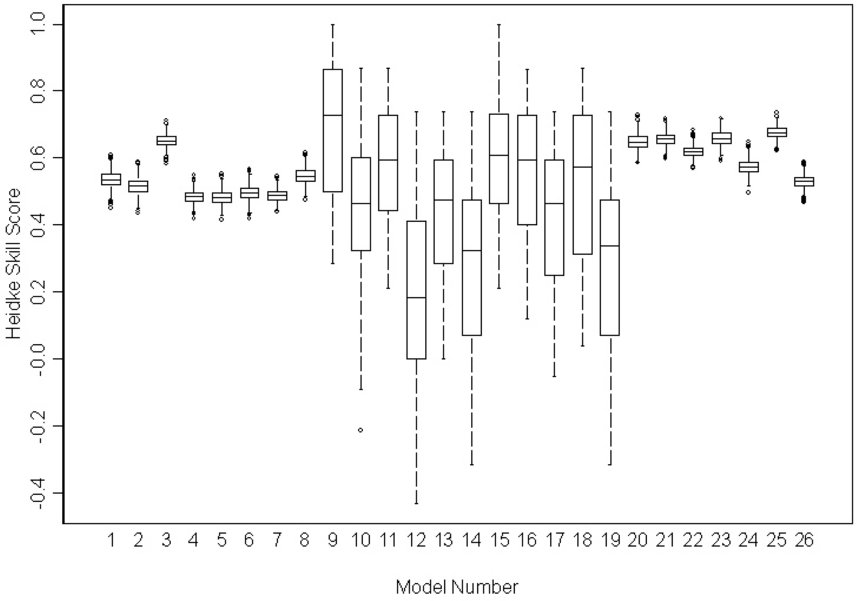


Fig. 23. Same as Fig. 20, but for HSS.

4) Synthesis

As a means of comparing the three statistical techniques, the best covariate combination (determined subjectively to be combination 25 for all boxplot results) was compared between the three methods. The individual techniques were judged based on the median value for the four contingency statistics and on the size of the IQR. Table 11 shows confidence limits of the bootstrap results for each of the four contingency statistics.

Table 11. Inter comparison of the three methods employed for classification.

	5% Limit	Median	95% Limit
	HR		
SVM	0.847	0.862	0.877
LogR	0.829	0.843	0.856
LR	0.824	0.839	0.853
	POD		
SVM	0.810	0.832	0.854
LogR	0.813	0.840	0.863
LR	0.791	0.814	0.837
	FAR		
SVM	0.096	0.116	0.139
LogR	0.123	0.141	0.159
LR	0.097	0.114	0.132
	HSS		
SVM	0.698	0.728	0.785
LogR	0.652	0.681	0.708
LR	0.662	0.691	0.718

The results in Table 11 reveal SVM as the optimal method when considering the HR (median SVM HR is larger than the 95% limit HR in LogR and LR), FAR (SVM is superior to LogR only), and HSS (SVM is clearly better than both other methods). Hence, the confidence limits show to a 90% confidence that SVM has the best bootstrapped contingency statistics and is the best method to use.

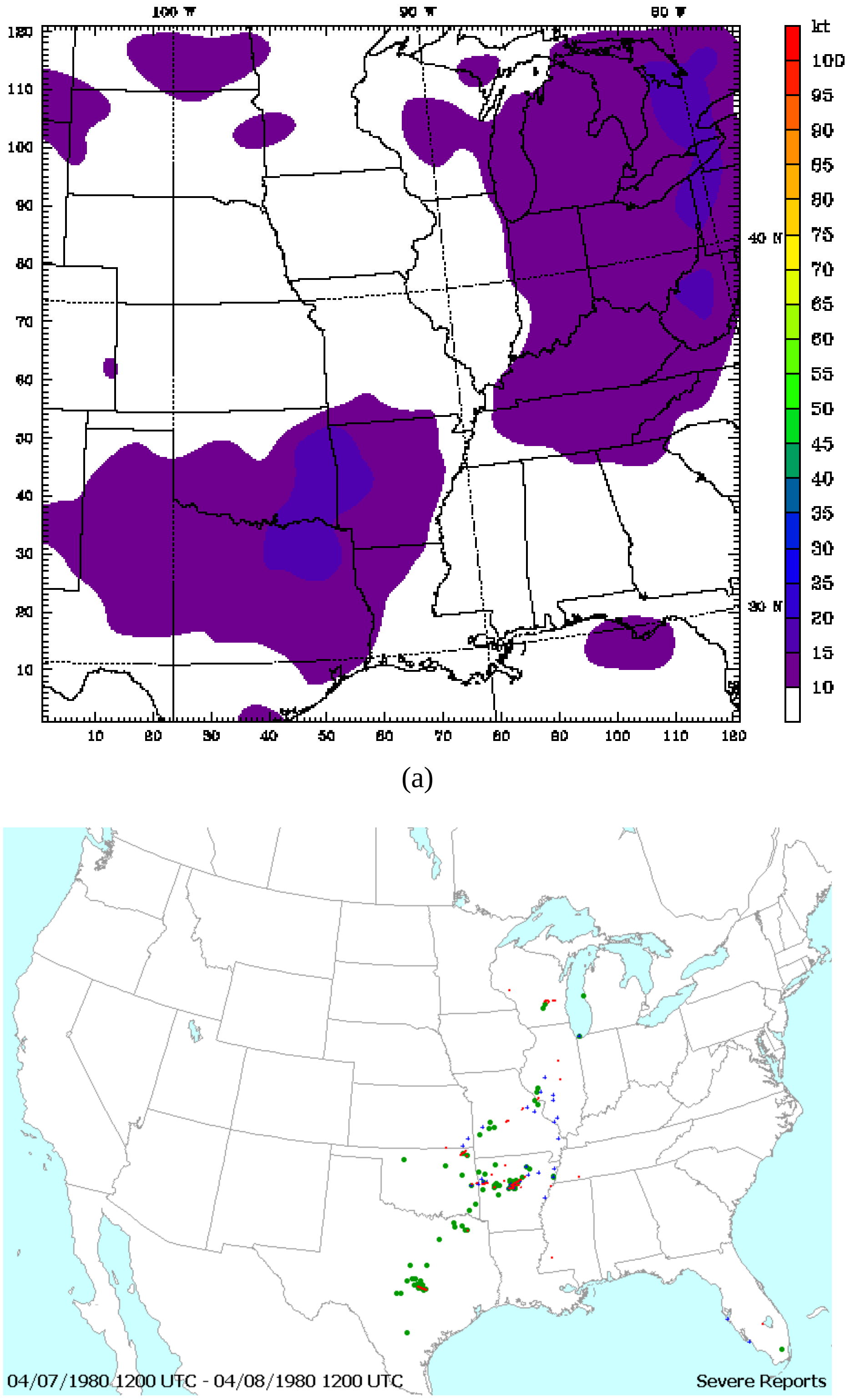
5) Case-by-case performance assessment

The performance of each statistical technique on each individual case was assessed as well. This analysis shows which cases were the most difficult to classify, which allows for subsequent investigation into the reasons for these difficulties. Each case was tested 15 times in the jackknifing methodology, and 26 covariate combinations were used. This resulted in 390 individual classification attempts by the statistical techniques for each case.

SVMs classified nine cases with 100% accuracy (i.e. the cases were classified correctly each time they were used for testing in all 26 covariate combinations). A majority of these were TOs, as only two NTOs were classified with 100% accuracy. Analysis of the worst 10 cases revealed that five TOs and five NTOs were handled poorly by SVMs. LogR and LR were more successful than SVMs with the more obvious outbreak types (over 15 classified with 100% accuracy by both methods). However, 24% of the bottom 10 cases were correctly classified by SVM, whereas only 14% were classified correctly with LogR and 9% with LR. Since SVMs perform better on the marginal cases, by this analysis, SVMs are still the best classification method at 24-hours.

Since individual case results were retained, investigation into causes of the poor classification of these worst 10 cases was performed. This investigation involved analysis of each case's covariate fields to determine of it was obvious which outbreak type was progged to occur by the WRF output. The bulk shear 24-hour WRF forecast, which was shown previously to be a good discriminator of outbreak type, was compared to the eventual outbreak storm reports (e.g. Fig. 24) to determine if the placement of the covariate suggested the correct outbreak location. If this location is incorrect, it is likely that the statistical method will not be able to distinguish the outbreak type correctly, since the input data into the method are not correctly located. This error type was denoted herein as “WRF error”. Additionally, if the atmosphere produced conditions which appeared as one outbreak type (i.e. small shear values should result in an NTO, provided an outbreak occurs) and the other occurred, this was denoted as “WRF error” as well. However, if atmospheric conditions appeared like an outbreak type that later developed, and that event was misclassified by the statistical methods, these errors were denoted as “statistical model error.”

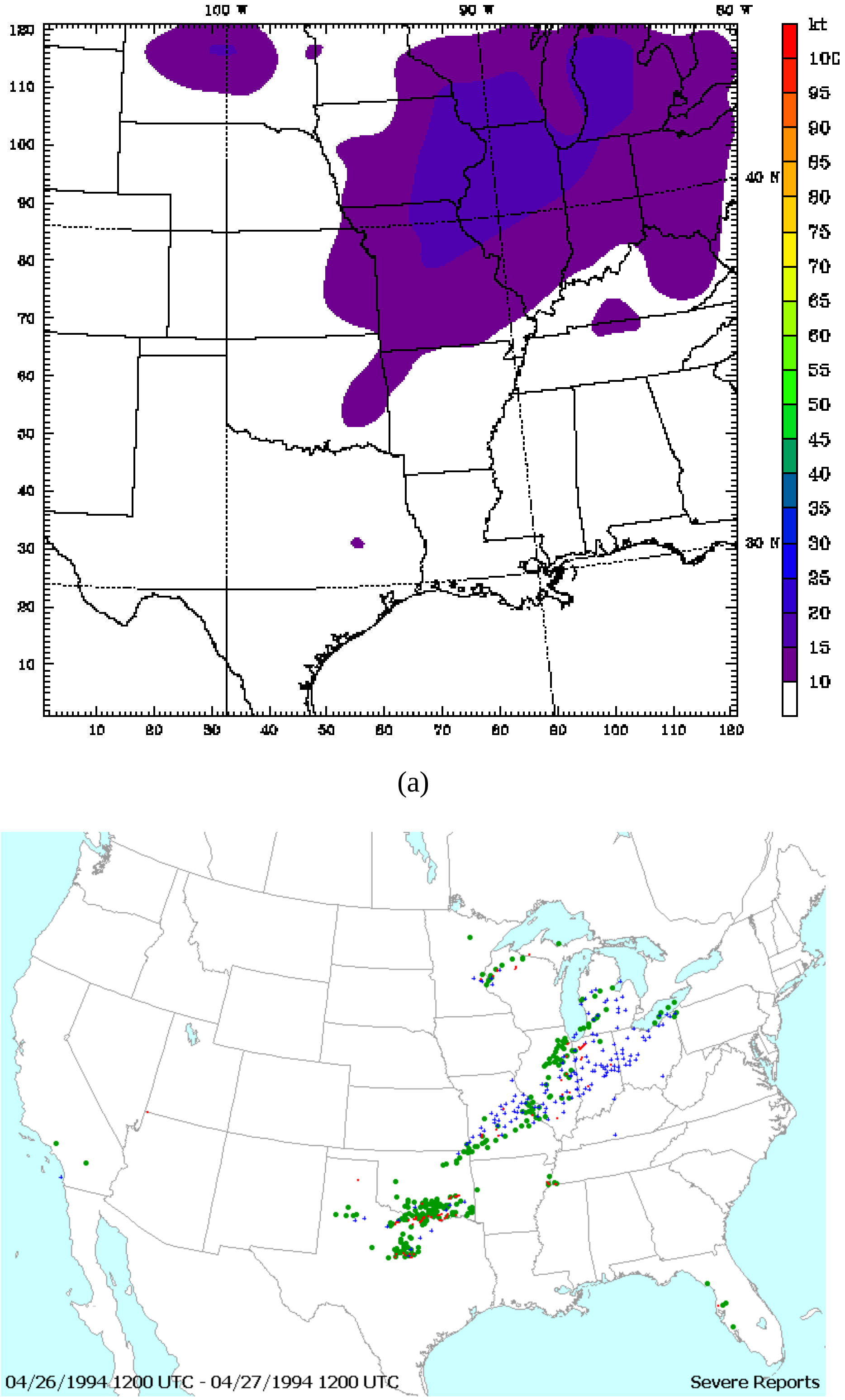
To show examples of these error types, three sample events are provided. One event (7 April 1980 – Fig. 24) was classified perfectly by SVMs at 24 hours. Noticeably large bulk shear values over the outbreak region indicated an eventual TO, and the SVMs detected this feature and classified this outbreak correctly each time. To contrast the 7 April 1980 event, a poorly classified case by SVMs (26 April 1994 – Fig. 25) was considered. On this day, large bulk shear values existed over the Ohio Valley, but no shear was forecast by the WRF over the Red River Valley, where many tornadoes occurred. Hence, the WRF produced conditions which were not expected to be associated with a TO. This error type was classified as a “WRF error.” The final event (26 September 1973 – Fig. 26) given was also classified poorly by SVMs. This TO showed large magnitudes of 0-1 km bulk shear over the eventual outbreak region 24-hours prior to the outbreak. In spite of this evidence of a looming TO, SVMs classified this event as a NTO over 50% of the time. These analyses were conducted for the worst 10 cases for each statistical technique.



(a)

(b)

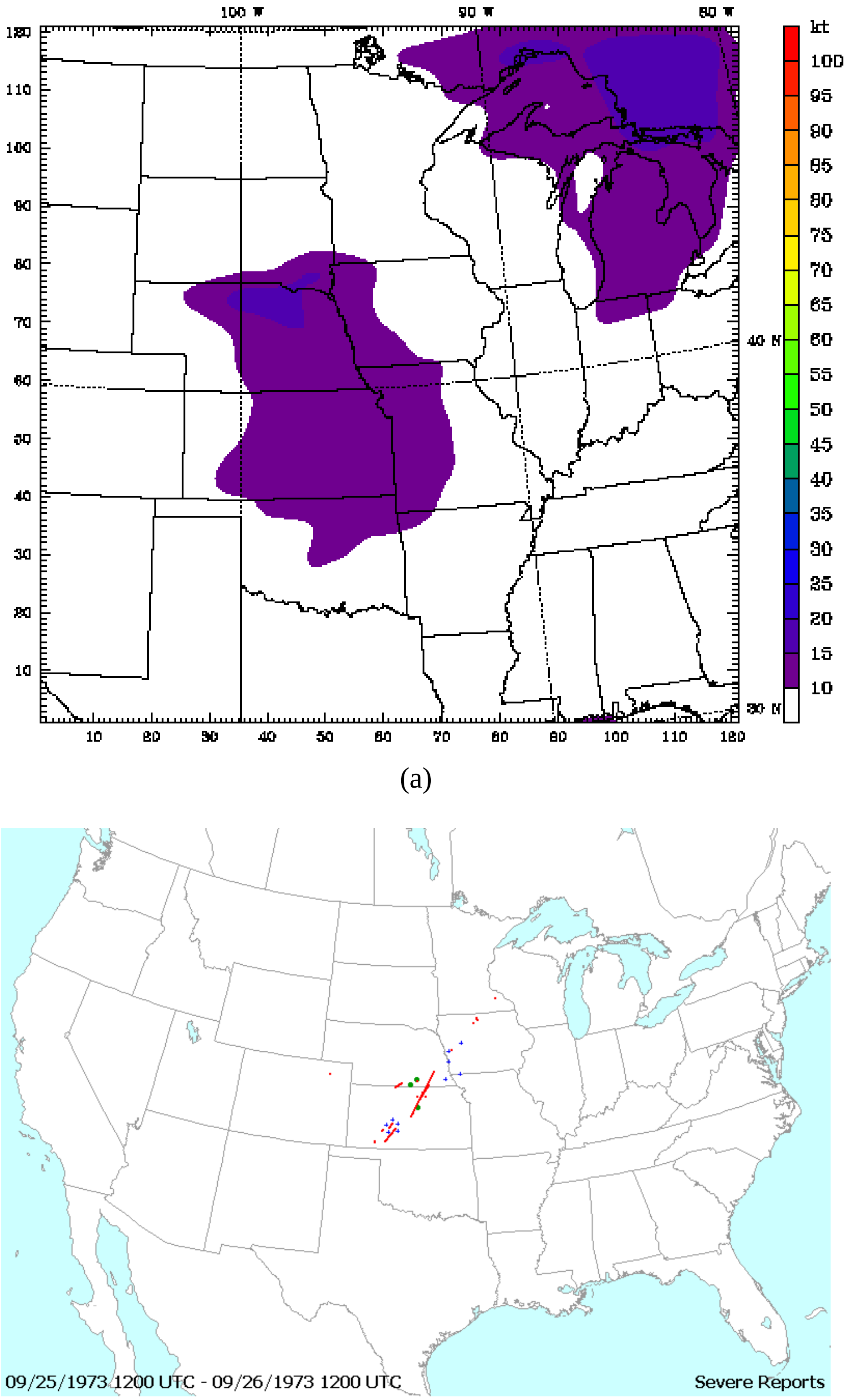
Fig. 24. Plot of 0-1 km bulk shear valid at the time of the outbreak for 07 April 1980. Contributed by Shafer (2007).



(a)

(b)

Fig. 25. Same as Fig. 24, but for 26 April 1994



(a)

Fig. 26. Same as Fig. 24, but for 26 September 1973.

The case error analysis results (Table 12) reveal that of the 30 bottom 10 cases for the three statistical techniques; only 14 cases are seen, revealing significant overlap between the worst 10 cases for each techniques. These results did not present a distinct source of error, since seven cases were classified poorly due to "WRF error" and seven were classified poorly owing to statistical model error. Model verification, which is outside of the scope of this project, combined with further training of marginal cases, will provide more insight as to the true source of these errors.

Table 12. Source of errors for the 14 cases that were in the bottom 10 for each statistical technique at 24-hours lead time.

Case	WRF Error?	Statistical Model Error?	Which Technique?
010409		x	LR, LogR, SVM
020508		x	LR, LogR, SVM
700417		x	LR, LogR, SVM
930507	x		LR, LogR, SVM
891120	x		SVM
940410		x	LR, LogR, SVM
030506	x	x	SVM
940426	x		SVM
000423	x		LR, LogR
990503	x		LR, LogR, SVM
010414		x	LR
730526		x	LR, LogR
990408		x	LR
020816	x		SVM, LogR

b) 48 Hour Results

At 48-hours, the results from the permutation testing suggested different shear covariates (0-6 km and 0-3 km bulk shear and BRN shear) and fewer instability covariates were best for 48-hour classification. Backward elimination of covariates was conducted as well, yielding 25 covariate combinations.

1) SVM CONTINGENCY AND CONFIDENCE LIMIT RESULTS

The contingency statistics (Table 13) for SVMs are somewhat degraded from those obtained at 24-hours, as might be expected. In the initial analysis, the best POD and FAR results result from the culling of 0-3 km SREH, consistent across all three statistical methods. The second stage culled 0-3 km SREH and an additional covariate (Table 13b). The best contingency statistics are associated with model 22.

The SVM boxplots are similar to those at 24-hours, as many of the covariate sets which considered only one or two covariates resulted in large variability. The HR results (Fig. 27) reveal that model 22 had the smallest variability and highest median of all of the combinations, consistent with the POD boxplots (Fig. 28). However, model 22 was subject to higher FAR than either models 7 or 23 (Fig. 29), and the HSS results (Fig. 30) showed either 1, 22, or 23 had the best overall performance as input into the SVMs. This result implied no single covariate combination was best for 48-hour classification. Additionally, median values of the contingency statistics at 48-hours were generally less than 10% worse than those at 24-hours, revealing only modest degradation of results with 24 hours more lead time. Since numerical weather prediction simulations generally worsen with increased lead time, this result was expected, but the relatively slow degradation is encouraging and obviously warranted further investigation into the classification performance at 72-hours.

Table 13. Contingency table results for SVMs. Table (a) represents the variables as stated, while Table (b) removes the stated variable and 0-3 km SREH.

Model #	Variable(s)	HR	POD	FAR	HSS	BIAS
1	All	0.808	0.810	0.196	0.616	1.007
2	No LCL	0.770	0.805	0.251	0.541	1.075
3	No shear (0-6 km)	0.810	0.815	0.196	0.620	1.014
4	No shear (0-3 km)	0.817	0.829	0.193	0.634	1.027
5	No shear (0-1 km)	0.768	0.795	0.248	0.537	1.057
6	No SREH (0-1 km)	0.789	0.795	0.218	0.577	1.016
7	No SREH (0-3 km)	0.808	0.838	0.212	0.616	1.064
8	No BRN shear	0.812	0.804	0.186	0.624	0.988
9	No shear variables	0.801	0.824	0.216	0.602	1.052
10	No SREH variables	0.822	0.833	0.187	0.644	1.024
11	Only shear variables	0.770	0.815	0.256	0.540	1.095
12	Only SREH variables	0.711	0.774	0.316	0.423	1.132
13	Just LCL	0.749	0.778	0.269	0.498	1.064
14	Just shear (0-6 km)	0.721	0.709	0.278	0.441	0.982
15	Just shear (0-3 km)	0.770	0.812	0.254	0.541	1.088
16	Just shear (0-1 km)	0.752	0.714	0.232	0.503	0.931
17	Just SREH (0-3 km)	0.714	0.766	0.310	0.428	1.110
18	Just SREH (0-1 km)	0.689	0.752	0.336	0.379	1.133
19	Just BRN shear	0.692	0.709	0.319	0.383	1.041

(a)

Model #	Variable(s)	HR	POD	FAR	HSS	BIAS
20	No LCL (all)	0.724	0.782	0.303	0.448	1.122
21	No shear (0-6) (all)	0.801	0.820	0.214	0.601	1.044
22	No shear (0-3 km)	0.830	0.844	0.182	0.659	1.031
23	No shear (0-1 km)	0.820	0.863	0.208	0.641	1.088
24	No SREH (0-1 km)	0.822	0.833	0.187	0.644	1.024
25	No BRN shear (all)	0.807	0.797	0.189	0.615	0.984

(b)

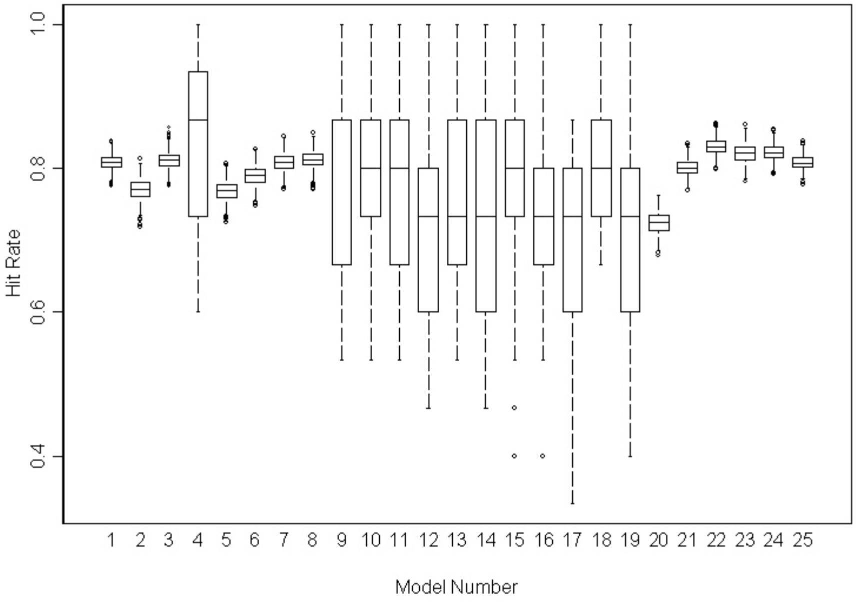


Fig. 27. Boxplots for 48 hour SVM HR. Model numbers correspond with the row number in Table 13.

