Statistical Verification Results for the

Collaborative Convective Forecast Product

Jennifer Luppens Mahoney[1], Barbara G. Brown[2], and Joan Hart[3]

March 2000

[1]Forecast Systems Laboratory, Office of Oceanic and Atmospheric Research, National Oceanic and Atmospheric Administration, 325 Broadway R/FS5, Boulder, CO 80303.

[2]Research Applications Program, National Center for Atmospheric Research, Boulder, CO.

[3]Joint collaboration with Cooperative Institute for Research in the Environmental Sciences, University of Colorado at Boulder, Boulder, CO.

Statistical Verification Results for the Collaborative Convective Forecast Product

Jennifer Luppens Mahoney, Barbara G. Brown, Joan Hart

Abstract. This report summarizes the verification results for the Collaborative Convective Forecast Product (CCFP) that were collected during the 1999 summer Convective Intercomparison Exercise. The forecasts included in the study were those used in the collaborative process as well as the operational products produced by the Aviation Weather Center (AWC). The exercise took place from 1 June through 31 August 1999. The evaluation was funded by the Federal Aviation Administration (FAA) Aviation Weather Research Program (AWRP).

The CCFP forecasts were issued at 1500 and 1900 UTC with valid times of 1600, 1800, 2000, and 2200, 0000, and 0200 UTC, respectively. The forecasts were verified using individual lightning and radar observations, as well as a convective field in which both lightning and radar observations were combined. The forecasts were evaluated as Yes/No forecasts of convection.

This study covers a near real-time evaluation of the forecasts generated by the Real-Time Verification System (RTVS) developed by the Forecast Systems Laboratory. A web-based interface (http://www-ad.fsl.noaa.gov/afra/rtvs/RTVS-project_des.html), including contingency tables of statistical results, time series and scatterplots, and graphical displays, was developed to provide an efficient and easy way for users to access the results in near real time.

Results of the exercise suggest that forecasting convection is difficult. However, forecasts improve when convection is associated with long-lived convective cells. The CCFP discriminates well between convective and nonconvective areas. However, the properties of the false alarms for the convective areas are large. In comparison with the convective SIGMET Outlooks, the CCFP convective area is smaller, at least partly due to the shorter valid period.

Plans are underway to continue this convective exercise through the summer of 2000. The exercise will again intercompare the CCFP with various convective forecasts. RTVS will generate statistical displays and provide output on the World Wide Web. The verification methods will be enhanced to allow a more thorough evaluation of coverage and probability forecasts.

1. Introduction

Each summer, convective weather is responsible for numerous air traffic delays, reroutes, and cancellations. In an attempt to mitigate this disruption, an experimental collaborative convective weather forecasting process was developed during the summer of 1999. This process allowed meteorologists at the Aviation Weather Center (AWC), Center Weather Service Unit (CWSU), and participating airlines to work together to produce a convective forecast product that is different from any other currently being produced in the National Weather Service. This forecast, known as the Collaborative Convective Forecast Product (CCFP), was provided to the Traffic Flow Management Unit (TFM) from May through August 1999 to be used as a decision-making tool when rerouting airline traffic.

During this 4-month period, the quality of the CCFP was evaluated both objectively and subjectively. This paper summarizes only the objective verification results produced for the CCFP during a convective forecast intercomparison exercise; the subjective results are described by Phaneuf and Nestoros (1999). The objective evaluation was funded by the Federal Aviation Administration (FAA) Aviation Weather Research Program (AWRP).

The goals of the evaluation were to 1) assess the quality of the CCFP and provide objective feedback to the decision-makers, forecasters, and administrators; 2) demonstrate progress so far in the development of the CCFP; and 3) examine the strengths and weaknesses of this experimental forecast. To accomplish these goals, first, statistical results were generated in near real time for a 3-month period, with Web-based displays provided to users. Second, several different convective forecasts were included in the evaluation to provide a baseline for measuring success. These products include the operational forecasts known as Convective Significant Meteorological Advisory (c-SIGMETs) Outlooks; c-SIGMETs produced by the AWC; and the experimental First Guess (FG) forecast, which is produced as guidance prior to formulation of the CCFP. Third, the analyses presented here will help determine the strengths and weaknesses of the CCFP.