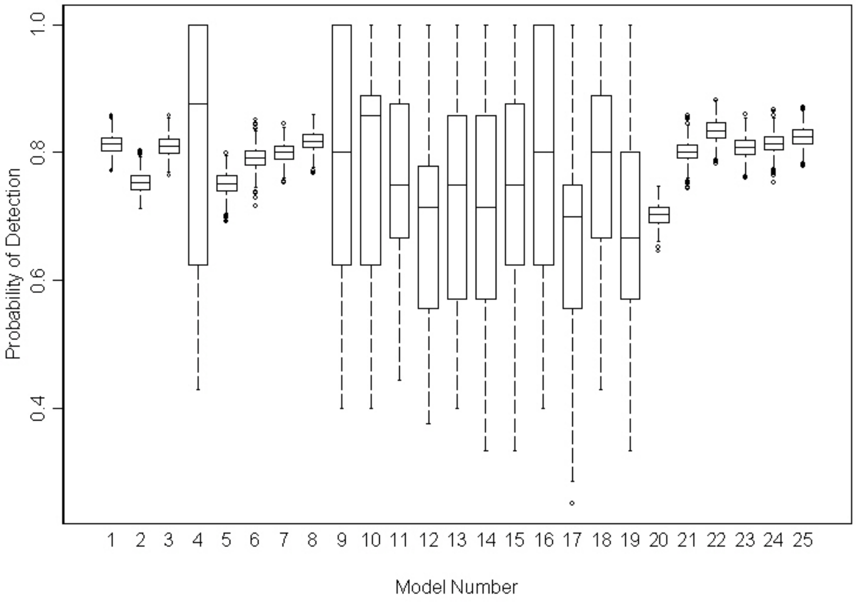


Fig. 28. Same as Fig. 27, but for POD.

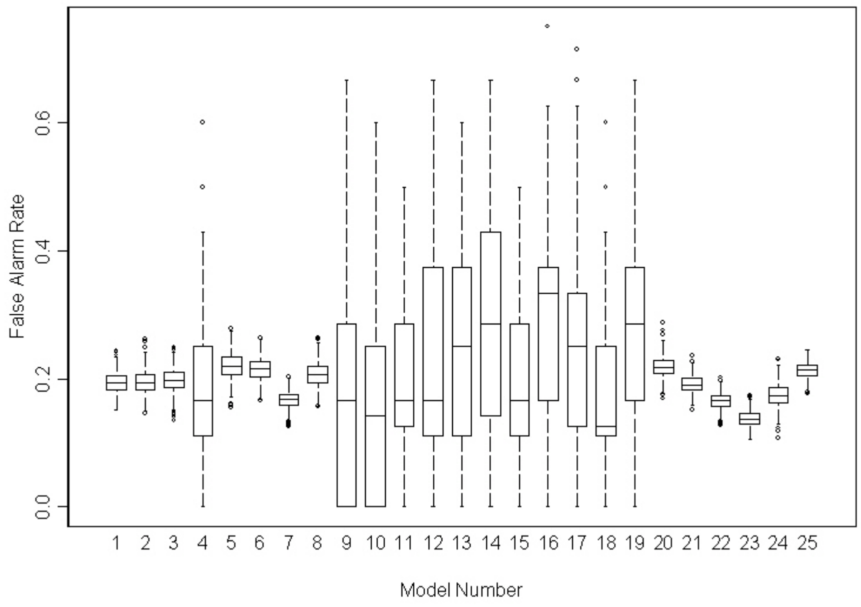


Fig. 29. Same as Fig. 27, but for FAR.

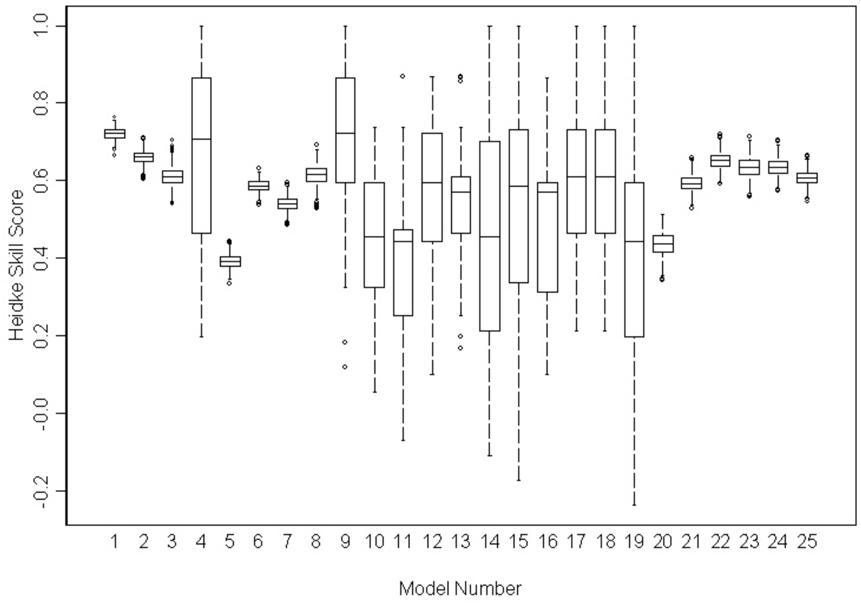


Fig. 30. Same as Fig. 27, but for HSS.

2) LogR contingency and confidence limit results

The initial jackknife contingency statistic results for LogR (Table 14) were consistent with the SVM results, as rejecting 0-3 km SREH improved results. In contrast to the SVM results, further rejection of covariates besides 0-3 km SREH resulted in reduced accuracy, so no additional culling of covariates was done. Model seven was the best model in terms of the jackknifed contingency statistics.

The HR and POD statistics (Figs. 31 and 32) show the narrowest IQR and highest median value in model seven, consistent with the jackknife contingency results. The FAR calculations (Fig. 33) showed models two or seven as having the lowest FAR, and the HSS results (Fig. 34) supported the conclusion that model seven provided the superior results. One feature of the boxplots not seen with SVM was the large variability in model four, which only rejected 0-3 km bulk shear. This result supports 0-3 km bulk shear as a good covariate, but is not consistent with the previous conclusion that more covariates input into the model produced less variability in the boxplots. As with SVM, the boxplot and jackknife contingency results were within 10% of the 24-hour results, which further supports the conclusion that little drop-off of the classification ability of these methods was noted at 48-hours prior to the outbreak.

Table 14. Same as Table 13, but for LogR.

Model #	Variable(s)	HR	POD	FAR	HSS	BIAS
1	All	0.817	0.839	0.200	0.634	1.049
2	No LCL	0.811	0.844	0.211	0.622	1.069
3	No shear (0-6 km)	0.810	0.819	0.198	0.620	1.022
4	No shear (0-3 km)	0.805	0.808	0.201	0.609	1.011
5	No shear (0-1 km)	0.785	0.829	0.242	0.569	1.093
6	No SREH (0-1 km)	0.803	0.834	0.217	0.607	1.065
7	No SREH (0-3 km)	0.838	0.844	0.169	0.675	1.015
8	No BRN shear	0.801	0.822	0.215	0.601	1.046
9	No shear variables	0.804	0.822	0.209	0.608	1.039
10	No SREH variables	0.776	0.818	0.248	0.553	1.087
11	Only shear variables	0.826	0.845	0.189	0.651	1.042
12	Only SREH variables	0.760	0.810	0.267	0.520	1.105
13	Just LCL	0.691	0.684	0.311	0.382	0.993
14	Just shear (0-6 km)	0.772	0.780	0.236	0.543	1.020
15	Just shear (0-3 km)	0.786	0.808	0.230	0.572	1.049
16	Just shear (0-1 km)	0.763	0.740	0.228	0.526	0.959
17	Just SREH (0-3 km)	0.754	0.786	0.265	0.509	1.069
18	Just SREH (0-1 km)	0.733	0.788	0.294	0.466	1.116
19	Just BRN shear	0.741	0.774	0.278	0.482	1.072

(a)

Model #	Variable(s)	HR	POD	FAR	HSS	BIAS
20	No LCL (all)	0.776	0.812	0.245	0.553	1.076
21	No shear (0-6) (all)	0.803	0.800	0.199	0.605	0.999
22	No shear (0-3 km)	0.807	0.808	0.196	0.615	1.005
23	No shear (0-1 km)	0.791	0.818	0.227	0.583	1.057
24	No SREH (0-1 km)	0.776	0.818	0.248	0.553	1.087
25	No BRN shear (all)	0.795	0.800	0.212	0.589	1.015

(b)

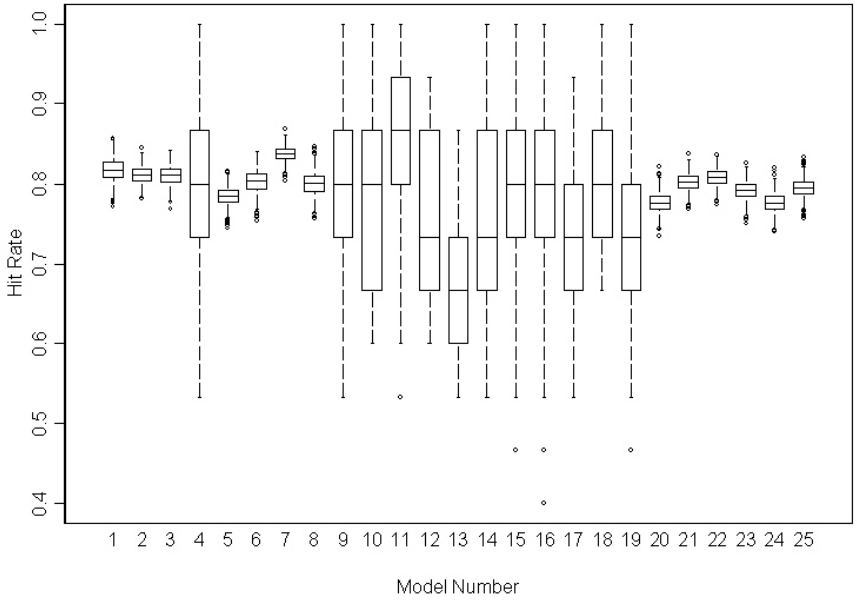


Fig. 31. Same as Fig. 27, but for LogR.

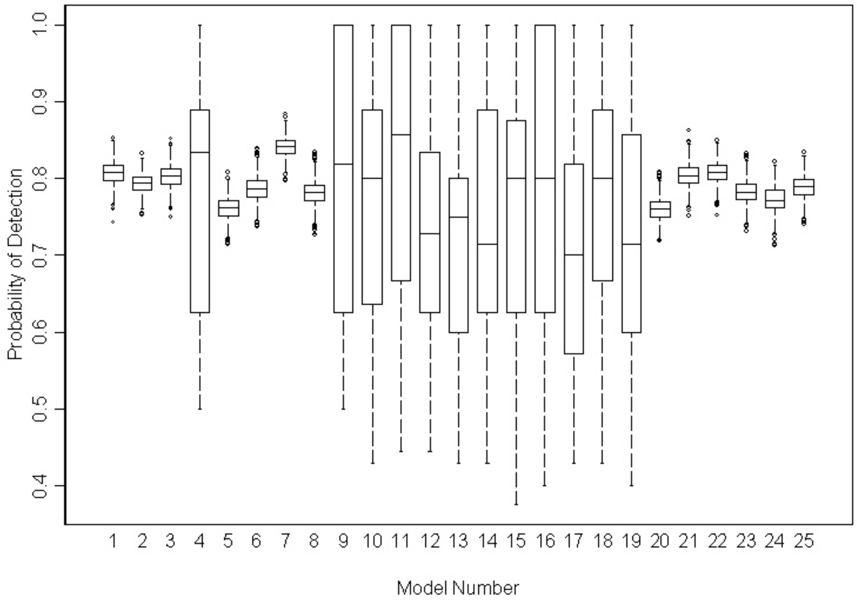


Fig. 32. Same as Fig. 31, but for POD.

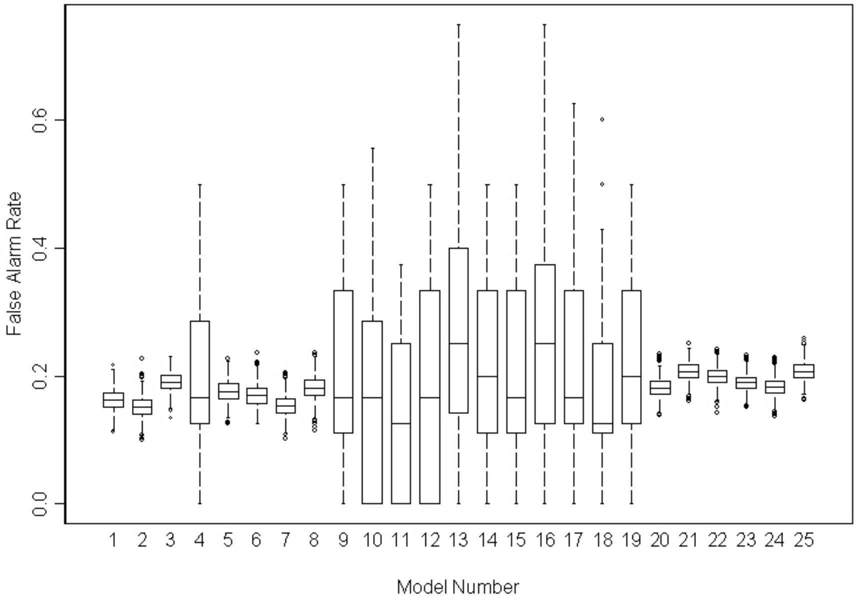


Fig. 33. Same as Fig. 31, but for FAR.

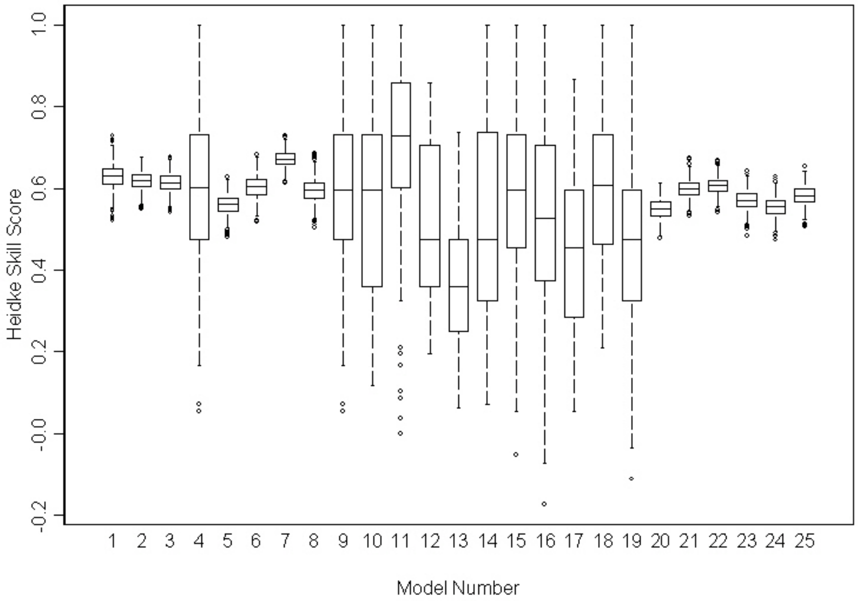


Fig. 34. Same as Fig. 31, but for HSS.

3) LR CONTINGENCY AND CONFIDENCE LIMIT RESULTS

The initial LR jackknife contingency statistics (Table 15) show the best results when 0-1 km SREH was rejected. This result differs from both LogR and SVM, which suggested removing 0-3 km SREH. However, consistent with LogR, additional rejection of covariates with LR led to accuracy reductions. Since three different covariate sets were determined from the three statistical methods, as lead time increases, the best covariate combination becomes more dependent on the statistical method being tested.

For the boxplot LR results, the HR (Fig. 35) and POD (Fig. 36) show the highest medians when considering model six, consistent with the jackknife contingency statistic results. Numerous covariate combinations resulted in low FAR results (Fig. 37), and it was not possible to distinguish which one was superior. The HSS results provided an overall performance measure of LR (Fig. 38), and show model six with the highest median value. These results are consistently within 10% of those at 24-hours, a result common to the three statistical methods tested.

Table 15. Same as Table 13, but for LR.

Model #	Variable(s)	HR	POD	FAR	HSS	BIAS
1	All	0.818	0.856	0.208	0.635	1.080
2	No LCL	0.776	0.856	0.265	0.552	1.165
3	No shear (0-6 km)	0.809	0.822	0.202	0.618	1.030
4	No shear (0-3 km)	0.807	0.823	0.205	0.615	1.035
5	No shear (0-1 km)	0.785	0.842	0.247	0.571	1.118
6	No SREH (0-1 km)	0.828	0.857	0.193	0.655	1.063
7	No SREH (0-3 km)	0.816	0.860	0.212	0.633	1.091
8	No BRN shear	0.807	0.827	0.207	0.615	1.044
9	No shear variables	0.807	0.841	0.216	0.614	1.072
10	No SREH variables	0.774	0.829	0.256	0.548	1.114
11	Only shear variables	0.758	0.839	0.281	0.517	1.167
12	Only SREH variables	0.754	0.837	0.286	0.508	1.171
13	Just LCL	0.686	0.683	0.318	0.371	1.001
14	Just shear (0-6 km)	0.745	0.793	0.280	0.490	1.102
15	Just shear (0-3 km)	0.752	0.796	0.272	0.503	1.094
16	Just shear (0-1 km)	0.723	0.754	0.293	0.447	1.067
17	Just SREH (0-3 km)	0.766	0.842	0.273	0.532	1.158
18	Just SREH (0-1 km)	0.743	0.830	0.296	0.488	1.178
19	Just BRN shear	0.744	0.808	0.287	0.489	1.133

(a)

Model #	Variable(s)	HR	POD	FAR	HSS	BIAS
20	No LCL	0.763	0.831	0.272	0.527	1.141
21	No shear (0-6 km)	0.795	0.814	0.218	0.591	1.041
22	No shear (0-3 km)	0.798	0.815	0.215	0.596	1.038
23	No shear (0-1 km)	0.785	0.827	0.241	0.569	1.090
24	No SREH (0-1 km)	0.774	0.829	0.256	0.548	1.114
25	No BRN shear	0.784	0.804	0.230	0.568	1.045

(b)

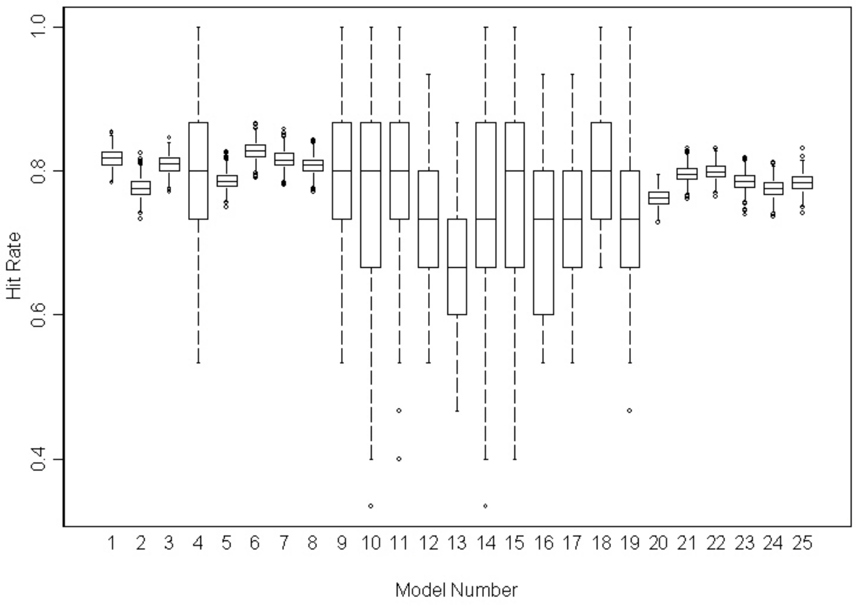


Fig. 35. Same as Fig. 27, but for LR.

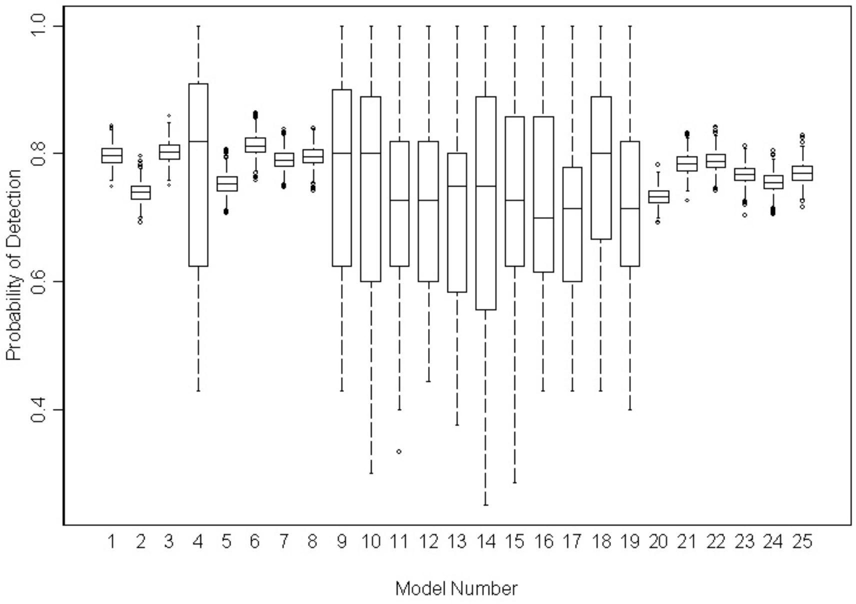


Fig. 36. Same as Fig. 35, but for POD.