The report presents the evaluation approach in Section 2, forecast products in Section 3, data in Section 4, verification methodology in Section 5, results in Section 6, and summary and conclusions in Section 7.

2. Evaluation Approach

Four forecast products were included in this evaluation: the CCFP, the FG forecast, the c-SIGMET Outlook, and the c-SIGMET. Each forecast was provided by AWC in real time to the Real-Time Verification System (RTVS; Mahoney et al. 1997). The CCFP was issued at 1500 and 1900 UTC with valid times of 1600, 1800, and 2000 UTC

and 2200, 0000, and 0200 UTC, respectively. The FG forecast was issued at 1400 and 1715 UTC with valid times of 1600, 1800, and 2000 UTC and 2200, 0000, and 0200 UTC, respectively. The matching valid times for the c-SIGMET Outlooks and c-SIGMETs were used for the comparison. All four forecasts were produced as text products that were translated into latitude/longitude points, producing polygons of convective regions (Fig. 1).

The collaborative forecasting process started in May with forecasters generating the CCFP and the FG forecasts each day. During the first month, however, significant changes to the forecast lead and issue times as well as the forecast domain were introduced. As a result, the objective verification process using RTVS started in June and continued for three months, ending 31 August 1999.

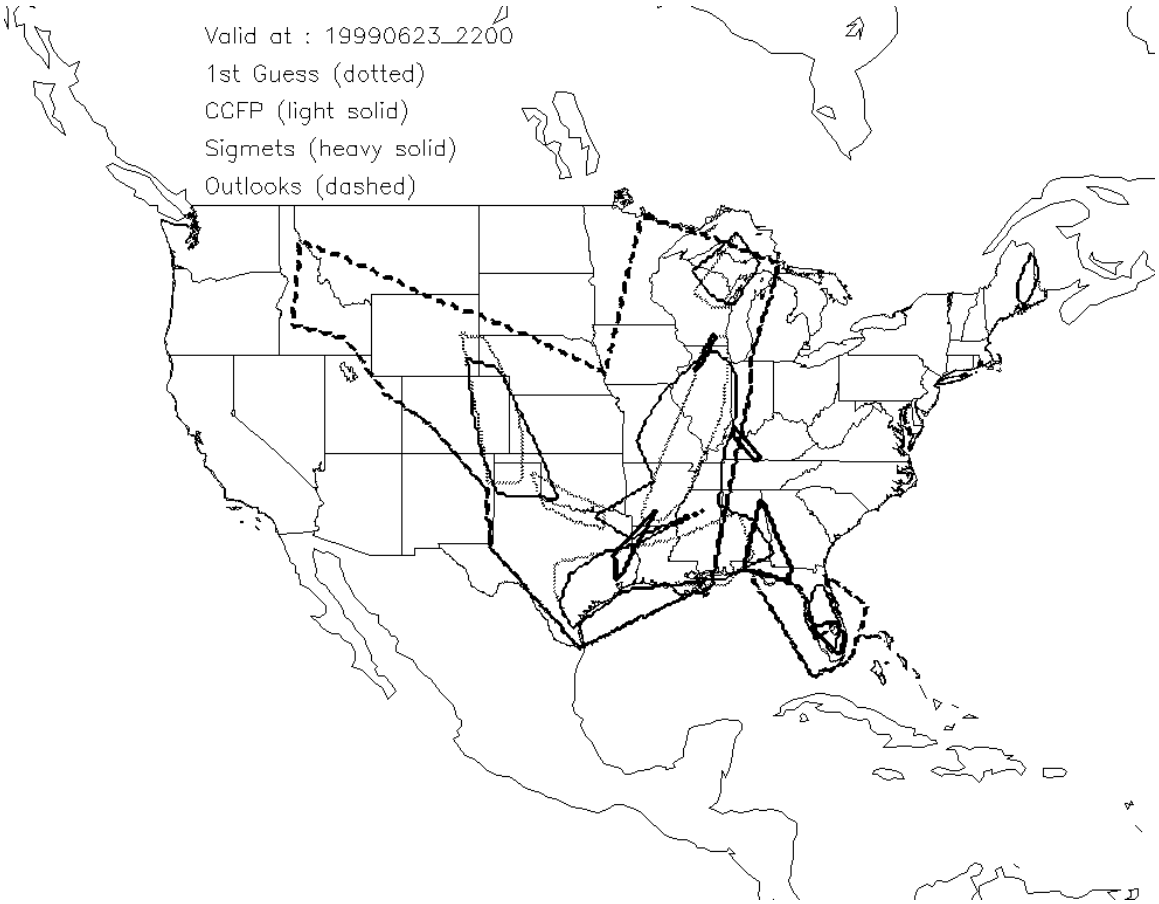

Figure 1. Example of convective forecast polygons for the FG forecast (dotted), CCFP forecast (light solid), the c-SIGMETs (heavy solid), and the c-SIGMET Outlooks (dashed).

The forecasts were verified using individual lightning and radar observations with specific thresholds set to indicate convection. A third type of observation used to assess the convective forecasts was the National Convective Weather Detection Product (NCWDP, also labeled as the NCDP; Mueller et al. 1999) developed by the FAA Convective Weather Product Development Team (PDT; Sankey et al. 1997). Radar and lightning observations are combined to produce this detection field.

A web-based interface[1] containing contingency tables of statistical results, time series and scatterplots, and graphical displays was developed to provide efficient and user-friendly access the results in near real time.

3. Forecast Products