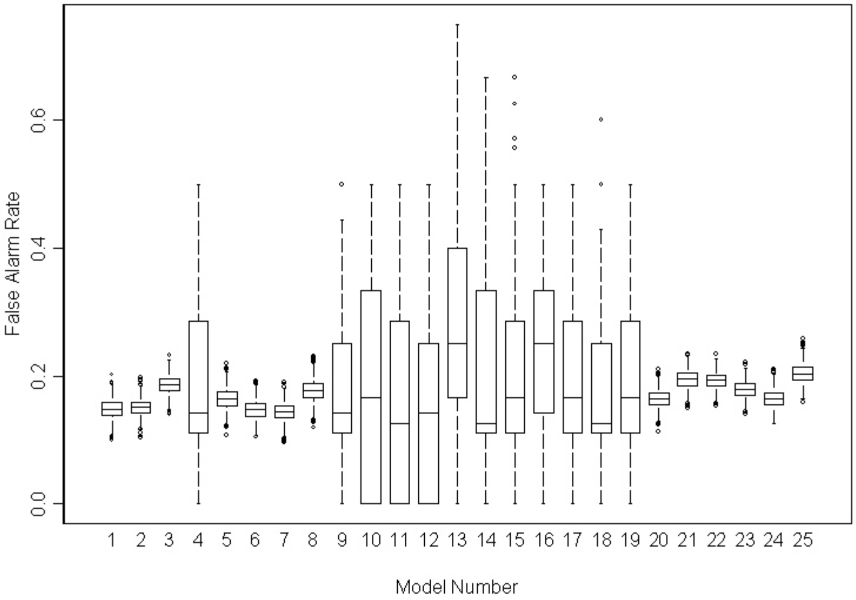


Fig. 37. Same as Fig. 35, but for FAR.

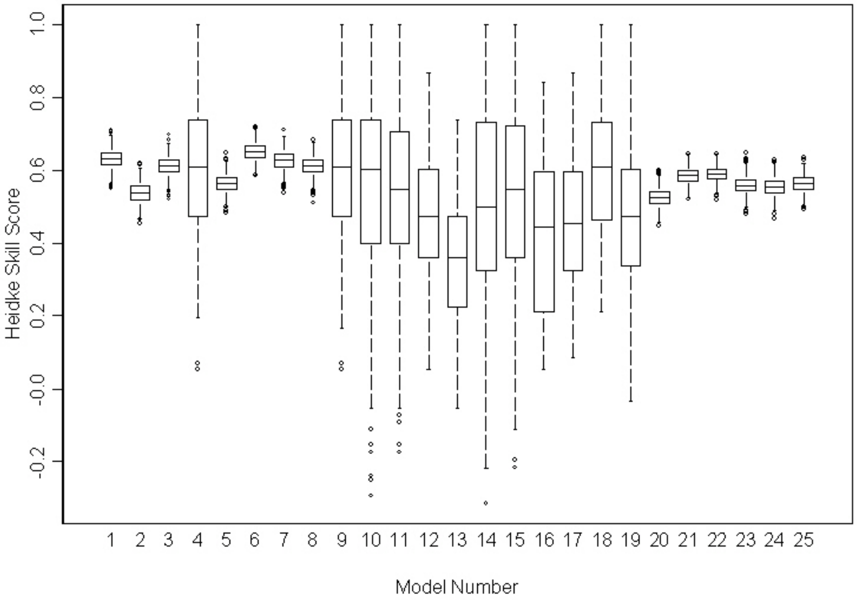


Fig. 38. Same as Fig. 35, but for HSS.

4) SYNTHESIS

Since the optimal covariate sets are consistently different between the three statistical techniques considered, the best set from each technique (model 22 for SVM, model seven for LogR, and model six for LR) was compared to determine the most favorable technique at 48-hours. Other than the LogR POD result being statistically significantly larger than the LR POD result (Table 16), no significant differences between the three methods are observed at 48-hours. Hence, either SVM or LogR, since they are in a statistical tie for the four contingency statistics, are the best methods to use.

Table 16. Same as Table 10, but for 48-hours lead time.

	5% Limit	Median	95% Limit
	HR		
SVM	0.812	0.830	0.847
LogR	0.822	0.838	0.854
LR	0.808	0.827	0.846
	POD		
SVM	0.806	0.835	0.862
LogR	0.821	0.840	0.861
LR	0.786	0.813	0.838
	FAR		
SVM	0.146	0.165	0.184
LogR	0.127	0.152	0.180
LR	0.124	0.147	0.171
	HSS		
SVM	0.615	0.652	0.686
LogR	0.640	0.672	0.703
LR	0.611	0.650	0.688

5) CASE-BY-CASE PERFORMANCE ASSESSMENT

The first result, which was surprising owing to the increased lead time, was that more cases were classified with 100% accuracy by all three methods at 48 hours lead time (17 for SVMs, 31 for LogR, and 28 for LR). For SVMs, roughly the same number of TOs and NTOs were classified with 100% accuracy, which showed that SVMs increased in ability to classify NTOs by 48-hours. This was not the case for LogR or LR, which classified both types equally at 24 and 48-hours. The worst cases were classified best by SVM (6% correct), but these results were only trivially better than LogR (5% correct) and LR (4% correct). Since LogR classified the most cases with 100% accuracy, LogR was deemed the best method for 48-hour classification of outbreak type.

Table 17 compares the type of error for the bottom 10 cases for each technique, where the error definitions are consistent with those provided at 24-hours. In contrast to 24-hours, the 48-hour results showed that “WRF error” was consistently responsible for erroneous outbreak classification of these bottom 10 cases. Since statistical model error was not as prevalent at 48-hours, the selection of covariates at 48-hours may have been responsible for the better case-by-case results at 48-hours. Of the covariates selected at 48-hours, no measure of instability was included. This covariate, which was seen to introduce error into the results at 24-hours, may have worsened case-by-case performance results, since all 26 covariate combinations at 24-hours were considered in the performance analysis.

Table 17. Source of errors for the 14 cases that were in the bottom 10 for each statistical technique at 48-hours lead time.

Case	WRF Error?	Statistical Model Error?	Which Technique?
010409	x		LR, LogR, SVM
800712	x		LR, SVM
700417	x		LR, LogR, SVM
930507	x		LR, LogR
891120	x		LR, LogR, SVM
890521	x		LR, LogR, SVM
730925	x		LR, LogR, SVM
840607	x		LogR, SVM
900416		x	LR, LogR, SVM
990503	x		LR, LogR, SVM
730526		x	LR, LogR, SVM

c) 72 Hour Results

Since only relatively little degradation of results was found between 24 and 48 hours prior to the outbreak, an analysis at 72-hours prior to the outbreak was conducted. The permutation testing covariate set included a different combination of six of the same covariates used at 24- and 48-hours lead time. Since fewer covariates were used, fewer model combinations (22) were tested at 72-hours.

1) SVM CONTINGENCY AND CONFIDENCE LIMIT RESULTS

The SVM jackknife contingency statistic results (Table 18) showed a larger drop-off (exceeding 10%) of contingency statistics from the values noted at 48-hours. The best POD results were noted in model 13, which only used 0-3 km SREH when classifying outbreak type. This result was surprising since single covariates did not classify well at shorter lead times. The initial analysis also showed that culling 0-1 km EHI produced lower FAR but also lower HSS. As a result, 0-1 km EHI was rejected, and a secondary analysis was conducted. The best overall contingency results were obtained through rejection of 0-1 km EHI and 0-3 km SREH (model 18).

The boxplot results for SVM HR and POD (Figs. 39 and 40) reveal large median values of the statistics in models 13 and 18. However, model 13, which includes a single covariate, produced significantly large IQR, which is undesirable. This large IQR might not have been detected without a boxplot analysis, which could have lead incorrectly to recommending model 13 for outbreak classification. Model 18, which rejected 0-1 km EHI and 0-3 km SREH, shows the highest HR results in the jackknife contingency results. The FAR (Fig. 41) is smallest with model 13, but model 13 shows large IQR on the FAR results as well. Models 1, 8, 18, and 22 produce the best FAR results of those boxplots with small IQR. The highest median HSS results (Fig. 42) are observed from model 18, implying that for SVMs, model 18 is the best covariate combination to use.

Table 18. Contingency table results for SVMs. Table (a) represents the variables as stated, while Table (b) removes the stated variable and 0-1 km EHI.

Model #	Variable(s)	HR	POD	FAR	HSS	BIAS
1	All	0.669	0.770	0.363	0.339	1.210
2	No LCL	0.657	0.729	0.366	0.315	1.151
3	No bulkshear (0-6 km)	0.668	0.766	0.363	0.337	1.203
4	No bulkshear (0-3 km)	0.652	0.751	0.377	0.305	1.205
5	No SREH (0-1 km)	0.670	0.755	0.358	0.341	1.177
6	No SREH (0-3 km)	0.675	0.705	0.339	0.350	1.067
7	No EHI (0-1 km)	0.706	0.735	0.309	0.413	1.063
8	No Shear	0.657	0.765	0.375	0.315	1.223
9	No SREH	0.696	0.727	0.319	0.393	1.067
10	Only LCL	0.650	0.731	0.375	0.301	1.169
11	Only bulkshear (0-3 km)	0.698	0.759	0.327	0.397	1.128
12	Only bulkshear (0-6 km)	0.721	0.725	0.285	0.442	1.014
13	Only SREH (0-1 km)	0.713	0.819	0.327	0.427	1.218
14	Only SREH (0-3 km)	0.662	0.781	0.373	0.325	1.245
15	Only EHI (0-1 km)	0.616	0.793	0.418	0.235	1.362
16	Only SREH variables	0.636	0.750	0.393	0.273	1.235
17	Only shear variables	0.682	0.758	0.345	0.365	1.158

(a)

Model #	Variable(s)	HR	POD	FAR	HSS	BIAS
18	No SREH (0-3 km)	0.738	0.762	0.276	0.476	1.053
19	No bulkshear (0-3 km)	0.662	0.702	0.354	0.324	1.087
20	No bulkshear (0-6 km)	0.668	0.693	0.344	0.336	1.056
21	No SREH (0-1 km)	0.696	0.727	0.319	0.393	1.067
22	No LCL	0.679	0.754	0.348	0.359	1.156

(b)

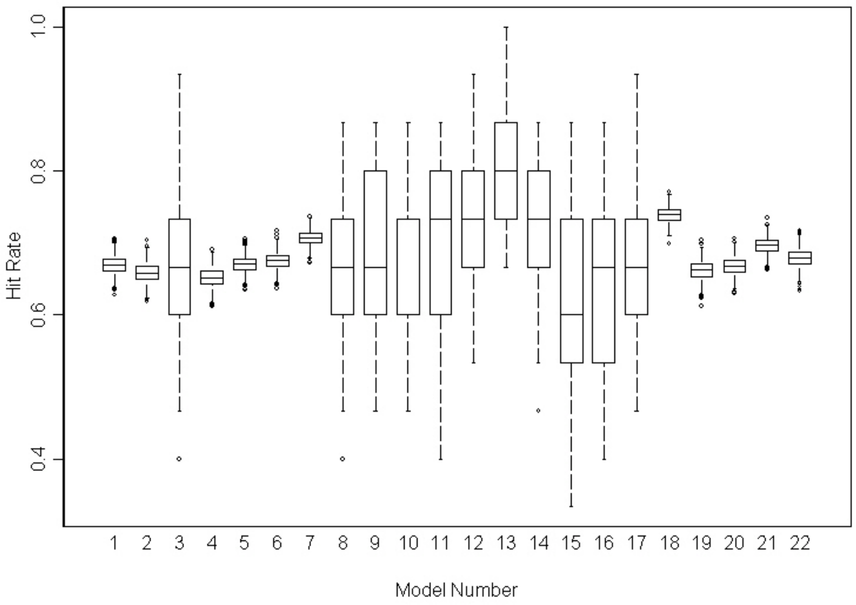


Fig. 39. Boxplots for 72 hour SVM HR. Model numbers correspond with the row number in Table 18.

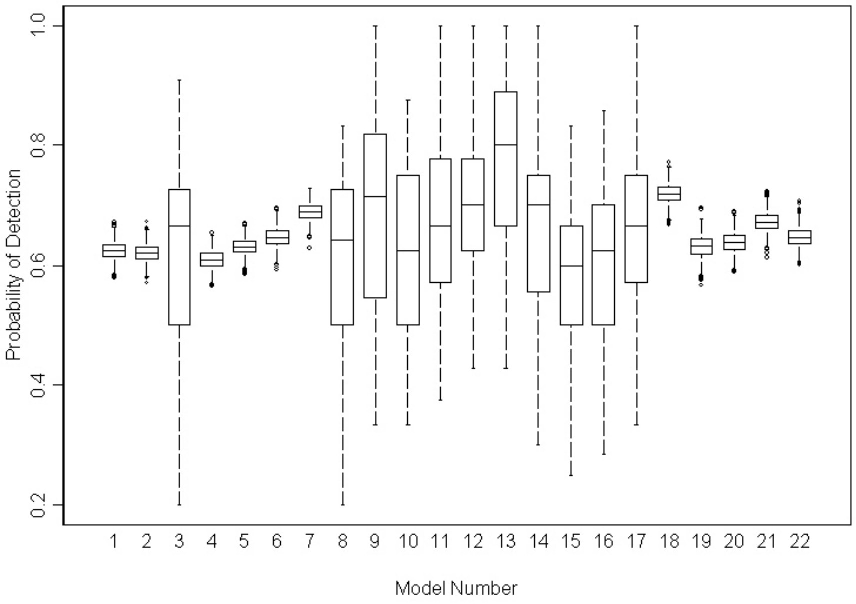


Fig. 40. Same as Fig. 39, but for POD.

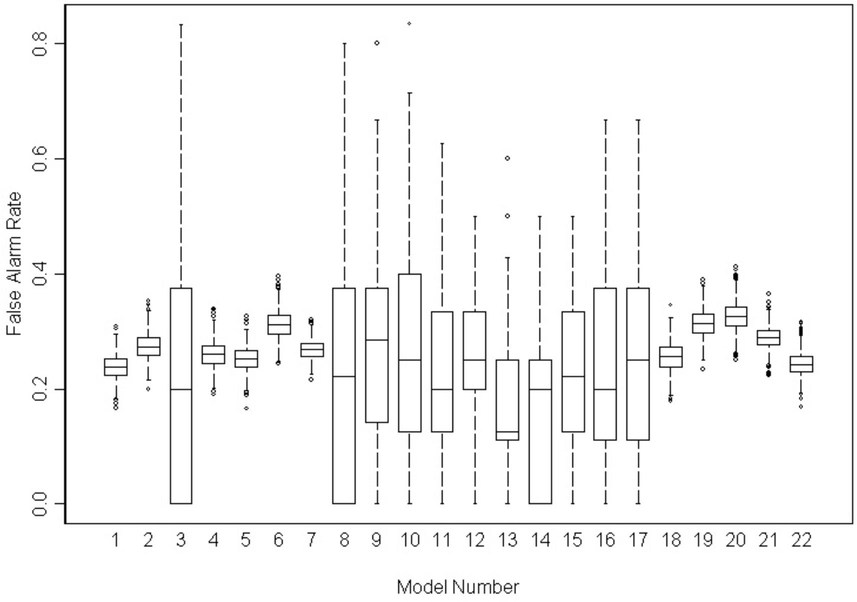


Fig. 41. Same as Fig. 39, but for FAR.

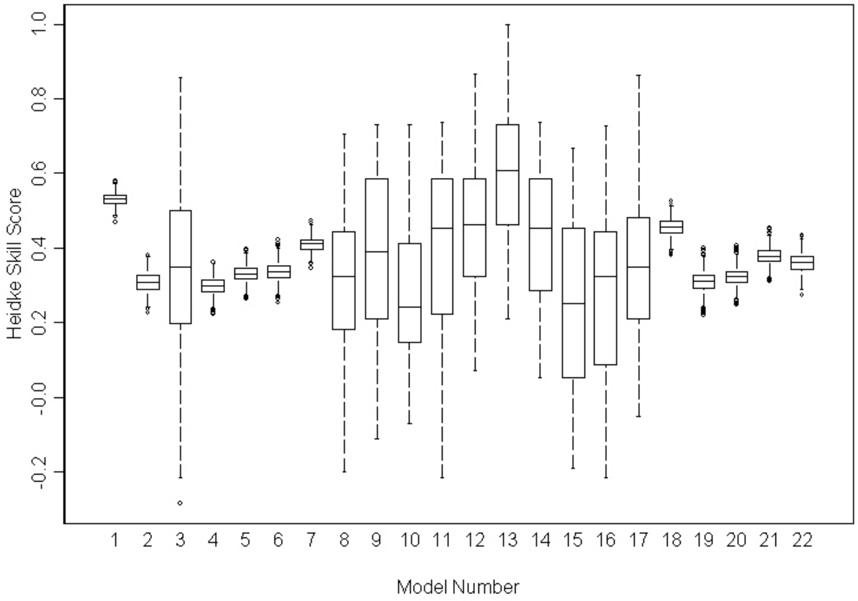


Fig. 42. Same as Fig. 39, but for HSS.

2) LogR contingency and confidence limit results

The LogR results (Table 19) reveal the best HR (0.748) and FAR (0.259) values when only considering 0-6 km bulk shear. These values are slightly better than the SVM results, although their differences are likely not statistically significant. This single covariate optimality is consistent with model 13 classifying with the highest POD for SVMs, but may be subject to large variability. The highest probability of detection values are obtained from culling the LCL (0.812). The best HSS values are obtained from only using 0-6 km bulk shear (0.496), which is consistent with FAR and HR.

In the HR and POD boxplot results (Figs. 43 and 44), models 12 and 13 (median overlaps the 3rd quartile in model 12) clearly have the highest median values. However, large variability associated with these two models led to their rejection. Numerous models with smaller IQR values have high medians for HR and POD (models 7, 19, and 20), and it was not possible to determine the best with any statistical confidence. The FAR median values (Fig. 45) are lowest in models two, seven, and 18, while the HSS medians (Fig. 46) are highest in model seven. These results indicate model seven, which culls 0-1 km EHI, produces the best results for all contingency statistics and is the best covariate combination for LogR. As was true with SVMs, the contingency statistics at 72-hours worsen by 10-20% with LogR compared to 48-hours lead time, a result that can be attributed to diminished WRF performance with increased lead time.

Table 19. Same as Table 18, but for LogR.

Model #	Variable(s)	HR	POD	FAR	HSS	BIAS
1	All	0.703	0.773	0.325	0.407	1.146
2	No LCL	0.701	0.812	0.339	0.403	1.229
3	No bulkshear (0-6 km)	0.706	0.771	0.322	0.412	1.137
4	No bulkshear (0-3 km)	0.709	0.770	0.317	0.419	1.128
5	No SREH (0-1 km)	0.669	0.722	0.351	0.339	1.113
6	No SREH (0-3 km)	0.729	0.777	0.295	0.458	1.102
7	No EHI (0-1 km)	0.734	0.804	0.298	0.469	1.146
8	No Shear	0.702	0.754	0.320	0.405	1.109
9	No SREH	0.677	0.724	0.342	0.355	1.099
10	Only LCL	0.621	0.595	0.377	0.241	0.955
11	Only bulkshear (0-3 km)	0.732	0.767	0.287	0.464	1.076
12	Only bulkshear (0-6 km)	0.748	0.755	0.259	0.496	1.019
13	Only SREH (0-1 km)	0.691	0.732	0.328	0.382	1.088
14	Only SREH (0-3 km)	0.676	0.767	0.355	0.353	1.189
15	Only EHI (0-1 km)	0.615	0.725	0.410	0.231	1.229
16	Only SREH variables	0.700	0.766	0.327	0.401	1.137
17	Only shear variables	0.706	0.717	0.302	0.413	1.027

(a)

Model #	Variable(s)	HR	POD	FAR	HSS	BIAS
18	No SREH (0-3 km)	0.727	0.808	0.308	0.455	1.167
19	No bulkshear (0-3 km)	0.731	0.777	0.292	0.462	1.098
20	No bulkshear (0-6 km)	0.731	0.774	0.291	0.462	1.093
21	No SREH (0-1 km)	0.677	0.724	0.342	0.355	1.099
22	No LCL	0.675	0.774	0.357	0.352	1.204

(b)

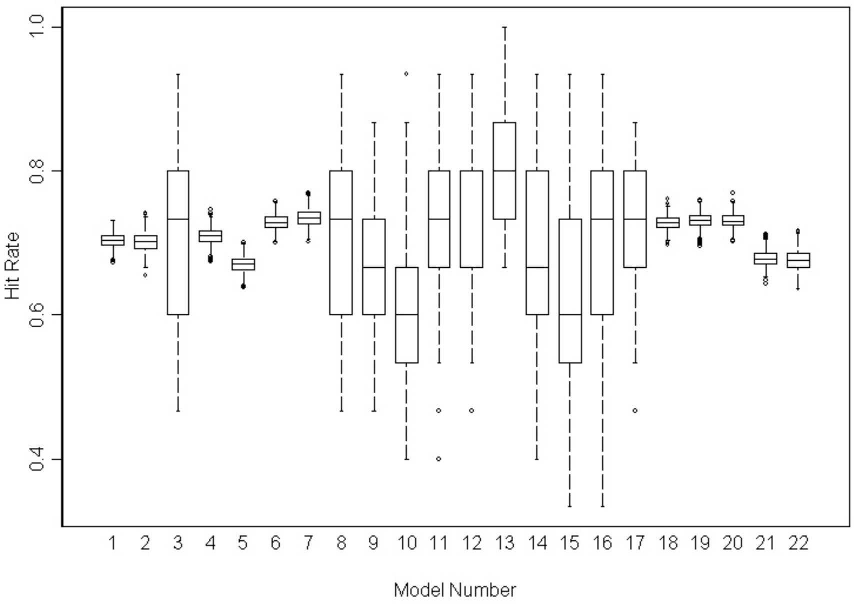


Fig. 43. Same as Fig. 39, but for LogR.

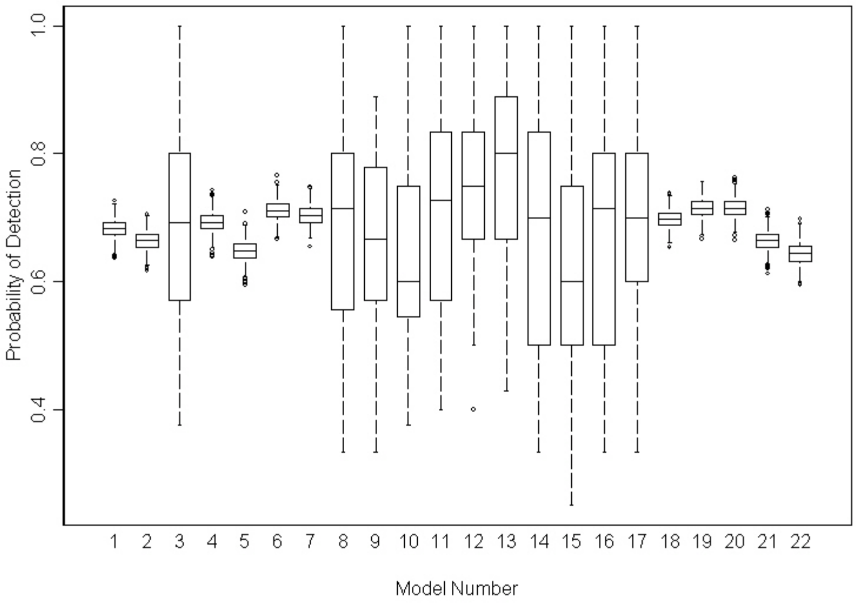


Fig. 44. Same as Fig. 43, but for POD.

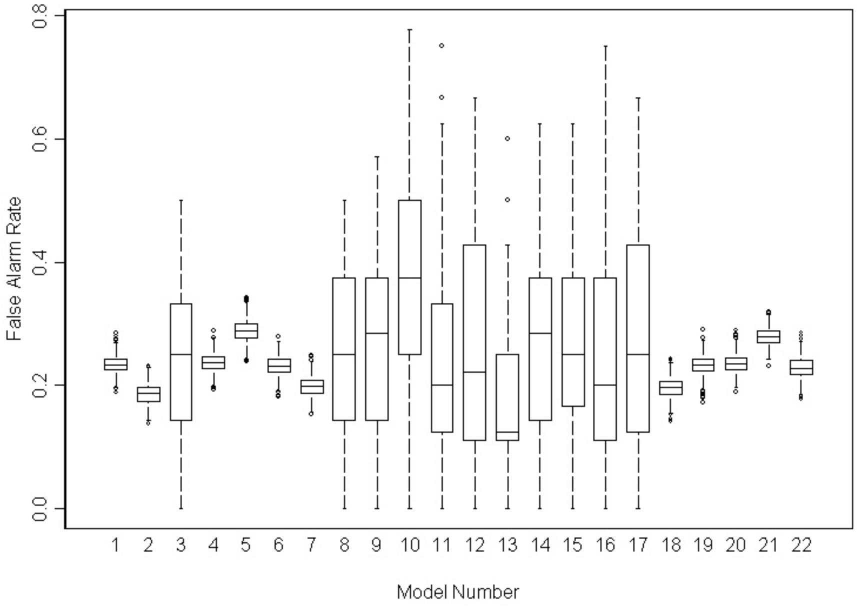


Fig. 45. Same as Fig. 43, but for FAR.

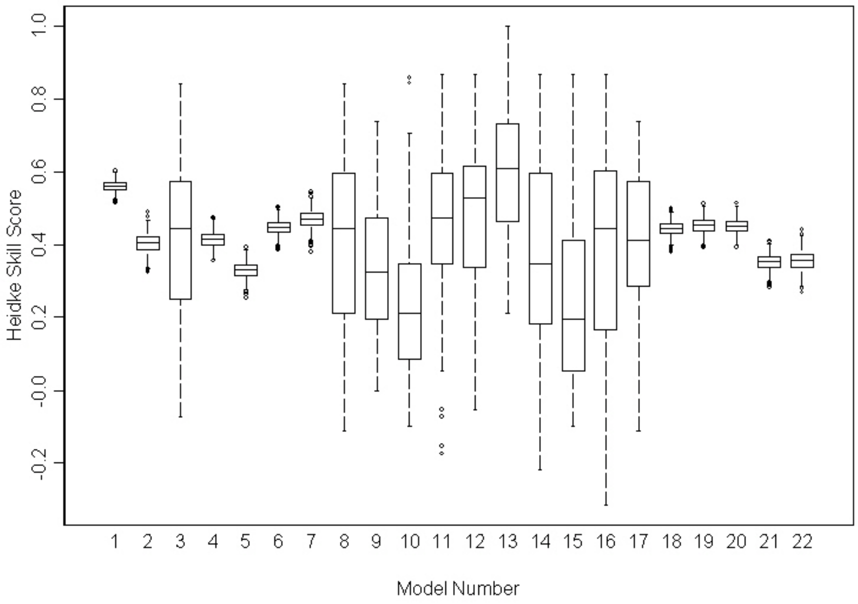


Fig. 46. Same as Fig. 43, but for HSS.

3) LR CONTINGENCY AND CONFIDENCE LIMIT RESULTS

The 72-hour LR jackknife contingency results (Table 20) are similar to the LogR or SVM results. The best HR results are noted when only considering 0-3 km bulk shear (0.735). Rejecting only the LCL produced the highest POD of all three methods (0.844), while the best FAR results (0.284) for LR were obtained when only considering 0-6 km bulk shear. The best 72-hour LR results for HSS were seen when 0-1 km EHI was rejected (0.614). In essence, numerous covariate combinations produced the best contingency statistics for LR.

The boxplot results provided additional insight as to the best covariate combination or combinations for LR. The HR results (Fig. 47) showed the highest median value with model 13, which was consistent with SVM and LogR. This model had large IQR though, so it was not selected as the best covariate combination. Many combinations with low IQR had high median HR and POD (Fig. 48) values, including 7, 18, 19, and 20. The FAR results (Fig. 50) revealed the lowest medians associated with small IQR from models 2, 7, 18, and 22. Model 7 produced superior HSS results (Fig. 50) as well. As a result, the best covariate combination for LR at 72-hours lead time was model 7. The results at 72-hours degraded from those at 48-hours for LR, but the magnitude of degradation was the smallest for LR versus the other two methods (only 10-15%). Thus, as lead time increased, simpler statistical models were able to adjust to the more suspect WRF input.

Table 20. Same as Table 18, but for LR.

Model #	Variable(s)	HR	POD	FAR	HSS	BIAS
1	All	0.697	0.777	0.334	0.395	1.166
2	No LCL	0.690	0.844	0.358	0.382	1.313
3	No bulkshear (0-6 km)	0.700	0.776	0.330	0.400	1.158
4	No bulkshear (0-3 km)	0.704	0.778	0.325	0.410	1.154
5	No SREH (0-1 km)	0.669	0.739	0.355	0.340	1.146
6	No SREH (0-3 km)	0.723	0.790	0.307	0.447	1.140
7	No EHI (0-1 km)	0.741	0.833	0.300	0.482	1.189
8	No Shear	0.703	0.776	0.326	0.407	1.151
9	No SREH	0.685	0.743	0.338	0.370	1.122
10	Only LCL	0.618	0.584	0.378	0.236	0.939
11	Only bulkshear (0-3 km)	0.735	0.774	0.286	0.470	1.084
12	Only bulkshear (0-6 km)	0.733	0.762	0.284	0.466	1.064
13	Only SREH (0-1 km)	0.694	0.777	0.337	0.390	1.171
14	Only SREH (0-3 km)	0.694	0.814	0.346	0.390	1.245
15	Only EHI (0-1 km)	0.611	0.736	0.415	0.223	1.259
16	Only SREH variables	0.727	0.844	0.319	0.454	1.239
17	Only shear variables	0.712	0.729	0.299	0.424	1.041

(a)

Model #	Variable(s)	HR	POD	FAR	HSS	BIAS
18	No SREH (0-3 km)	0.716	0.801	0.319	0.433	1.177
19	No bulkshear (0-3 km)	0.719	0.786	0.311	0.438	1.141
20	No bulkshear (0-6 km)	0.715	0.778	0.313	0.431	1.132
21	No SREH (0-1 km)	0.685	0.743	0.338	0.370	1.122
22	No LCL	0.699	0.815	0.342	0.399	1.238

(b)

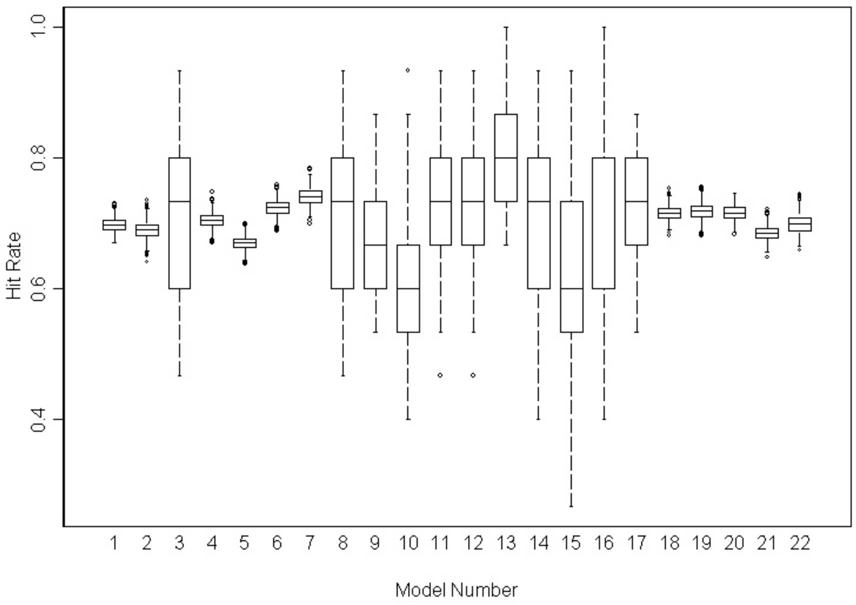


Fig. 47. Same as Fig. 39, but for LR.

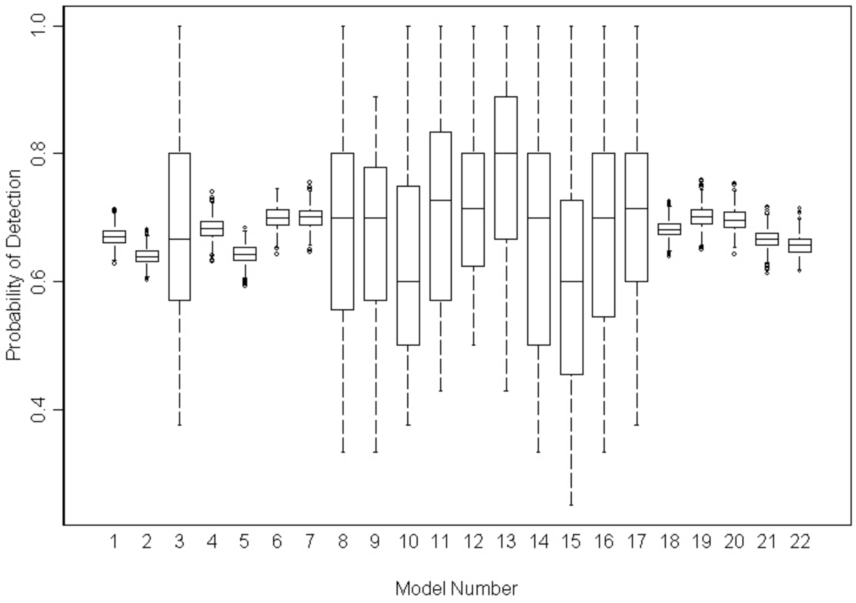


Fig. 48. Same as Fig. 47, but for POD.

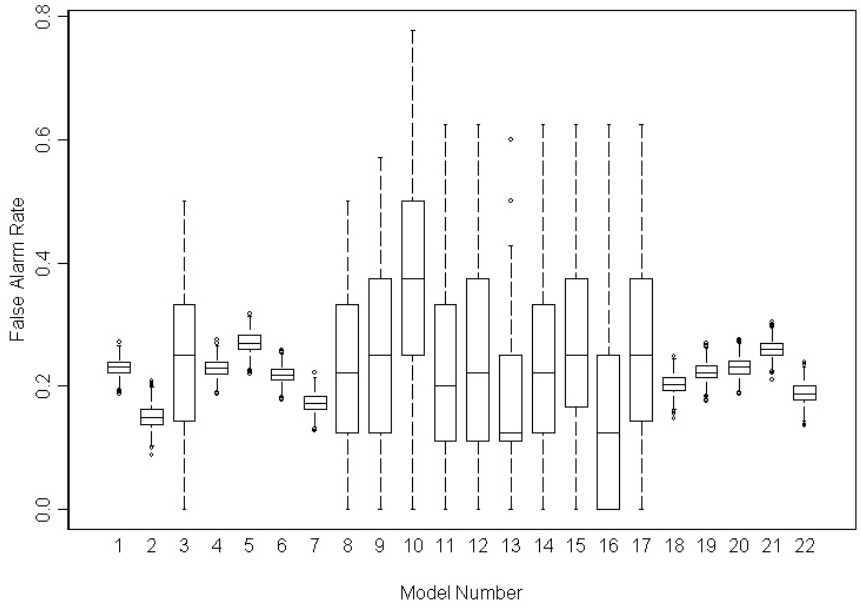


Fig. 49. Same as Fig. 47, but for FAR.

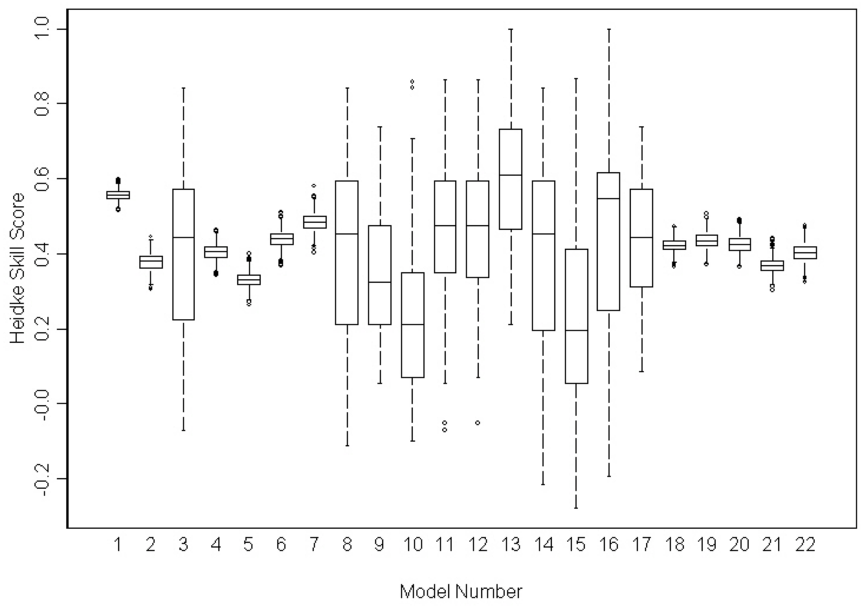


Fig. 50. Same as Fig. 47, but for HSS.

4) SYNTHESIS

In order to determine which statistical technique performed best at 72-hours, the results from the best covariate combination for each method were compared. One optimal set of covariates was obtained for each method, (models 18 for SVM, seven for LogR and LR), which was a good result. The results of the statistical technique comparison (Table 21) show LR as statistically superior to SVM when regarding FAR, indicating that SVM is not the superior method at 72-hours lead time. All other contingency statistics are tied with all three methods, so either LogR or LR is the best method to use for 72-hour classification.

	5% Limit	Median	95% Limit
	HR		
SVM	0.720	0.738	0.755
LogR	0.714	0.734	0.751
LR	0.719	0.740	0.762
	POD		
SVM	0.691	0.718	0.745
LogR	0.677	0.703	0.728
LR	0.673	0.700	0.726
	FAR		
SVM	0.218	0.258	0.299
LogR	0.174	0.199	0.224
LR	0.150	0.174	0.198
	HSS		
SVM	0.416	0.455	0.491
LogR	0.431	0.469	0.507
LR	0.444	0.484	0.525

Table 21. Same as Table 10, but for 72-hours lead time.

5) Case-by-case performance assessment

At 72-hours, SVM only classified seven outbreak cases perfectly (two TOs and five NTOs severe), which is a significant degradation from 48-hours (17 correct) and 24-hours (nine correct). The LogR method discriminated 18 cases with 100% accuracy at 72-hours, consistent with results at 24-hours but significantly worse than results at 48 hours (31 classified with 100% accuracy). The LR results classified 19 cases with 100% accuracy, consistent with 24-hours but a degradation of the 48-hour results. Of the cases classified with 100% accuracy by the three techniques, less than 40% were TOs. This shows that the difficulty of discriminating TOs at 72-hours is larger than that at 48-hours (about a 50-50 spread) and 24-hours (more TOs were classified with 100% accuracy). The bottom 10 cases for each method were mostly TOs as well (over 70%). Clearly, as lead time increases, the ability to discriminate TOs deteriorates. SVM continues its tendency to classify the bottom 10 cases best (8% accuracy versus 2% accuracy for both LogR and LR), although this difference is small. Overall, multiple methods were needed to produce the best results at 72-hours.

Table 23 shows the breakdown of "WRF error" versus statistical model error for the bottom 10 cases of each statistical technique. For this error analysis, "WRF error" continues to be the primary source for classification failure of the statistical techniques,. However, more statistical model error appeared at 72-hours than at 48-hours, implying that statistical classification performance degrades from 48-hours to 72-hours, an implication confirmed by the contingency statistics.

Table 22. Source of errors for the 14 cases that were in the bottom 10 for each statistical technique at 72-hours lead time.

Case	WRF Error?	Statistical Model Error?	Which Technique?
700417		x	SVM
730526	x		LR, LogR
730527	x		LR, LogR, SVM
800712	x		LR, LogR
840607	x		LR, LogR
850531		x	LR, LogR, SVM
890521		x	LR, LogR
900416		x	LR, LogR
920615	x		LR, LogR, SVM
930507	x		LR, LogR, SVM
950527		x	SVM
990408		x	LR, SVM
990503	x		SVM
010409	x		SVM
030506	x		LogR, SVM

4. COMPOSITE RESULTS

To complement the aforementioned objective discrimination results, synoptic storm types of TOs and NTOs were created using the methods described in Chapter 2. A cluster analysis of the principal components loadings revealed two groupings of loadings for each outbreak type, indicating two storm types. Storm type fields were created over a domain encompassing the United States at the 17 NCEP/NCAR reanalysis vertical levels.

In order to determine objectively those regions where the storm types exhibit different features, permutation testing was conducted on the raw case data for the different events within each cluster following the method described in Chapter 2. Fields of *p*-values resulting from the permutation testing were used to assess regions of significant difference in the different raw fields. Examples of the composites are presented to complement the permutation testing discussion. In order to assess the low-level and mid-level differences or similarities between the storm types, gridpoint permutation testing was conducted on the 850 mb and 500 mb height fields.

a) TOs

Two distinct storm types (Fig. 51) resulted from the cluster analysis of the TO loadings at 24-hours lead time. The cluster analyses at other lead times resulted in very similar clusters to those obtained at 24-hours, so 24-hours is presented. The corresponding dendrogram (Fig. 52) shows two main clusters separated by a Euclidian distance that is larger than 0.5, supporting the conclusion of two distinct TO map types.

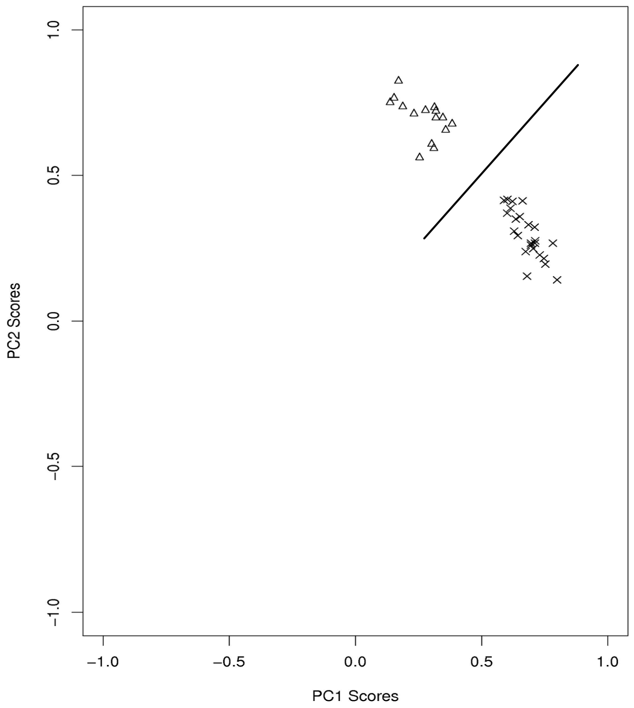


Fig. 51. Scatterplot of PC1 loadings versus PC2 loadings for the 50 TO cases at 24-hours. The triangles represent TO type 1, and the crosses represent type 2. The Fig. illustrates the clustering of the two storm types and the separation between them at 24-hours.