The forecast products that were included in the evaluation are briefly described here.

Collaborative Convective Forecast Product (CCFP): This experimental forecast was generated from input provided by participating airline, CWSU and AWC meteorologists and the staff at the FAA Air Traffic Control System Command Center. The products were generated as a graphic depicting forecasts of convective activity valid at specific valid times. The forecast product was ultimately used by TFM decision-makers for routing traffic around convective areas (FAA 1999). Forecasts were issued at 1500 and 1900 UTC with 1-, 3-, and 5-h and 3-, 5-, 7-h lead times, respectively.

First Guess Forecast (FG): The FG forecast (FAA 1999) was generated by AWC meteorologists as a precursor to the CCFP. The forecast was made available to the participants (listed above) who evaluated the forecast and provided feedback that was ultimately incorporated into the CCFP. Forecasts were issued at 1400 and 1715 UTC with 2-, 4-, and 6-h and 5-, 7-, and 9-h lead times, respectively.

Convective SIGMET (c-SIGMET): This product is an operational forecast of convective activity that is generated at the AWC. The forecast is produced hourly and is valid for up to 2 h (NWS 1991). We assumed a 1-h forecast length valid at the end of the forecast period.

Convective SIGMET Outlook (c-SIGMET Outlook): The convective outlook is an operational forecast of convective activity, generated by AWC meteorologists, issued hourly, and valid from 2-6 h after the issuance time of the c-SIGMET (NWS 1991). The size of the forecast area encompasses moving and changing weather over the 4-h period. For this exercise, the outlooks were evaluated in two ways: 1) as a forecast of length 6h, valid at the end of the period (referred to as the 6-h Outlook) and 2) as a forecast length of 4 h, valid throughout the 2-to-6-h period after issuance (referred to as the 4-h Outlook).

[1.] http:www-ad.fsl.noaa.gov/afra/rtvs/RTVS-project_des.html; select Convective Intercomparison Exercise: June - August 1999)

4. Data

Data that were used in the evaluation include convective forecasts, lightning reports, radar data, and the NCWDP. These data were obtained and used in real time by the RTVS and archived for future analysis and comparison capabilities.