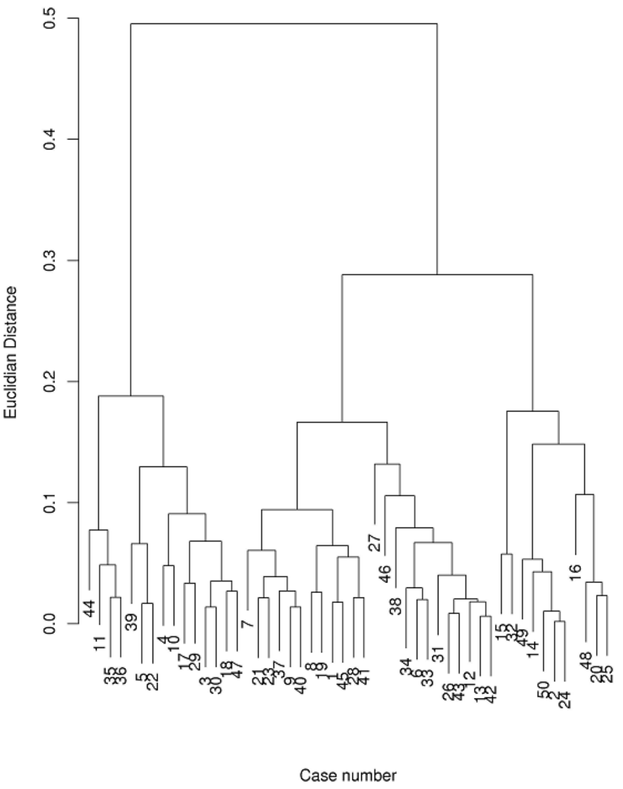
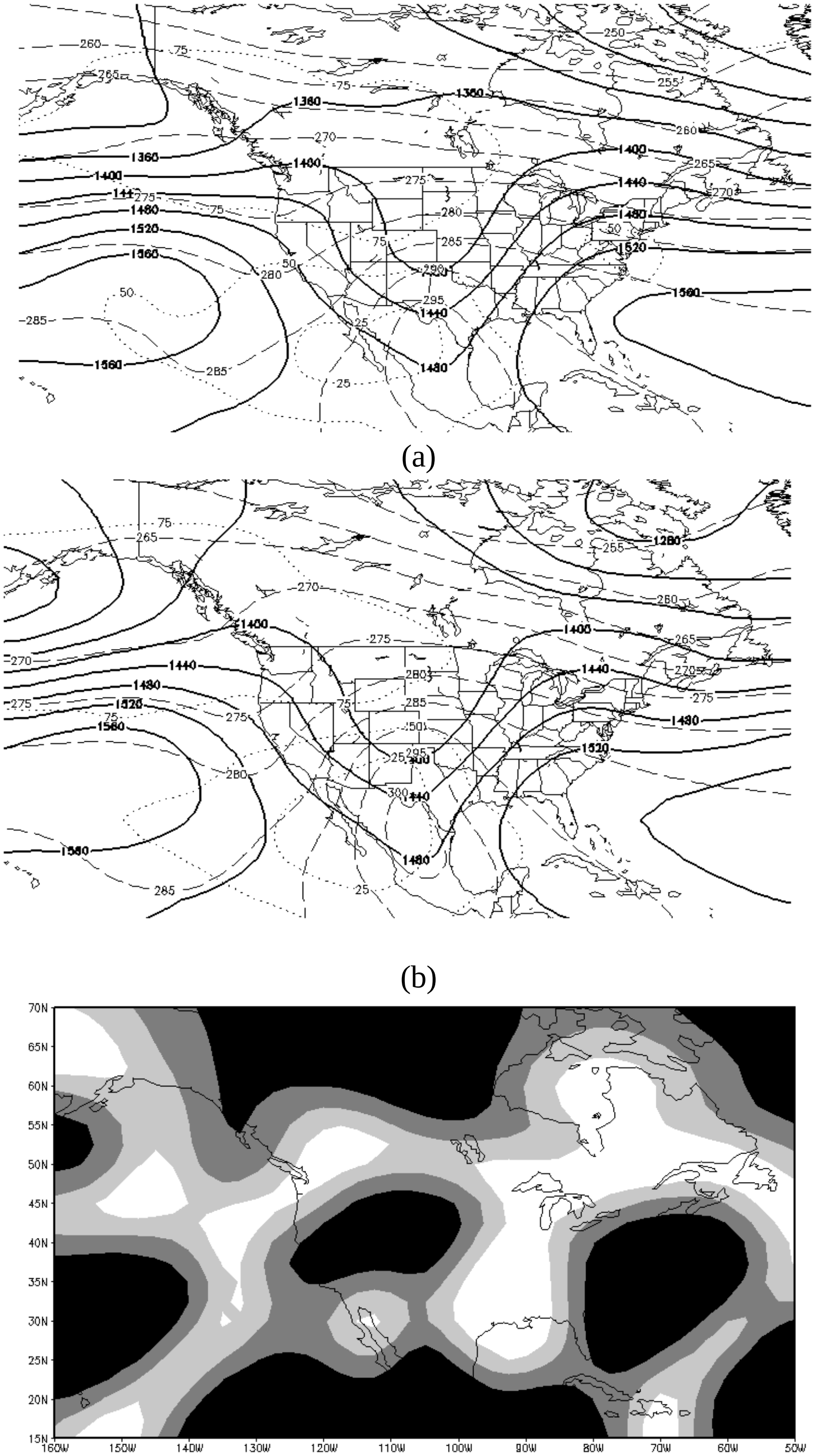


Fig. 52. Dendrogram of Euclidian distance of PC loadings for TOs at 24-hours lead time. The Fig. suggests two main branches spanning from the merge at the 0.5 Euclidian distance level.

The low-level features for both TO types showed large regions of difference significant to a *p*-value of 0.01 (Fig. 53c). The two 850 mb composites (Figs. 53a and b) both show some similar synoptic characteristics (i.e. trough over the western portion of the domain, thermal gradient deforming around the 850 mb trough), although the magnitudes of these characteristics are slightly different. The 850 mb permutation fields suggest that statistically significant differences in the two map types exist in numerous locations. These locations appear similar in a visual inspection in the two map types (and these maps have a correlation of 0.864). Since the fields are similar spatially, magnitude differences must have resulted in the regions of low *p*-values.



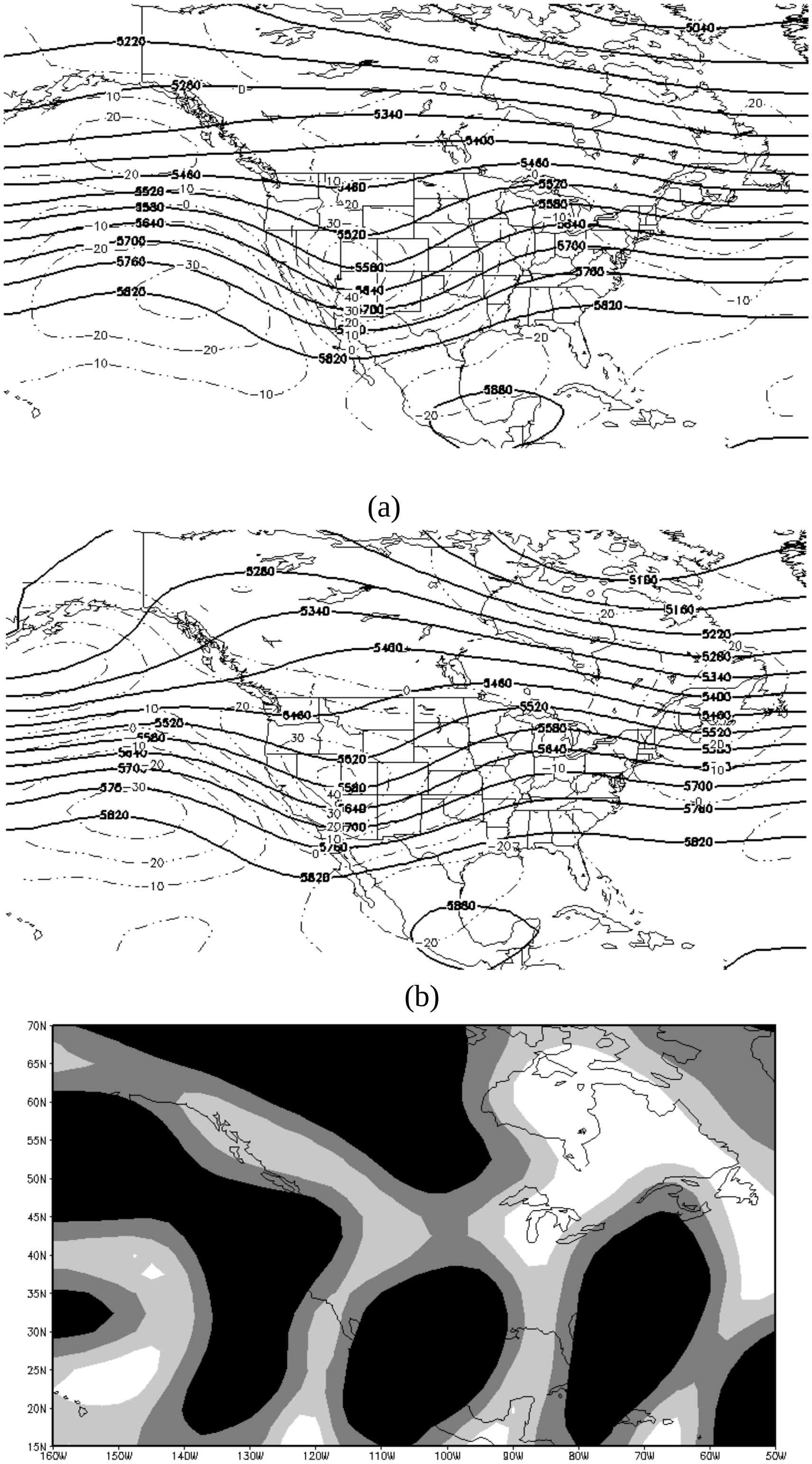
(a)

(b)

(c)

Fig. 53. The 850 mb TO1 and TO2 map types (panels a and b) and the permutation testing results, showing differences between them. In panels a and b, solid lines are height lines, dashed lines are isotherms, and dotted lines are isohumes. In panel c, white areas indicate *p* > 0.1, light gray areas represent *p* < 0.1, dark gray areas represent *p* < 0.05, and black areas represent *p* < 0.01.

The mid-level analyses (Fig. 54) support the results from the low-level analyses, showing some regions of significant difference. As was true at low-levels, the patterns of the 500 mb heights are highly correlated (0.931). Since the patterns are highly correlated and the permutation testing is showing numerous regions of low *p*-values, the differences between the two map types must be in magnitude. Visual inspection (confirmed by computing the difference of the gridpoint magnitudes between the two map types, not shown) of the two map types shows that in the regions that are synoptically active (i.e. near the trough over the western third of the domain, near the ridge in the eastern portion of the domain), the differences suggest a more curved trough-ridge system (lower heights in the trough and higher heights in the ridge). However, these two map types, while significantly different in terms of magnitude, are not easily distinguished by inspection since the patterns appear so similar.



(a)

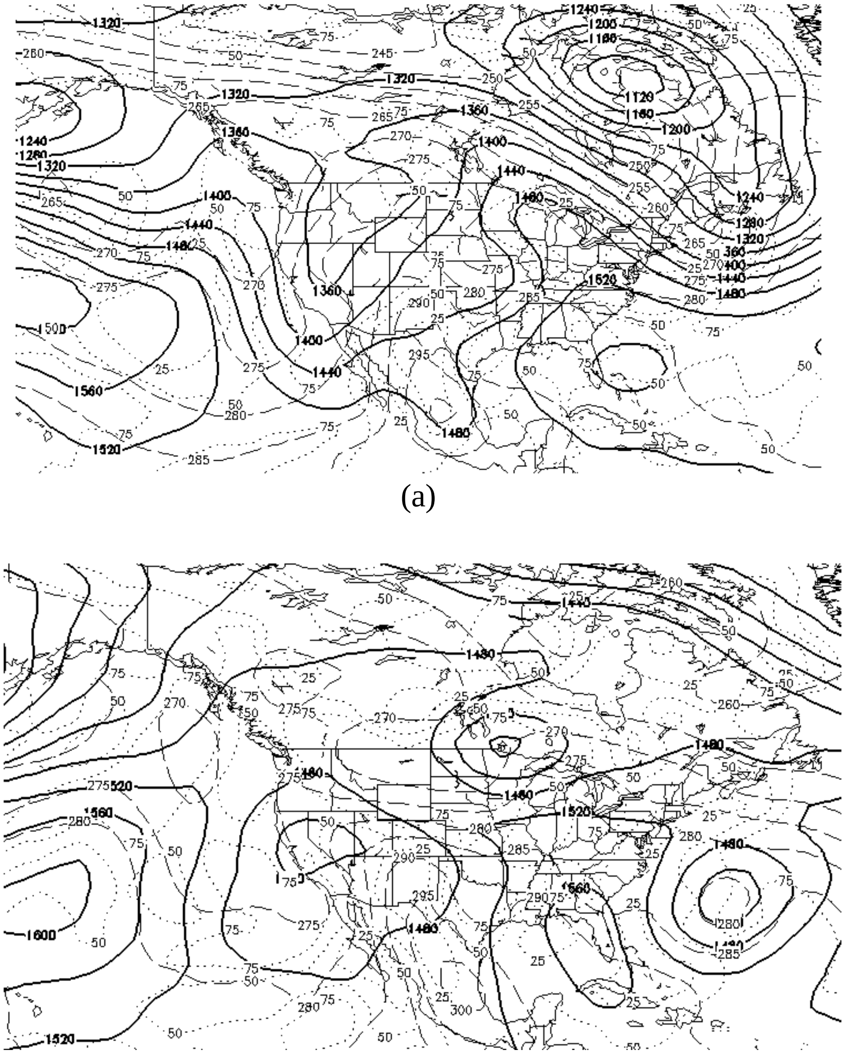
(b)

(c)

Fig. 54. Same as Fig. 53, but for TO2.

In order to assess the similarities and differences between the two TO types further, the two TO cases nearest the cluster dividing line in Fig. 51 (denoted herein as “marginal cases”) and the two cases farthest from the line in the different groups (known herein as “extreme cases”) are considered. Since composite analyses are essentially mean fields, small details unique to each outbreak are damped out by the mean. Thus, the general patterns of these cases are compared to the composites.

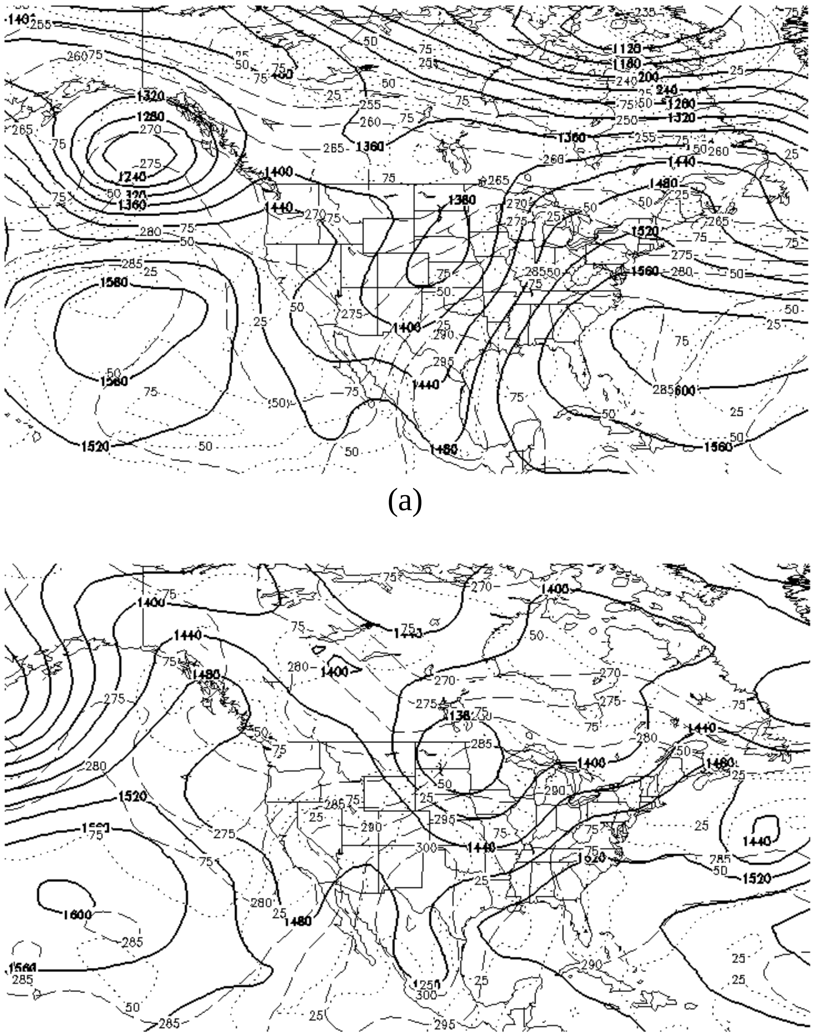
The two marginal events (27 March 1994 for TO1 and 17 April 1970 for TO2 – Fig. 55) have numerous similar synoptic characteristics (i.e. the cyclone over eastern Canada, a weak cyclone in the Great Basin), yet are poorly correlated (0.671). This poor correlation implies that the spatial differences between these two cases are significant, which contrasts the composite results that reveal magnitude differences are the main discrepancy. The TO1 outbreak is much more synoptically dynamic (tighter gradients, stronger low-level flow), which is reversed from what is observed in the composites (i.e. TO2 is slightly more synoptically dynamic in the composites). A forecaster analyzing these two marginal cases might have classified them as either map type, supporting their close Euclidian distance in the cluster analysis.


(a)

(b)

Fig. 55. 850 mb plots from 15 March 1982 (panel a) and 17 April 1970 (panel b), providing the “marginal” TO1 (a) and TO2 (b) cases. Solid lines are isohypses, dashed lines are isotherms, and dotted lines are isohumes.

Numerous visual differences are apparent in the extreme events (1 March 1997 for TO1 and 31 May 1985 for TO2 – Fig 56); yet, they are more highly correlated than the marginal cases (0.787). The higher correlation in these extreme cases indicates that their large Euclidian distance is likely a result of the significant magnitude differences between the two events. The height gradient in the TO1 case over the East that is significantly tighter than the gradient observed in the TO2 event is an example of these magnitude differences. Hence, these two events exhibit the differences which are observed in the composites, an expected result owing to the large Euclidian distance between the events. Overall, the spatial similarities of the two TO types do not allow for an easy discrimination between the two storm types, in spite of the magnitude differences in the height fields that are suggested by the permutation testing and confirmed by the two extreme events.



(a)

(b)

Fig. 56. Same as Fig. 56, but for two extreme events from the cluster analysis. The 1 March 1997 TO (panel a) was used to illustrate an extreme TO1 outbreak, while the 31 May 1985 (panel b) outbreak was used to demonstrate an extreme TO2 outbreak.

b) NTOs

Two NTO types were specified by the cluster analysis (Figs. 57 and 58) as well. Seasonal biases in the NTO case set were resolved by the cluster analysis, as all summer NTOs but two (06 June 1985 and 07 June 1985) grouped into the first NTO type (hereafter NTO1) and all non-summer NTOs but one (11 September 2000) clustered into the second NTO type (hereafter NTO2). The scatterplot of the 24-hour PC loadings (Fig. 57) shows separation (albeit less pronounced than with the 24-hour TO cluster analysis) between the two groups, as was the case with TOs, and are consistent with the dendrogram (Fig. 58) results for NTOs. The Euclidian distance between these two NTO groups (> 0.5) was larger than what was observed with TOs (~0.5).

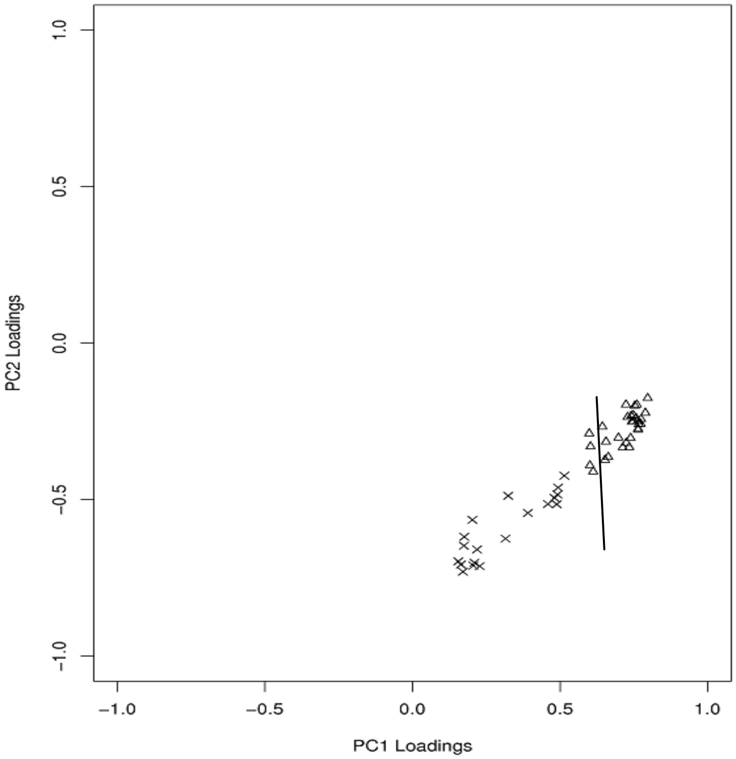


Fig. 57. Same as Fig. 51, but for NTOs at 24-hours.

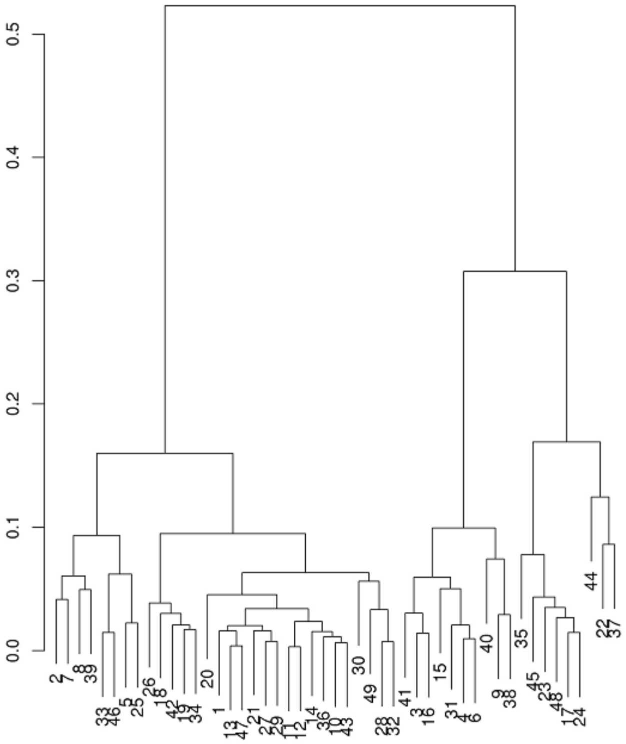
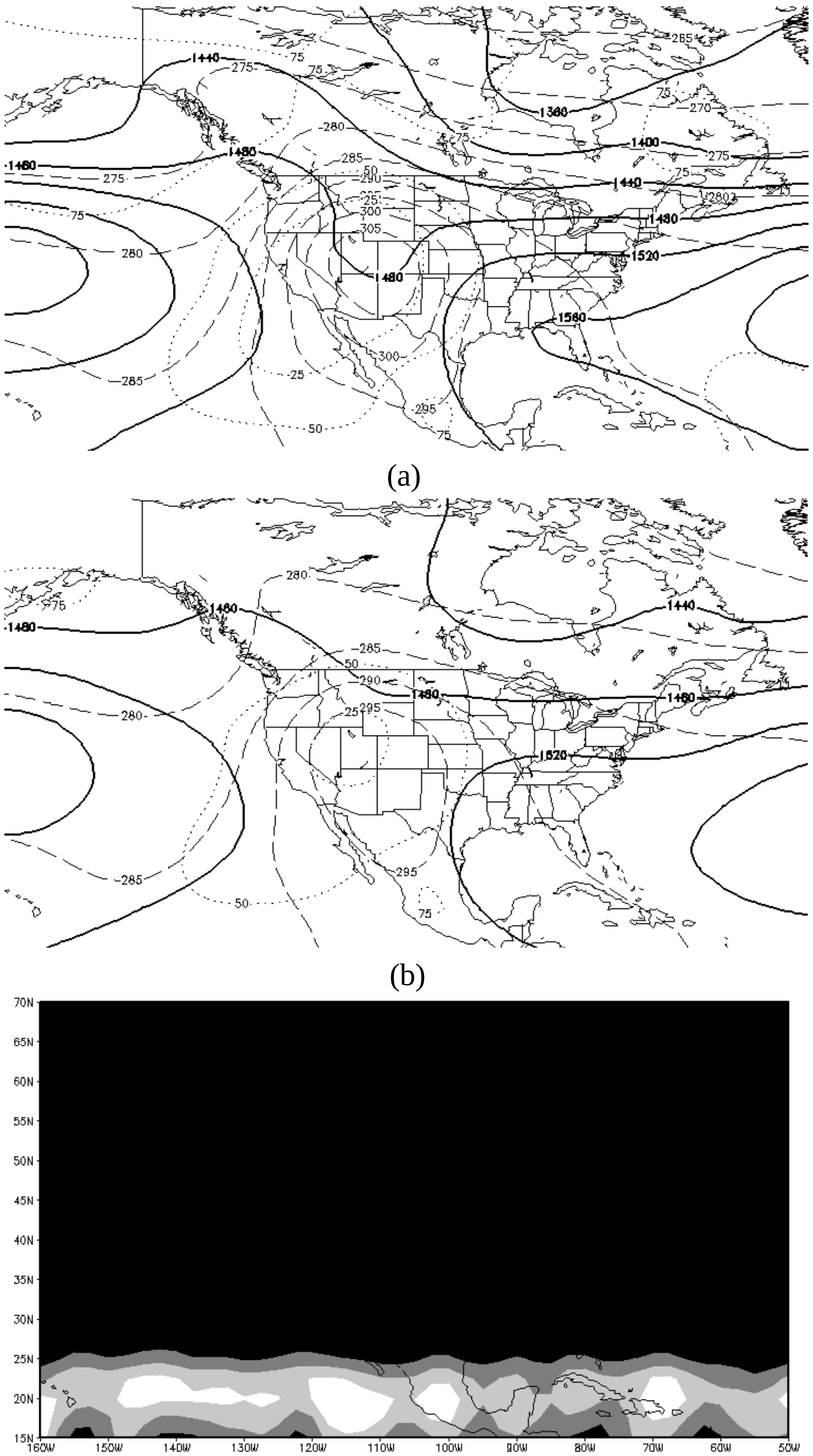


Fig. 58. Same as Fig. 52, but for NTOs at 24-hours.

The two NTO 850 mb map types (Fig. 59) exhibit some similar synoptic characteristics (thermal maximum over the western third of the domain as well as a weak trough, weak synoptic flow throughout the domain in both map types), but their correlation is considerably lower than was observed with TOs (0.538). This result implies some spatial differences between the two NTO map types exists, which is a reasonable assessment from analyzing the composite fields (Figs. 59a and b). The permutation testing results (Fig. 59c) confirm these implications, since *p*-values smaller than 0.01 are present over all but a small portion of the southern quarter of the domain.



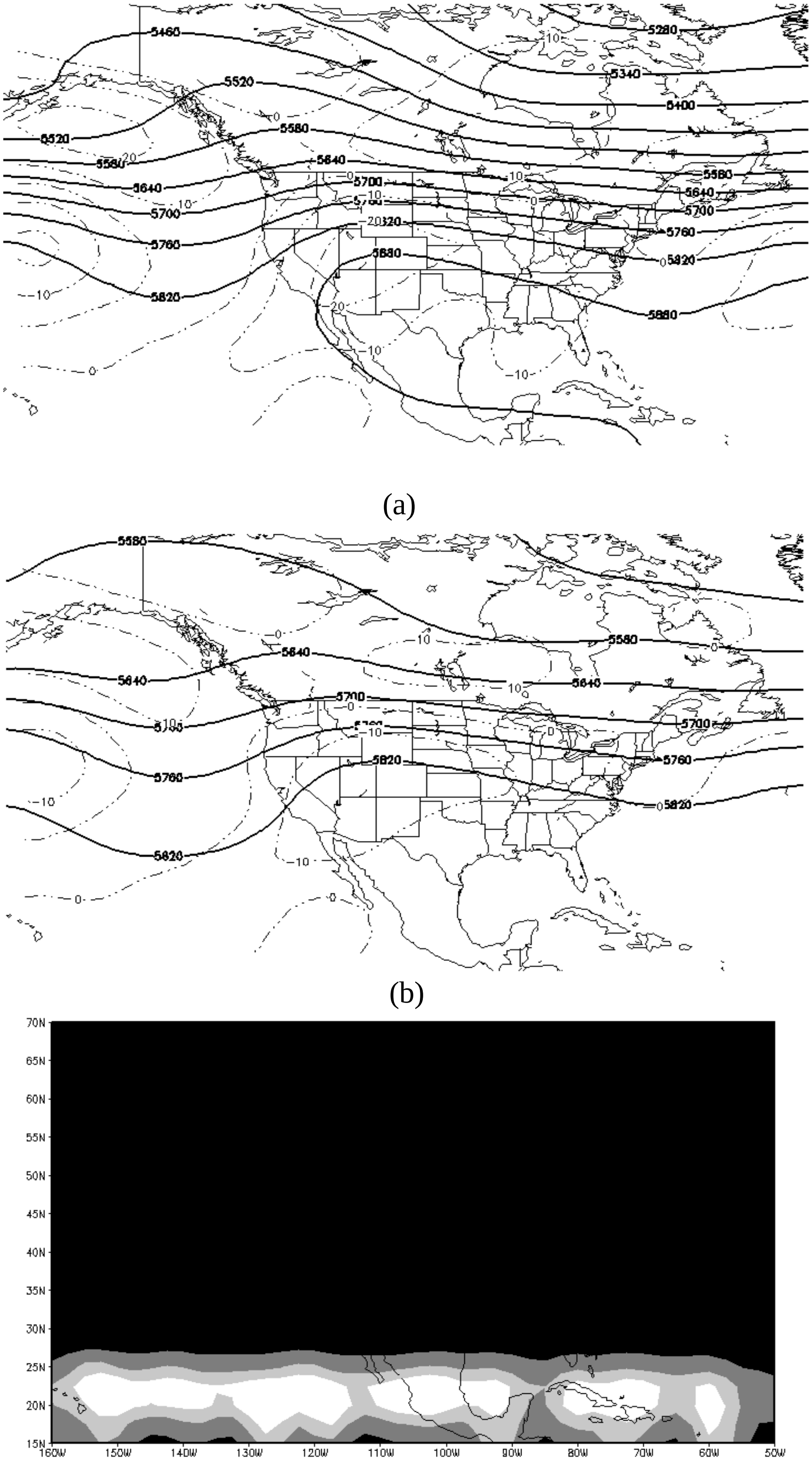
(a)

(b)

(c)

Fig. 59. Same as Fig. 53, but for NTO1.

Mid-level features (Fig. 60) appear similar as well, although the height gradient is slightly tighter over the center of the domain for NTO1. The correlation between these map types is much higher than was observed at 850 mb (0.717) but is still considerably lower than those observed for TOs. Thus, some noticeable spatial differences in the two NTO types should exist and are manifest in the tighter height gradient in NTO1. The permutation testing field (Fig. 60c) appears nearly identical to that at 850 mb, which is expected since the height magnitudes are nearly constant across the southern quarter of both the 850 mb and 500 mb composite domains. These permutation test results and low correlations suggest that some large differences exist between these two storm types, which are apparent in the different height gradients and height magnitudes in NTO1 and NTO2.



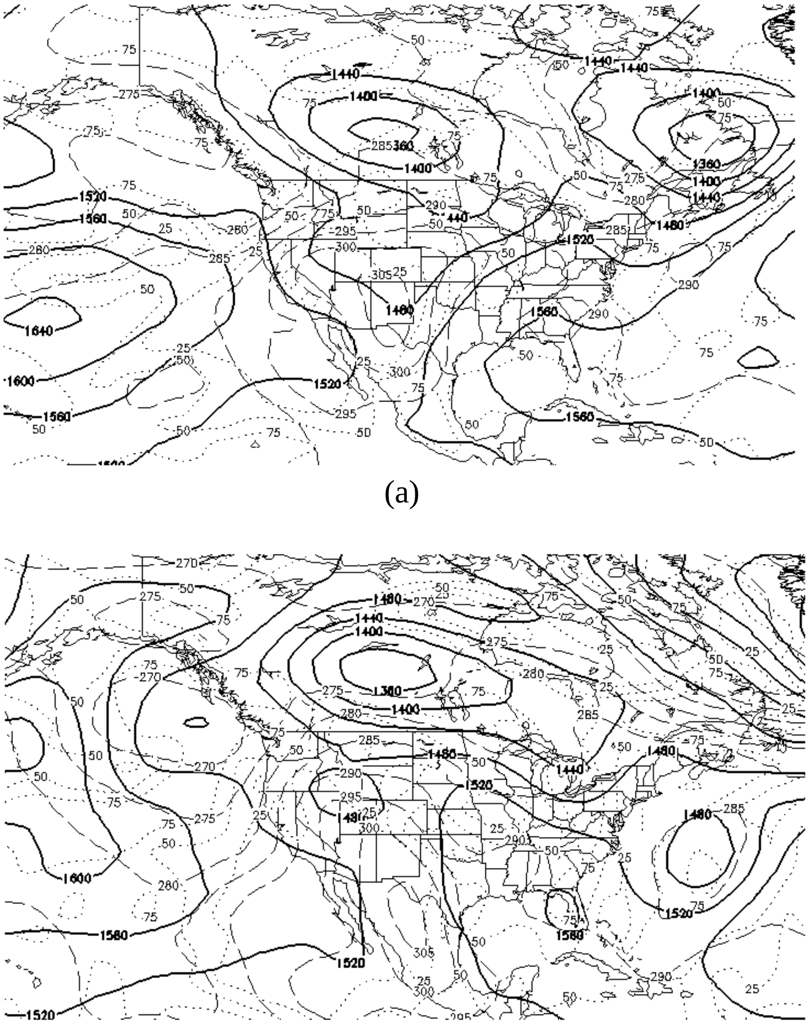
(a)

(b)

(c)

Fig. 60. Same as Fig. 59, but for NTO2.

As done with TO types, NTO cases nearest the separation line in Fig. 57 (marginal cases) and those which are the most distant (extreme cases) are compared. The two marginal cases selected (20 June 1997 for NTO1, 21 May 1989 for NTO2 – Fig. 61) have numerous synoptic similarities (i.e. closed low over central Canada, second low in the eastern third of the domain) and are marginally correlated (0.759). Since both magnitude and spatial differences are present in the two storm types, it is expected that these two marginal cases be similar in magnitude and orientation since their Euclidian distance is so small (0.162), and the correlation results and visual inspection of the fields confirms this expectation.

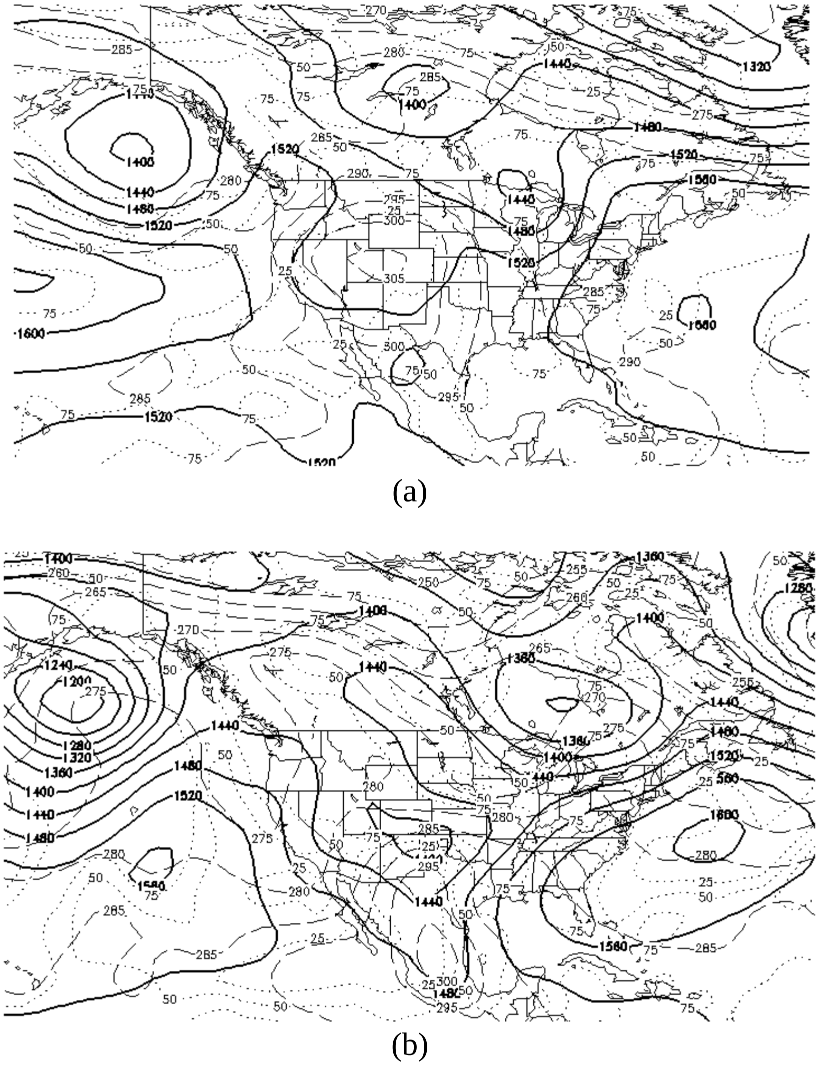


(a)

(b)

Fig. 61. 850 mb fields for the 20 June 1997 NTO (panel a) and the 21 May 1989 outbreak (panel b) over the central Plains. These cases are examples of marginal outbreaks.

The extreme events (6 August 1985 for NTO1 and 10 April 1994 for NTO2 - Fig. 62), which have a large Euclidian distance (0.839), are more strongly correlated than the marginal events (correlation of 0.798). Since the orientations of these two fields are similar, as indicated by their relatively high correlation, magnitude differences (which are present on the fields, note the shortwave over the Great Lakes as an example) must have led to their large Euclidian distance. Thus, as the Euclidian distance between NTO cases increases, magnitudes of the synoptic features in the storm types become more separated, whereas the orientations of the fields change only slightly.



(a)

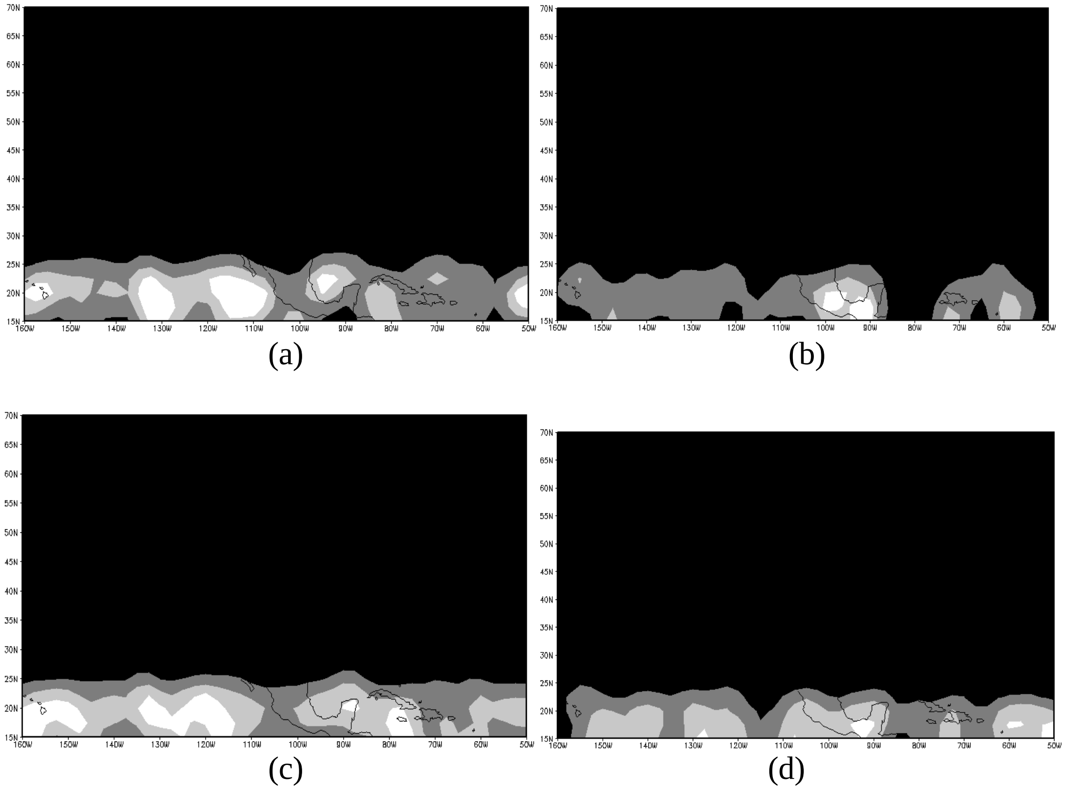
(b)

Fig. 62. 850 mb plots of the 6 August 1985 NTO (panel a) and the 10 April 1994 NTO (panel b) used to illustrate the extreme cases from the NTO cluster analysis.

c) Outbreak Type Synthesis

Since one of the main goals of this study is to distinguish between TOs and NTOs, an inter-comparison of these four map types is needed. Thus, permutation testing on the four possible combinations of the composites (i.e. NTO1 and TO1, NTO1 and TO2, NTO2 and TO1, and NTO2 and TO2 – Fig. 63) was conducted to determine if these composites showed statistically significant differences. It is apparent from this permutation testing that all four combinations have noticeable similarities. Over 90% of the permutation fields are significant to a *p*-value less than 0.01, except for the extreme southern portion of each domain. These non-significant regions represent the tropics, which tend to remain near mean conditions throughout the year (so no composite differences would be apparent in these regions).

The correlations between NTO1 and both TO1 and TO2 are nearly identical (0.743 and 0.754, respectively), which is an expected result since the two TO types are similarly oriented but have different magnitudes. Interestingly, the NTO2 composite has a strong negative correlation with both TO1 and TO2 (-0.827 and -0.835, respectively), which again shows the similar orientation of the two TO types. This negative correlation is not surprising, since the key synoptic features in NTO2 are consistently west of those observed in the TO types, and instead some weak ridging exists in NTO2 where the significant shortwave trough resides in the TO1 and TO2 composites. Overall, similar orientations of the four map types derived from the synoptic storm typing are noted, but large magnitude differences in the different composites lead to statistically significant differences between the TO and NTO composites and allow for their discrimination.



(a) (b)

(c) (d)

Fig. 63. *P*-value plots for the four inter-map type comparisons described above. Panel (a) represents the comparison of TO1 and NTO1, panel (b) represents the comparison of TO2 and NTO1, panel (c) represents the comparison of TO1 and NTO2, and panel (d) represents the comparison of TO2 and NTO2.

5. SUMMARY AND CONCLUSIONS

a) Summary

On a yearly basis, TOs and NTOs affect detrimentally numerous lives and cause extensive property damage each year. Previously, studies have considered individual outbreak types, but no work has objectively investigated the differences between TOs and NTOs in an effort to determine the ability to discriminate between the two types. Thus, a need for advancing our knowledge of outbreak predictability exists, and this need motivates the current work. One key hypothesis of the current study is that the synoptic signal would provide an unknown degree of discrimination ability between outbreak types. To test this hypothesis, synoptic-scale data were used as input for two types of statistical analyses, a statistical objective classification, and synoptic storm typing. Large discrimination ability was achieved by conducting these analyses on the synoptic scale input data, setting a baseline for future work on this topic.

1) OBJECTIVE METHODOLOGY AND RESULTS SUMMARY

Statistical modeling of the two outbreak types was accomplished by simulating the top 100 TO and NTO cases (50 of each outbreak type) from D06 using the WRF model. Three lead times were considered in the objective discrimination analysis (24-, 48-, and 72-hours before outbreak). These lead times were chosen to determine the point prior to an outbreak that objective discrimination significantly worsens. WRF was initialized with the NCEP/NCAR reanalysis data, which were available at a 2.5º by 2.5º latitude-longitude grid spacing and included 17 vertical levels. This dataset was selected owing to its synoptic-scale spacing, since the synoptic-scale signal's ability to distinguish outbreak type was one question being investigated. The WRF simulations used a nested grid approach, with 5 nests being tested (152, 54, 18, 6 and 2 km grid spacing). Seventeen commonly used severe weather parameters or covariates, were computed from the domain 3 WRF output for use in the statistical models. Domain 3 was selected to provide a large number of gridpoints of each covariate for each case. Since the domain 3 output considered thousands of gridpoints, a subdomain of 21 X 21 gridpoints centered on the given outbreak was obtained from domain 3. The data on these subdomains was used in permutation testing, which allowed for the reduction of covariates to those which discriminated TOs and NTOs optimally. The permutation test determines if the means of two data distributions are different; furthermore, it does not require that the initial distributions of the data be known. P-values (the probability that the null hypothesis H_o, which says that the means of the two distributions are the same, should not be rejected) from the permutation testing were computed at each domain gridpoint of the covariate. Low p-values corresponded to larger differences between outbreak types of the particular covariate, which was desirable. After this testing, a smaller subset of covariates (6 or 7) was retained for the statistical modeling.