Convective forecasts were obtained in near real time from the AWC. The CCFP, FG forecasts, c-SIGMETs, and c-SIGMET Outlooks are text forecasts that are decoded into area forecasts bounded by latitude and longitude vertices. The CCFP forecast was generated from input provided by numerous meteorological groups within the aviation community, while the remaining three forecasts were generated by individual forecasters at AWC. Forecasters used any available observational data to generate the forecasts. However, as a result of offset valid and issue times, data used to make a forecast were not used to verify the same forecast.

Lightning data were obtained from the National Lightning Data Network (NLDN; Orville 1991). These data were available every 1 h with the locations of specific lightning strikes identified using latitude and longitude. The lightning observations are used alone as well as in combination with radar data to infer areas of active convection for verification of the forecasts. Radar reflectivity (dBZ) fields were available on a 4-km grid scale and were also used as one type of observed convective field.

The NCWDP combines a 2-dimensional radar mosaic of VIL (Vertically Integrated Liquid) with radar cloud top data and a grid of lightning detections from the National Lightning Detection Network. Cloud top data are primarily used to remove anomalous propagation and ground clutter. Lightning data help to keep the NCWDP current, since lightning data have a lower latency than radar data. The data were made available on the 4-km grid scale with a threshold of 40 dBZ or more than 3 lightning strokes in 10 minutes used to delineate storms.

5. Methodology

This section describes the various methods used to match the forecasts and observations, the statistical verification measures computed to evaluate the convective forecasts, and some criteria used to stratify the forecasts.

5.1. Matching Methods

Before forecasts were matched to observations, a 20-km grid was laid over the observation field. Each box on the overlay grid was assigned a *Yes* or *No* value depending on whether a positive observation fell within the 20-km box. For each 20-km box, the criteria used in this study to define a positive observation for each type of

verification observation included: 1) 4 strikes of lightning in the 20-km box, 2) one 4-km box of radar reflectivity greater than 40 dBZ that fell in the 20-km box, and 3) one 4-km box of NCWDP with a dBZ greater than 40 that fell in the 20-km box. The same procedure was performed for the forecasts, with a 20-km box labeled with a *Yes* forecast when any part of the forecast polygon intersected that box. If a forecast polygon did not intersect the 20-km box, then a *No* forecast was assigned to that box.

For some analyses, a filter was applied to the NCWDP observations in an attempt to screen out isolated short-lived convection. In this case, a 20-km box was assigned a *Yes* observation only when 12 or more 4-km NCWDP boxes meeting the 40 dBZ and greater criteria were activated. Otherwise, a *No* observation was assigned to the 20-km box.

Once this process was complete, each box on the 20-km observation grid was matched to each 20-km box on the forecast grid. This technique produced the forecast/observation pairs used to generate the verification statistics. For example, a *Yes* forecast box and a *Yes* observation box would produce a *Yes-Yes* pair. Similarly, a *Yes* forecast and *No* observation would produce a *Yes-No* pair, and so on, filling in the four cells of the statistical contingency table (described further in Section 5.2).

The forecasting domain defined for the evaluation extends west from the Atlantic Ocean to a north-south line east of Denver, Colorado (Fig. 2). Only statistics computed on this domain are presented in this report.