Since statistical models can be subject to errors due to multiplicity in the data, a method of reduction from the gridded covariate fields to individual variables was accomplished. A PCA was applied to these data, and the subsequent rejection of higher-order eigenvalues thought to be associated with noise led to less than 7 PCs being retained for each PCA. The associated PC scores were used as input into statistical models. Three statistical models were trained and tested using these PC scores, including a linear regression model, a logistic regression model, and a support vector machine. In order to obtain the best set of covariates from the base sets of 6 or 7 (depending on lead time), a backward elimination of covariates was conducted for each statistical model. This backward elimination method improved results, and provided up to 26 covariate combinations for each lead time which were trained and tested with the statistical models. The statistical models yielded classes, either a 0 for a NTO or a 1 for a TO. A contingency table was created from the resulting classes, and contingency statistics were computed from the statistical technique forecasts, allowing the results to be objectively ranked in terms of their ability to discriminate outbreak type. A jackknifing cross-validation procedure was applied in the training and testing of each statistical technique, providing a set of 99 statistical models for each method and 99 sets of results. The 99 result sets were bootstrapped to determine their distribution. Finally, the performance of the statistical methods on each individual case was assessed, providing sets of cases which were discriminated poorly. Reasons for the poor classification of these cases were investigated. Poorly classified cases were classified as subject to “WRF error” (the WRF or atmosphere produced conditions dissimilar from the eventual outbreak) or statistical model error (the environment was conducive for a type of outbreak, but the statistical models classified it incorrectly).

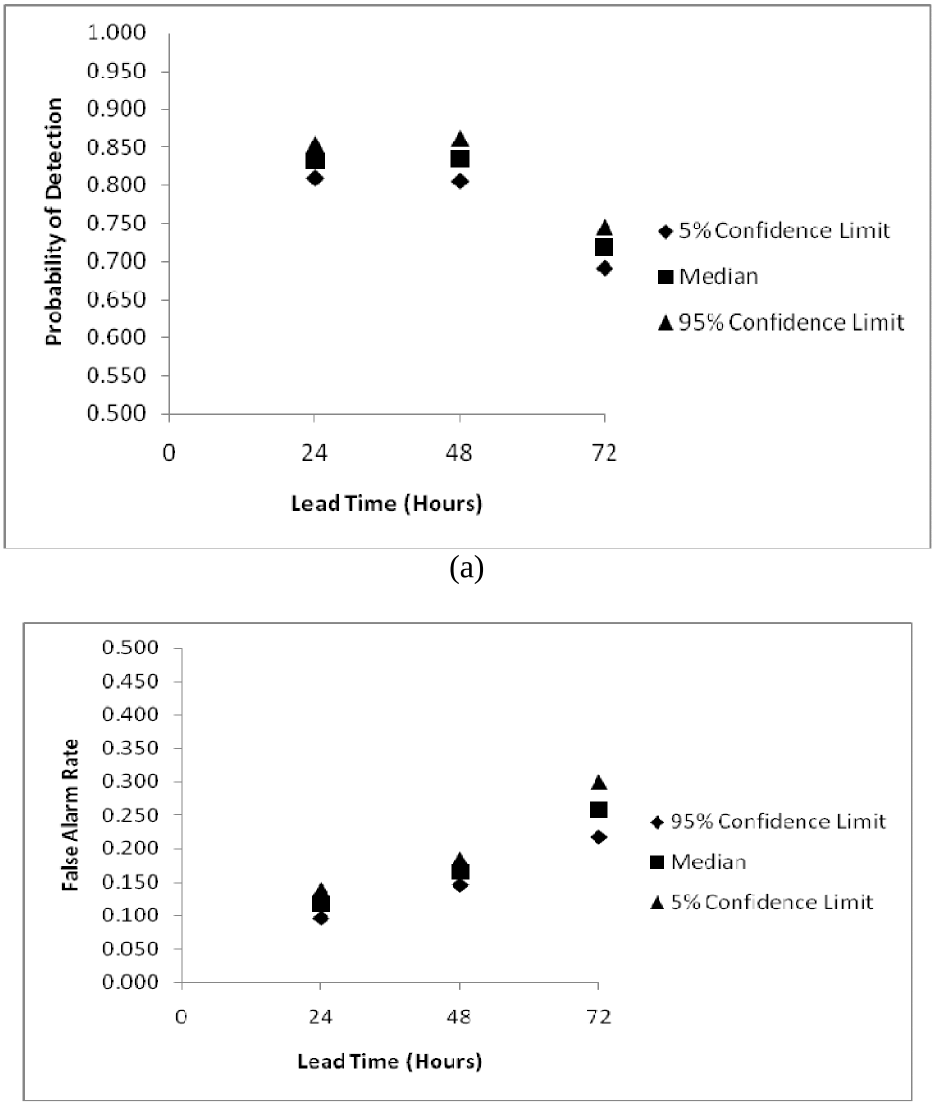
At 24-hours lead time, the SVM results for POD, FAR, and HSS are statistically significantly superior to the other two methods at a 90% confidence. Rejecting two instability covariates, (the product of CAPE and 0-1 km shear and surface based CIN) from the initial set of 7 covariates provided the best classification in all three methods. This result supports the conclusion that instability parameters classify outbreak type poorly (although the computation of the instability parameters may be suspect due to coarse vertical grid spacing from the WRF output in the boundary layer – Table 3).

POD values of 0.9 and FAR values of 0.15 to 0.2 were achieved with SVM, and median values were slightly lower than these for LogR and LR. The boxplot results for this covariate combination support the contingency results, as a small IQR and high median from SVM for the four contingency statistics (hit rate, POD, FAR, and HSS) tested was observed.

The classification evaluation indices at 48-hours did not deteriorate greatly (5-10% - Table 16 in Chapter 3), and the resulting distributions had the largest medians and smallest IQRs for LogR. However, the confidence limits of SVM and LogR were within the 90% confidence limit, so neither method was proven superior. POD values of 0.8 and FAR values of 0.2 to 0.25 were noted from all three statistical techniques, which were significant for a 48-hour lead time classification.

LR results were better at 72-hours than with the other two techniques, with POD's of 0.75 to 0.8 and FAR results of 0.25 to 0.3, which were still significant at 72-hours lead time. Additionally, the FAR from SVM was statistically inferior to the other two methods, so SVM was rejected at 72-hours. The successful outbreak classification observed at 72-hours suggested further lead times should be investigated in future work to determine at which temporal interval the classification ability significantly drops off.

To demonstrate the need for additional time intervals in a future analysis, the POD and FAR performance with lead time for SVM is presented below (Fig. 64). It is apparent that these contingency statistics do not worsen significantly between 24- and 48-hours since the medians remain within the 90% confidence limit. However, by 72-hours, the FAR and POD are both statistically inferior to a 90% confidence with respect to the 24-hour and 48-hour values. Hence, some evidence of a significant drop-off in performance is present by 72-hours, and investigating another 24 to 48 hours prior to the outbreak may provide a more significant drop-off of the contingency statistics. Therefore, further investigation of additional lead times is needed to determine how far in advance the capability to successfully discriminate outbreak type exists.



(a)

(b)

Fig. 64. Median and confidence intervals of POD and FAR with lead time for SVMs. Panel (a) represents POD, and panel (b) represents FAR.

In addition to the contingency analysis, an assessment of the performance of the statistical techniques on each case was conducted. Each case was tested 15 times in the jackknife methodology, and over 20 covariate combinations were considered. Thus, a percentage that each technique classified each case correctly was formulated. At 24 hour leads, LogR and LR classified the most cases with 100% accuracy (17 and 18, respectively), while SVM classified 9 cases with 100% accuracy. At 48-hours, the number of cases classified with 100% accuracy nearly doubled, but by 72-hours the results were similar to those at 24-hours. The increase of 100% accurately classified cases at 48-hours is an unexpected result attributed to the covariate set chosen at 48-hours, which contained no instability measure.

At 24-hours, a larger percentage of the worst classified cases (the bottom 10 performing cases for each statistical technique and each lead time) were classified correctly with SVM (24%) versus LogR (14%) and LR (9%). This result supports SVM as the best technique at 24-hours, since perfectly classified cases are the "classic" outbreak scenarios that forecasters will likely be able to classify correctly. At 48- and 72- hours, no statistical technique correctly classified a significantly larger percentage of these marginal cases. A subjective analysis of the source for the classification errors was conducted, and the marginal cases are classified as either "WRF" error or statistical model error (defined in Chapter 3). Knowledge of the different error types allows for further fine tuning of the statistical models or the WRF to improve results. The "WRF" errors increased with increasing lead time, which was expected since WRF forecasts degrade with increasing lead time. Statistical model error was lowest at 48-hours, a result attributed to the covariate selection at this lead time.

2) COMPOSITING METHODOLOGY AND RESULTS SUMMARY

In addition to the objective statistical discrimination of outbreak type, a compositing methodology was used to reveal physical features of the outbreak types. The composite fields utilized five raw NCEP/NCAR reanalysis meteorological variables (temperature, height, relative humidity, *u* and *v* wind). The composites were created using a PCA. Since the NCEP/NCAR reanalysis data reside on a latitude-longitude grid, converging longitude lines with increased latitude artificially inflated the correlation values at northern latitudes. A Fibonaaci grid, which provides equally spaced gridpoints in the latitudinal and longitudinal directions, was used to eliminate this bias. An O-mode principal component analysis (one in which the correlation matrix is computed along the observation (case) axis) was conducted on the standardized (mean removed) input data matrix.

Once the PCA was complete, a cluster analysis of the resulting PC loadings was conducted for each individual outbreak to determine how the individual cases grouped together. Two main groups resulted from the cluster analysis for both TOs and NTOs. The mean of the PC loadings from these groups was computed and squared, which provided a measure of the percentage of the distribution described by the PC, and these served as weights for the PC scores in the storm type creation. The cluster analyses were visualized using scatterplots of the PC loadings and dendrograms (see Chapter 4).

To assess the differences between the two map types for each outbreak type, gridpoint permutation testing using the same method as Chapter 2 was conducted on the two map types at different vertical levels. The TO composites had several individual regions of statistical significance which corresponded with the individual areas of enhanced synoptic activity (i.e. near troughs and ridges). The correlations between the two TO map types were larger than 0.9, suggesting their orientations were virtually indistinguishable. It was concluded that magnitude differences between the two TO map types in the regions of enhanced synoptic activity led to the low *p*-values in these regions, and these magnitude differences were evident in a visual inspection of the TO composites. Overall, a slightly deeper composite cyclone and more highly curved troughs and ridges were noted in the TO2 type versus the TO1 type.

The two NTO types were not as highly correlated at low-levels (near 0.5) as the two TO types, but at mid-levels the NTOs showed similar orientations (correlation over 0.75). The permutation testing showed the entire domain, except for the tropics in the extreme southern portion of the permutation field, significant to $p < 0.01$. Since moderate correlations were present, these significant differences throughout the entire field were largely attributed to magnitude differences, as were observed in the TO composites. This attribution was confirmed by analyzing marginal and extreme events, both of which were highly correlated, but the marginal cases had similar magnitudes and the extreme cases had vastly different magnitudes. In essence, the increased Euclidian distance between cases resulted in larger magnitude differences without significantly changing the orientations of the height fields.

In comparing TOs to NTOs, the *p*-value fields were significant to 0.01 everywhere but in the tropics for all possible TO and NTO map type combinations. The correlations between an individual NTO type and the two TO types were nearly identical, further supporting the conclusion of similarly oriented TO fields. One interesting finding revealed a highly negative correlation between NTO2 and the two

TO types, which was consistent with the placement of significant synoptic features in these composites (i.e. regions of low heights in TO1 or TO2 were regions of high heights in NTO2). Overall though, the high correlations between all of these composites suggest that large magnitude differences separate the individual map types, and that increased Euclidian distance between cases is a result of increased magnitude difference, not of significantly different orientations of the case height fields. Both of these statistical analyses provided a substantial amount of classification capability between TOs and NTOs, accomplishing the primary goals of this study.

b) Conclusions

The goal of this work was to use strictly objective methods to discriminate TOs from NTOs. One hypothesis tested herein was that the synoptic-scale signal provided information useful in distinguishing outbreak type. Two statistical analyses were used to quantitatively assess these goals, including a statistical objective classification and synoptic storm typing. This study successfully developed methods which were used to discriminate outbreak type objectively, allowing for additional outbreak scenarios to be considered in future work.

As is the case in most research endeavors, new research questions arise from the results which can be addressed in future studies. The seasonal dependence of the results of the compositing suggest the statistical classification results may be artificially inflated (since the summertime NTO is distinctly different from the spring TO and the spring NTO); hence, it is important to remove this seasonal dependence by analyzing spring NTOs when comparing spring TOs. Hence, additional spring cases should be added to the case set to address this problem.

Statistical technique errors observed from the statistical classification results can be improved by further training the statistical models on less ideal cases. Since this work was conducted on the top 50 of each outbreak type, numerous cases which are not distinctly one type or another could fall into a third category of “marginal” outbreaks. Classification of these marginal outbreaks from TOs and NTOs could be attempted in future work, as well as attempting to classify null cases (those in which no outbreak occurs). A larger case set will yield more robust composite fields and objective discrimination results as well, which is important further developing the ability to classify between TOs and NTOs.

Forecast applications of this classification method should be investigated, as these methods can provide powerful tools for forecasters in providing outbreak type classification with a substantial lead time. Some method of converting numerical model anomaly patterns into the composite fields, allowing them to be compared objectively for forecasters, should be considered. The composite fields could be used in a data assimilation package to support the numerical model in simulation of outbreaks. The statistical classification methods should be modified so that they can serve as an operational forecast tool by the Storm Prediction Center (SPC). An algorithm which takes covariates output from a numerical simulation could use the statistical classification methods to warn SPC forecasters of a looming outbreak type. If SVMs are used, their training can be modified to output a probability of a given outbreak type, as opposed to one class or another. This would be useful in supporting the issuing of convective outlooks by the SPC, since it would provide another idea as to the eventual outbreak type of the given day. If knowledge of outbreak type exists with some certainty up to 72-hours in advance, forecasters can warn the public to take steps to prepare for an impending outbreak. Future studies in this area will be challenging, as the current study assumes that an outbreak will occur. When adapting these ideas to a forecast application, it will not be possible to know in advance (besides an educated guess) whether an outbreak will even develop. Additionally, the probability of a tornado outbreak on any given day is very small, so adapting this problem to include null cases will prove to be difficult. Overall, the present study shows that a large ability to discriminate outbreak type exists, and these results can be applied to future studies to improve the overall understanding of these dangerous events, as well as the ability to forecast the outbreaks.

REFERENCES

Araneo, D. C., and R. H. Compagnucci, 2004. Removal of systematic biases in S-mode principal components arising from unequal grid spacing. *J. Climate*, **17**, 394-400.

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

Barnston, A.G., and C.F. Ropelewski, 1992: Prediction of ENSO episodes using canonical correlation analysis. *J. Climate*, **5**, 1316–1345.

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.

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

Bluestein, H. B., 1992. *Synoptic-Dynamic Meteorology in Midlatitudes: Volume I.* Oxford University Press, 431 pp.

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

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, USA, 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: *Support Vector Machines and other kernel-based learning methods.* Cambridge University Press, Cambridge, England, 189 pp.

Daley, R., and R. M. Chervin, 1985: Statistical significance testing in numerical weather prediction. *Mon. Wea. Rev.*, **113**, 814–826.

Davies, J. M., 1993. Hourly helicity, instability, and EHI in forecasting supercell tornadoes. Preprints, *17 Conf. on Severe Local Storms*, St. Louis, MO, Amer. Meteor. Soc., 107-111.

——, J.M., 2004: Estimations of CIN and LFC Associated with Tornadic and Nontornadic Supercells. *Wea. Forecasting*, **19**, 714–726.

——, J.M., 2006: Tornadoes in Environments with Small Helicity and/or High LCL Heights. *Wea. Forecasting*, **21**, 579–594.

Davies-Jones, R., 1984: Streamwise vorticity: the origin of updraft rotation in supercell storms. *J. Atmos. Sci.*, **41**, 2991–3006.

Doswell, C. A., 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.

——, and L .F. Bosart, 2001: Extratropical synoptic-scale processes and severe convection. *Severe Convective Storms, Meteor. Monogr .*, **28**, no. 50, Amer. Meteor. Soc., 27-69.

——, 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.

Dowell, D. C., and H. B. Bluestein, 1997: The Arcadia, Oklahoma, storm of 17 May 1981: analysis of a supercell during tornadogenesis. *Mon. Wea. Rev.*, **125**, 2562–2582.

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.

Edwards, R., S. F. Corfidi, R. L. Thompson, J. S. Evans, J. P. Craven, J. P. Racy, D. W. McCarthy, and M. D. Vescio, 2002: Storm Prediction Center Forecasting Issues Related to the 3 May 1999 Tornado Outbreak. *Wea. Forecasting*, **17**, 544–558.

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

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

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*. 2d ed. Amer. Meteor. Soc., 782 pp.

Grazulis, T. P., 1993: *Significant Tornadoes 1680-1991*. Environmental Films, St. Johnsbury, VT, 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**(**14**), Article 1693.

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

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.

Insightful Corporation 2007: Splus. Version 8.0. University of Oklahoma.

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

Jones, T. A., K. M. McGrath, and J. T. Snow, 2004: Association between NSSL mesocyclone detection algorithm-detected vortices and tornadoes. *Wea. Forecasting*, **19**, 872–890.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woolen, Y. Zhu, A. Leetmaa, B. Reynolds, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, R. Jenne, and D. Joseph, 1996. The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77**, 437-471.

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

Klemp, J. B., and R. Rotunno, 1983: A study of the tornadic region within a supercell thunderstorm. *J. Atmos. Sci.*, **40**, 359–377.

Koch, S.E., D. Hamilton, D. Kramer, and A. Langmaid, 1998: Mesoscale dynamics in the Palm Sunday tornado outbreak. *Mon. Wea. Rev.*, **126**, 2031–2060.

Lanicci, J. M., and T. T. Warner, 1991. A synoptic climatology of the elevated mixed-layer inversion over the southern Great Plains in spring. Part III: Relationship to severe-storms climatology. *Wea. Forecasting,* **6,** 214–226.

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.

Lemon, L.R., and C.A. Doswell, 1979: Severe thunderstorm evolution and mesocyclone structure as related to tornadogenesis. *Mon. Wea. Rev.*, **107**, 1184–1197.

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 Computation*, **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.

Mead, C. M., 1997: The discrimination between tornadic and nontornadic supercell environments: a forecasting challenge in the southern United States. *Wea. Forecasting*, **12**, 379–387.

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.

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, Washington, D. C., 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. Climatology,* **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., and C.A. Doswell III, 1984: Empirical orthogonal function expansion applied to progressive tornado outbreaks. *J. Meteor. Soc. Japan*, **62**, 929-936.

——, 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. Preprint, *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.

Schneider, R. S., J. T. Schaefer, and H. E. Brooks, 2004: Tornado outbreak days: an updated and expanded climatology (1875 – 2003). *Preprints*, 22nd Conf. Severe Local Storms, Hyannis MA. [PDF]

Shafer, C. M., 2007. Evaluation of WRF forecasts of tornadic and non-tornadic outbreaks when initialized with synoptic-scale input. Masters Thesis, University of Oklahoma, 200 pp. [Available from Bizzell Library, University of Oklahoma, Norman, Oklahoma 73019.]

——, A. E. Mercer, C. A. Doswell III, M. B. Richman, and L. M. Leslie, 2008: Evaluation of WRF forecasts of tornadic and nontornadic outbreaks when initialized with synoptic-scale nput. *Mon. Wea. Rev.*: In Press

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

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.

——, and S.J. Weiss, 2002: Mesoscale Model Ensemble Forecasts of the 3 May 1999 Tornado Outbreak. *Wea. Forecasting*, **17**, 526–543.

Swinback, R., and R. J. Purser, 2006. Fibonacci grids: A novel approach to global modeling. *Q. J. R. Meteor. Soc.*, **132**, 1769 - 1793.

Thompson, R.L., 1998: Eta model storm-relative winds associated with tornadic and nontornadic supercells. *Wea. Forecasting*, **13**, 125–137.

——, and M. D. Vescio, 1998. The destructive potential index – A method for comparing tornado days. Preprints, *19th Conference on Severe and Local Storms*, Minneapolis, MN, Amer. Meteor. Soc., 280-282.

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

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

Trafalis, T.B., B. Santosa, and M.B. Richman, 2005: Learning networks for tornado forecasting: A Bayesian perspective, *Sixth International Conference on Data Mining*, 25-27 May 2005, 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.

——, 1996: Statistical significance of long-range “optimal climate normal” temperature and precipitation forecasts. *J. Climate*, **9**, 827–839.

APPENDIX A: CASE LIST

Table A1. List of 50 severe weather outbreaks used in the study. The first column represents the date in yymmdd format, while the second and third columns are the subjectively determined outbreak centers for each outbreak.

Date	Lat	Lon
020615	35.5	-100
931012	33	-99
000721	36.5	-99
870617	38.5	-98.5
810508	33.5	-98
850512	35.5	-98
900416	36	-97
800806	46.5	-96.5
940410	36.5	-96
990522	36.5	-96
890521	37	-96
920704	39	-96
850806	40	-96
020816	43	-96
830829	45.5	-96
950725	37	-95
010414	37.5	-95
860801	36	-94.5
010614	35.5	-93.5
970620	41.5	-93.5
960518	45	-93
820608	39.5	-92.5
000911	41	-91.5
830719	44	-90.5
850704	40.5	-89.5
960505	37.5	-88
820803	44	-88
870705	37.5	-87.5
800702	38	-87.5
810428	40.5	-86
890805	39	-85
860506	41	-85
030707	41.5	-85
980721	42	-85
030502	33.5	-84.5
800708	39.5	-84.5
850607	35	-83.5
800705	40.5	-83.5
020502	36	-83
800712	40.5	-83
850709	41	-83
830704	41	-82.5
850710	36	-81.5
850605	36	-81
010409	40.5	-81
850604	35.5	-80.5
950715	41	-80
800716	40.5	-78.5
891120	41	-75.5
850624	41	-74.5

Table A2. Same as Table A1, but for the tornado outbreaks.

Day	Lat	Lon
700417	35	-102
930507	39.5	-98.5
790410	34	-98
920615	39.5	-98
990503	36	-97.5
910426	37.5	-97.5
940426	35	-97
900313	38.5	-97
730925	40	-96.5
930607	44	-96.5
740608	36	-96
840426	39.5	-95.5
920616	44	-95.5
990504	35	-95
820315	37.5	-95
770504	42	-95
000423	32.5	-94
820402	34	-94
881115	36.5	-94
950527	41.5	-94
840607	43	-94
800407	35	-92.5
990408	40	-92.5
970301	35	-92
990121	35	-92
760329	35.5	-92
711214	36	-92
011123	33.5	-91.5
880508	41.5	-91.5
921121	31.5	-90.5
710221	33.5	-90.5
010224	34	-90.5
030510	40.5	-90.5
730526	36	-90
030504	36	-90
030506	37	-89
960419	38.5	-89
760320	39.5	-87.5
011124	34.5	-87
950518	35.5	-86
980416	36	-86
900602	39	-86
940327	34.5	-85
021110	36	-85
740403	37.5	-85
730527	35	-83
921122	34	-82
850531	41.5	-79.5
840328	34.5	-79
980531	42	-75

APPENDIX B: DESCRIPTION OF THE COVARIATES

B.1 Surface Based Convective Available Potential Energy (CAPE)

A measure of thermodynamic instability is one essential ingredient to severe thunderstorm development (Stensrud et al. 1997, Johns and Doswell 1992, others). Many severe weather studies have used CAPE as a measure of instability, including Brooks et al. (1994) which analyzed CAPE in mesoscale environments associated with severe weather and tornadoes, and Koch et al. which (1998) used CAPE to describe convective instability associated with a Palm Sunday TO in the Southeast. CAPE measures positive buoyancy of air parcels, an indicator of instability, and is defined as:

$$CAPE = g\int_{LFC}^{EL} \frac{\theta(z) - \bar{\theta}(z)}{\bar{\theta}(z)} dz \tag{B.1}$$

where $\theta(z)$ represents the potential temperature as a parcel ascends a moist adiabat, $\bar{\theta}(z)$ represents the environmental potential temperature as a function of height, LFC represents the level of free convection (section B.4), and EL represents the equilibrium level, which is the level at which ascending parcels become negatively buoyant (parcel potential temperature is less than atmospheric potential temperature). The computation of CAPE from the WRF simulation requires vertically stacked gridpoints, and a vertical grid spacing of 31 levels (default for WRF) does not provide a dense vertical grid for its computation. Thus, computations of CAPE may be subject to errors from large vertical grid spacing. Units of CAPE are J kg^{-1}.

B.2 Surface Based Convective Inhibition (CIN)

In most severe weather environments, some measure of low-level stability is available to inhibit the formation of convection. This is known in the literature as CIN (Markowski 2002), and is defined in the AMS glossary (Glickman 2000) as:

$$CIN = -R_d \int_{p_0}^{LFC} (T_{vp} - T_{ve}) d \ln p \qquad \text{(B.2)}$$

where R_d is the gas constant for dry air, T_{vp} is the virtual temperature of the ascending parcel, T_{ve} is the virtual temperature of the environment, LFC represents the level of free convection (section B.4), and p_0 represents the reference pressure where parcel ascension begins. CIN is the negatively buoyant area below the LFC and inhibits convective development. Since CIN is generally confined to the lowest 1-1.5 km of the atmosphere, only a few vertical gridpoints (roughly 10 – Table 3) are used in its computation. This issue leads to artificial errors in the WRF calculation of CIN. CIN has units of J kg $^{-1}$.

B.3 Lifting Condensation Level (LCL)

The lifting condensation level (LCL) is the height at which an air parcel will become saturated if it is lifted dry adiabatically (Glickman 2000). Many studies have related the LCL to the likelihood for tornado development, including Rasmussen and Blanchard (1998), who found that significant tornado development was related to lower LCL values. Thompson et al. (2003) showed statistically significant differences between the mean-layer LCL values from model forecast soundings for different severe weather types. Davies (2006) analyzed a few weak tornado cases in which high LCL values were observed, which contrasts current research on the LCL. The LCL can be determined, using a thermodynamic diagram, as the level at which a dry adiabat originating from the surface temperature intersects a mixing ratio line originating from the surface dewpoint. The units on the LCL are hPa.

B.4 Level of Free Convection (LFC)

According to the AMS glossary (Glickman 2000), the LFC is the level at which a parcel of air, lifted dry adiabatically until it becomes saturated, and moist-adiabatically afterward, will become warmer than the environment. This level is determined on a thermodynamic diagram by following a moist adiabat from the LCL (B.3) until it intersects the temperature profile. Once an air parcel reaches the LFC, it becomes positively buoyant. Davies (2004) found that the threat for significant tornadoes decreased significantly with increased LFC height. He determined that the LFC is more useful for tornado prediction in high CIN environments. Inclusion of this covariate will account for these scenarios. The units on the LFC are hPa.

B.5 Bulk Shear

Another key ingredient required for tornadic supercell development is rotation. One common method which generates rotating flow in the atmosphere is called bulk shear, which the AMS glossary (Glickman 2000) defines as the “local variation of the wind vector or any of its components in a given direction”. For the current study, bulk shear was computed on the *u*-wind component over three commonly analyzed vertical layers, 0-6 km, 0-3 km, and 0-1 km. Tornadoes have been found to be associated with higher values of vertical bulk shear (Dowell and Bluestein 1997, Klemp and Rotunno 1983, others). Colquhoun and Riley (1996) used proximity soundings of tornadic thunderstorms to create mean soundings and hodographs that allowed for analysis of wind shear and thermodynamic parameters as a function of tornado intensity. They found that increased tornado intensity is correlated strongly with higher values of bulk shear. Weisman and Klemp (1984) performed comparisons using numerical simulations of supercell behavior as a function of directionally varying wind shear. They noted that weak shear environments were conducive to short-lived air mass thunderstorms, while high shear environments were better suited for supercell development. The units of bulk shear are m s^{-1}.

B.6 Storm Relative Environmental Helicity

A measure of the streamwise vorticity of the inflow environment of a convective thunderstorm is known as storm relative environmental helicity (SREH). Many studies have observed the relation between high SREH values and the threat for tornadoes, including Kerr and Darkow (1996) which found that deep layer strong SREH was essential in their tornadic supercell model. Colquhoun and Riley (1996) found a correlation of 0.56 between F-scale and SREH magnitude as well. Davies-Jones (1984) provided the theory of tornado development as it relates to streamwise vorticity, stating that a high correlation between vertical velocity and vertical vorticity was present in simulated tornadic supercells. SREH is expressed mathematically in Colquhoun and Riley (1996) as:

$$SREH = \int \omega \bullet (V - V_s)\, dz \tag{B.3}$$

where $\omega = \hat{k} \times d\overset{\omega}{V} / dz$, V represents the wind velocity vector, and V_s is the storm motion velocity vector. For the present study, SREH was computed over the 0-1 km layer and the 0-3 km layer. The units for SREH are m^2 s^{-2}.

B.7 Bulk Richardson Number Shear (BRN Shear)

The Bulk Richardson number (BRN) is used to determine the type of thunderstorms (multicells, supercells, etc.) expected to develop over a region. The BRN is a ratio of CAPE and a measure of the vertical wind shear. It is defined in Stensrud et al. (1997) as:

$$BRN = \frac{CAPE}{0.5(\overline{U}^2 + \overline{V}^2)} \tag{B.4}$$

where CAPE has been defined previously and $\overline{U}$ and $\overline{V}$ represent the density weighted mean wind components over the lowest 6000 m and the lowest 500 m in the atmosphere. For BRN shear, only the denominator of the BRN was considered. Droegemeier et al. (1993) found a high correlation (0.97) between vertical vorticity (a good indicator of storm rotation and possible tornadoes) and BRN shear. Thompson (1998) noted that BRN shear values typically ranged between 25 m^2 s^{-2} and 100 m^2 s^{-2} for tornadic thunderstorms, and only 6% of the tornado events he considered had magnitudes higher than 100 m^2 s^{-2}.

B.8 Storm Relative Flow

In order to include information about storm motion, the storm relative flow at low levels (roughly 2 km) was considered as a covariate. This parameter was the mean supercell motion from the model output, and was defined for the WRF as 75% the magnitude of and 30° to the right of the mean wind vector between 3 km and 10 km above ground. Many studies have analyzed storm motion within tornadic thunderstorms (Lemon and Doswell 1979, Kerr and Darkow 1996, others), but there is a lack of work in using storm relative flow to distinguish between different types of severe weather. The units of storm relative flow are m s^{-1}.

B.9 Energy Helicity Index (EHI)

In addition to considering covariates which contained information on instability or the shear and vorticity profiles of the severe weather environment, covariates which combined these two properties into a single index were considered. One such parameter is the energy helicity index (EHI), which is defined by Davies (1993) as:

$$EHI = CAPE\left(\frac{SREH}{160000}\right) \tag{B.5}$$

where CAPE and SREH have been defined previously. For this work, EHI was computed using SREH at 0-1 km and 0-3 km. Davies (1993) found that for EHI values greater than 1, tornadoes often occurred, and when the value of EHI exceeded 2.5, strong or violent tornadoes were possible. Many more recent severe weather studies examined EHI while observing severe weather environments, including McNulty (1995) who analyzed the use of EHI for tornado forecasting in the central United States, and Mead (1997) who found that for the southern United States, a tornadic supercell environment was characterized by a mean EHI of 3.6, while values less than 2.0 characterized a non-tornadic environment.

B.10 Vorticity Generation Potential (VGP)

Another covariate which considers shear and thermodynamic instability properties of the environment through a single index is known as the vorticity generation potential (VGP). The VGP is defined by the WRF as the total shear from 0 to 3 km above ground level multiplied by the square root of the CAPE of the parcel with maximum equivalent potential temperature below 3 km. According to Rasmussen and Blanchard (1998), VGP is given as:

$$\left(\frac{\partial \zeta}{\partial t}\right)_{tilt} = \eta \bullet \nabla w \qquad \text{(B.6)}$$

where ζ represents the vertical vorticity, *η* represents the horizontal vorticity vector, and *w* represents the vertical velocity. This parameter is intended to give a measure of the conversion of horizontal vorticity to vertical vorticity through tilting, a process thought to be crucial for tornado development. Rasmussen and Blanchard (1998) find significant differences in values in VGP between three different thunderstorm categories (tornadic supercells, non-tornadic supercells, and ordinary thunderstorms). Other studies have used this variable as a measure of the likelihood of tornado formation (Blanchard 1998, others). The units of VGP are s^{-2}.

B.11 Product of CAPE and Bulk Shear

The final covariate considered in this study is one not seen in the literature; the product of CAPE and bulk shear. This parameter is another index that considers values of CAPE and shear simultaneously. This product is considered at 0-1 km, 0-3 km, and 0-6 km.

APPENDIX C: JACKKNIFE SVM S-PLUS CODE

Below is a sample of the code used to run the SVM jackknife methodology.

```
svm.jackknife<-function(dataset,trainratio) {

#  if (length(case1)==ncases)
#      cbind(y,case1)->y
#   if (length(case2)==ncases)
#      cbind(y,case2)->y
#   if (length(case3)==ncases)
#      cbind(y,case3)->y
#
#This function will allow for a bootstrap using SVM.  It will bootstrap the
   different
#cases selected by the sample command with replacement, and do svm's on these.
   It will
#then train the different models and determine output statistics from the
   results.
#The kernel must be changed manually if necessary using the fix svm.bootstrap
   command.
#I may implement further kernel modification later on...
#
#This function is also only valid for classification, as of now regular.  It
#can be modified relatively easily for regression at a later time by using fix
   svm.bootstrap
#
#Determine the dimensions of the input dataset for easier manipulation


#These are required in order to capture which cases are failing
#the most

matrix(scan("case1.txt"),ncol=1,byrow=T)->case1
matrix(scan("case2.txt"),ncol=1,byrow=T)->case2
matrix(scan("case3.txt"),ncol=1,byrow=T)->case3

dim(dataset)[1]->ncases
dim(dataset)[2]->ncolumns
numeric(ncases*(ncolumns-1))->normdata
matrix(normdata,nrow=ncases)->normdata
Normalize(dataset[,1:ncolumns],0,1)->normdata
cbind(normdata,dataset[,ncolumns])->normdata

   if (length(case1)==ncases)
       cbind(case1,normdata)->normdata
    if (length(case2)==ncases)
       cbind(case2,normdata)->normdata
    if (length(case3)==ncases)
       cbind(case3,normdata)->normdata

#Determine the sizes of the training and testing datasets
trainratio*ncases->ntrainrows
as.integer(ntrainrows)->ntrainrows
ntestrows<-ncases-ntrainrows
#
#Declare the data to be manipulated
c(1:ncases)->cases
numeric(ntrainrows)->traincases
numeric(ntestrows)->testcases
numeric(3)->predictions
numeric(ntrainrows*(ncolumns+1))->traindata
numeric(ntestrows*(ncolumns+1))->testdata
predictions<-matrix(predictions,ncol=3,nrow=1)
traindata<-matrix(traindata,ncol=ncolumns+1,nrow=ntrainrows)
testdata<-matrix(testdata,ncol=ncolumns+1,nrow=ntestrows)


#Begin the jackknifing
for (i in 1:ncases) {

   c(i:ncases,1:ncases)->casedata
   casedata[1:ncases]->casedata
   traincases<-casedata[1:ntrainrows]
   testcases<-casedata[ntrainrows+1:ntestrows]

   #Determine the training dataset using the sampled values from above
   for (j in 1:ntrainrows) {
      traindata[j,]<-normdata[traincases[j],]
   }

   #Determine the testing dataset using the sampled values from above
   for (j in 1:ntestrows) {
      testdata[j,]<-normdata[testcases[j],]
   }

   svm(traindata[,2:ncolumns],y=traindata[,ncolumns+1],type="C-
   classification",cost=25000)->svm.model

   predict(svm.model,testdata[,2:ncolumns])->y.hat
#   ifelse(y.hat<0.65,0,1)->y.hat   #For regression combined with
   classification, uncomment this line

   testdata[,ncolumns+1]->y
   cbind(y.hat,y,testdata[,1])->y
   rbind(y,predictions)->predictions



}      # End of the big for loop for the bootstrap
dim(predictions)[1]->removed
predictions[-removed,]->predictions

#table(predictions[,1],predictions[,2])->contingency
#table.stats(contingency)->results

return(predictions)

}  #End of the function
```

APPENDIX D: SAMPLE STORM TYPE S-PLUS CODE

Below is a sample of the code used to create the synoptic storm types for each outbreak type.

```
matrix(scan("tdata_f_06.txt"),ncol=50,byrow=T)->tdata

tdata[1:10591,]->tdata.1
tdata[10592:21182,]->tdata.2
tdata[21183:31773,]->tdata.3
tdata[31774:42364,]->tdata.4
tdata[42365:52955,]->tdata.5

scale(tdata.1[1:623,])->tdata.1.10
scale(tdata.1[624:1246,])->tdata.1.20
scale(tdata.1[1247:1869,])->tdata.1.30
scale(tdata.1[1870:2492,])->tdata.1.50
scale(tdata.1[2493:3115,])->tdata.1.70
scale(tdata.1[3116:3738,])->tdata.1.100
scale(tdata.1[3739:4361,])->tdata.1.150
scale(tdata.1[4362:4984,])->tdata.1.200
scale(tdata.1[4985:5607,])->tdata.1.250
scale(tdata.1[5608:6230,])->tdata.1.300
scale(tdata.1[6231:6853,])->tdata.1.400
scale(tdata.1[6854:7476,])->tdata.1.500
scale(tdata.1[7477:8099,])->tdata.1.600
scale(tdata.1[8100:8722,])->tdata.1.700
scale(tdata.1[8723:9345,])->tdata.1.850
scale(tdata.1[9346:9968,])->tdata.1.925
scale(tdata.1[9969:10591,])->tdata.1.1000

rbind(tdata.1.10,tdata.1.20,tdata.1.30,tdata.1.50,tdata.1.70,tdata.1.100,tdata
   .1.150,tdata.1.200,tdata.1.250,tdata.1.300,tdata.1.400,tdata.1.500,tdata.1
   .600,tdata.1.700,tdata.1.850,tdata.1.925,tdata.1.1000)->tdata.1

scale(tdata.2[1:623,])->tdata.2.10
scale(tdata.2[624:1246,])->tdata.2.20
scale(tdata.2[1247:1869,])->tdata.2.30
scale(tdata.2[1870:2492,])->tdata.2.50
scale(tdata.2[2493:3115,])->tdata.2.70
scale(tdata.2[3116:3738,])->tdata.2.100
scale(tdata.2[3739:4361,])->tdata.2.150
scale(tdata.2[4362:4984,])->tdata.2.200
scale(tdata.2[4985:5607,])->tdata.2.250
scale(tdata.2[5608:6230,])->tdata.2.300
scale(tdata.2[6231:6853,])->tdata.2.400
scale(tdata.2[6854:7476,])->tdata.2.500
scale(tdata.2[7477:8099,])->tdata.2.600
scale(tdata.2[8100:8722,])->tdata.2.700
scale(tdata.2[8723:9345,])->tdata.2.850
scale(tdata.2[9346:9968,])->tdata.2.925
scale(tdata.2[9969:10591,])->tdata.2.1000

rbind(tdata.2.10,tdata.2.20,tdata.2.30,tdata.2.50,tdata.2.70,tdata.2.100,tdata
   .2.150,tdata.2.200,tdata.2.250,tdata.2.300,tdata.2.400,tdata.2.500,tdata.2
   .600,tdata.2.700,tdata.2.850,tdata.2.925,tdata.2.1000)->tdata.2

scale(tdata.3[1:623,])->tdata.3.10
scale(tdata.3[624:1246,])->tdata.3.20
scale(tdata.3[1247:1869,])->tdata.3.30
scale(tdata.3[1870:2492,])->tdata.3.50
scale(tdata.3[2493:3115,])->tdata.3.70
scale(tdata.3[3116:3738,])->tdata.3.100
scale(tdata.3[3739:4361,])->tdata.3.150
scale(tdata.3[4362:4984,])->tdata.3.200
scale(tdata.3[4985:5607,])->tdata.3.250
scale(tdata.3[5608:6230,])->tdata.3.300
scale(tdata.3[6231:6853,])->tdata.3.400
scale(tdata.3[6854:7476,])->tdata.3.500
scale(tdata.3[7477:8099,])->tdata.3.600
scale(tdata.3[8100:8722,])->tdata.3.700
scale(tdata.3[8723:9345,])->tdata.3.850
scale(tdata.3[9346:9968,])->tdata.3.925
scale(tdata.3[9969:10591,])->tdata.3.1000

rbind(tdata.3.10,tdata.3.20,tdata.3.30,tdata.3.50,tdata.3.70,tdata.3.100,tdata
   .3.150,tdata.3.200,tdata.3.250,tdata.3.300,tdata.3.400,tdata.3.500,tdata.3
   .600,tdata.3.700,tdata.3.850,tdata.3.925,tdata.3.1000)->tdata.3

scale(tdata.4[1:623,])->tdata.4.10
scale(tdata.4[624:1246,])->tdata.4.20
scale(tdata.4[1247:1869,])->tdata.4.30
scale(tdata.4[1870:2492,])->tdata.4.50
scale(tdata.4[2493:3115,])->tdata.4.70
scale(tdata.4[3116:3738,])->tdata.4.100
scale(tdata.4[3739:4361,])->tdata.4.150
scale(tdata.4[4362:4984,])->tdata.4.200
scale(tdata.4[4985:5607,])->tdata.4.250
scale(tdata.4[5608:6230,])->tdata.4.300
scale(tdata.4[6231:6853,])->tdata.4.400
scale(tdata.4[6854:7476,])->tdata.4.500
scale(tdata.4[7477:8099,])->tdata.4.600
scale(tdata.4[8100:8722,])->tdata.4.700
scale(tdata.4[8723:9345,])->tdata.4.850
scale(tdata.4[9346:9968,])->tdata.4.925
scale(tdata.4[9969:10591,])->tdata.4.1000

rbind(tdata.4.10,tdata.4.20,tdata.4.30,tdata.4.50,tdata.4.70,tdata.4.100,tdata
   .4.150,tdata.4.200,tdata.4.250,tdata.4.300,tdata.4.400,tdata.4.500,tdata.4
   .600,tdata.4.700,tdata.4.850,tdata.4.925,tdata.4.1000)->tdata.4

scale(tdata.5[1:623,])->tdata.5.10
scale(tdata.5[624:1246,])->tdata.5.20
scale(tdata.5[1247:1869,])->tdata.5.30
scale(tdata.5[1870:2492,])->tdata.5.50
scale(tdata.5[2493:3115,])->tdata.5.70
scale(tdata.5[3116:3738,])->tdata.5.100
scale(tdata.5[3739:4361,])->tdata.5.150
scale(tdata.5[4362:4984,])->tdata.5.200
scale(tdata.5[4985:5607,])->tdata.5.250
scale(tdata.5[5608:6230,])->tdata.5.300
scale(tdata.5[6231:6853,])->tdata.5.400
scale(tdata.5[6854:7476,])->tdata.5.500
scale(tdata.5[7477:8099,])->tdata.5.600
scale(tdata.5[8100:8722,])->tdata.5.700
scale(tdata.5[8723:9345,])->tdata.5.850
scale(tdata.5[9346:9968,])->tdata.5.925
scale(tdata.5[9969:10591,])->tdata.5.1000

rbind(tdata.5.10,tdata.5.20,tdata.5.30,tdata.5.50,tdata.5.70,tdata.5.100,tdata
   .5.150,tdata.5.200,tdata.5.250,tdata.5.300,tdata.5.400,tdata.5.500,tdata.5
   .600,tdata.5.700,tdata.5.850,tdata.5.925,tdata.5.1000)->tdata.5
rbind(tdata.1,tdata.2,tdata.3,tdata.4,tdata.5)->scaled.tdata
ifelse(is.na(scaled.tdata),0,scaled.tdata)->scaled.tdata
```

```
tmode.eigen<-eigen(cor(scaled.tdata))
plot(tmode.eigen$values[1:10])

tmode.load<-tmode.eigen$vectors[,1:2]%*%sqrt(diag(tmode.eigen$values[1:2]))

rotate(tmode.load)->tmode.rot
pc.scores(t(scaled.tdata),tmode.rot$rmat)->tmode.scores
tmode.scores->tmode.scores.group

matrix(scan("tdata_f06_group1.txt"),ncol=1,byrow=T)->group
length(group)->len.group

numeric(len.group * dim(tmode.rot$rmat)[2])->group.mat
matrix(group.mat,ncol=dim(tmode.rot$rmat)[2])->group.mat

for (i in 1:len.group) {
   group.mat[i,]<-tmode.rot$rmat[group[i],]
}

apply(group.mat,2,mean)->group.means
group.means<-group.means^2

for (i in 1:dim(tmode.rot$rmat)[2]) {
   tmode.scores[,i]*group.means[i]->tmode.scores.group[,i]
}

apply(tmode.scores.group,1,sum)->tmode.scores.group


stdev.tdata<-apply(tdata,1,stdev)
mean.tdata<-apply(tdata,1,mean)

tmode.scores.group * stdev.tdata + mean.tdata ->tmode.output.group


tmode.output.group ->finaldata

rm(group)
finaldata.pc1<-finaldata

finaldata.temp.pc1<-finaldata[1:10591]
finaldata.hgt.pc1<-finaldata[10592:21182]
finaldata.rh.pc1<-finaldata[21183:31773]
finaldata.ugrd.pc1<-finaldata[31774:42364]
finaldata.vgrd.pc1<-finaldata[42365:52955]
```