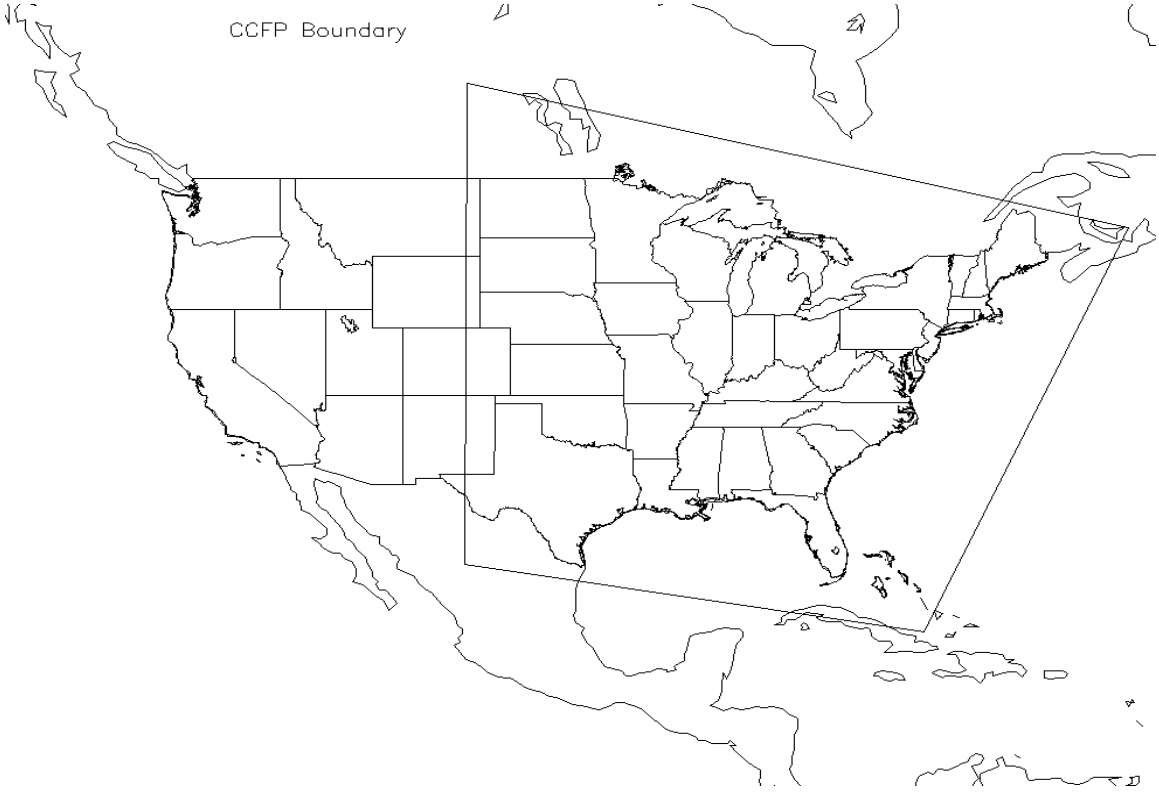

Figure 2. Solid line represents geographic boundary defined for the exercise.

Observations that fell within a 10-minute time window prior to the forecast valid time were mapped to the 20-km grid and used for verification. All forecast products, excluding the 4-h Outlook were subjected to this criterion to ensure consistency among results. Additional criteria were applied to the 4-h Outlook, where all observations within the 4-h period were mapped to the forecasts. However, we do recognize that each forecast is distinct and is often issued under different weather assumptions.

5.2. Statistical Verification Methods

The verification methods used in this study are based on standard verification concepts (Murphy and Winkler 1987). The methods were developed by the Quality Assessment Group of the FAA Aviation Gridded Forecast Systems PDT and the Convective Weather PDT. To ensure that the study was complete and fair, statistics were generated using various observational data.

As described in Section 5.1, the convective forecasts and observations are treated as Yes/No values. For instance, a convective forecast indicates convective activity inside the forecast polygon and an absence of convective activity outside the polygon. Observations are converted to *Yes/No* values by applying thresholds to the data fields. The verification methods are based on the *Yes/No* two-by-two contingency table (Table 1), where the forecasts are represented by the rows, and the observations are represented by the columns.

Table 1. Basic contingency table for evaluation of dichotomous (e.g., Yes/No) forecasts. Elements in the cells are the counts of forecast-observation pairs.

	Observation		
Forecast	***Yes***	***No***	***Total***
Yes	YY	YN	YY+YN
No	NY	NN	NY+NN
Total	YY+NY	YN+NN	YY+YN+NY+NN

Table 2 lists the verification statistics used in this evaluation. Based on the 2x2 table, PODy, PODn, and FAR are the primary verification statistics used in the analysis. PODy and PODn are estimates of the proportion of *Yes* observations that were correctly forecast and *No* observations that were correctly forecast, respectively (Brown et al. 1999; Brown et al. 1997). FAR is the proportion of *Yes* forecasts that were incorrect. The Bias is the ratio of the number of *Yes* forecasts to the number of *Yes* observations and is a measure of over and underforecasting. The Critical Success Index (CSI), also known as the Threat Score, is the proportion of hits that were either forecast or observed. The True Skill Statistic (TSS) (Doswell et al. 1990) is a measure of the ability of the forecasts to discriminate between *Yes* and *No* observations, and is also known as Hanssen-Kuipers discrimination statistic (Wilks 1995). The Heidke Skill Score is the percent correct,

corrected for the number expected to be correct by chance. The Gilbert Skill Score (Schaefer 1990), also known as the Equitable Threat Score, is the CSI corrected for the number of hits expected by chance. The % Area is the percentage of area of the forecast domain where convection is expected to occur (Brown et al. 1997). Area Efficiency is the ratio of PODy to % Area. Most of the results presented here will concern PODy, PODn, FAR, Bias, TSS, and % Area. Other statistics are included in the web-based results.

Table 2. Verification statistics used in this study.

Statistic	***Definition***	***Description***
PODy	YY/(YY+NY)	Probability of Detection of "Yes" observations
PODn	NN/(YN+NN)	Probability of Detection of "No" observations
FAR	YN/(YY+YN)	False Alarm Ratio
CSI	YY/(YY+NY+YN)	Critical Success Index
Bias	(YY+YN)/(YY+NY)	Forecast Bias
TSS	PODy + PODn – 1	True Skill Statistic
Heidke	[(YY+NN)-C1]/(N-C1), where N=YY+NY+NY+NN C1=[(YY+YN)(YY+NY) + (NY+NN)(YN+NN)] / N	Heidke Skill Score
Gilbert	(YY-C2)/[(YY-C2)+YN+NY], where C2=(YY+YN)(YY+NY)/N	Gilbert Skill Score
% Area	(Forecast Area) / (Total Area) x 100	% of the area of the forecast domain where convection is expected to occur
Area efficiency	(PODy x 100) / % Area	PODy (x 100) per unit % Area

5.3. Stratifications

The convective areas defined by the CCFP were stratified using 4 types of criteria: maximum tops (e.g. height), areal coverage, probability of occurrence, and growth rate. The statistical results were also stratified using these categories. The stratification criteria and their categories are:

1) Maximum Tops (Height)
- At or above 25,000 ft;
- 25 - 31,000 ft;
- 31 - 37,000 ft; and
- Above 37,000 ft.

2) Areal Coverage. Statistics were generated for each of the coverage categories, however, no attempt was made to vary the observation criteria within a specific coverage category.

From 1 - 26 June; the coverage categories were
- 25% and above;
- 25 - 49%; and
- 50% and above.

From 27 June - 31 August, the coverage categories
- 25% and above;
- 25 - 49%;
- 50 - 74%; and
- 75% and above.

3) Probability of Occurrence
- High (70 - 100%);
- Medium (40 - 69%); and
- Low (1 - 39%).

The method for identifying the high, medium, and low probability areas changed from single circles during the evaluation period to circles within circles. As a consequence, statistics were generated independently for each specific probability category. Results are presented here only for the *all* height and coverage categories combined.

6. Results

The overall and daily results presented here were generated using the NCWDP observations, since this detection product combines both radar reflectivity with lightning reports to produce a clear picture of the convective activity. These results represent the "*all*" height and coverage categories, with additional results (e.g., stratified by height and

coverage) available on the web, described in Section 2. As noted earlier, the c-SIGMET and the c-SIGMET Outlook results assume that both products are valid only at the end of the forecast period, with the exception of the 4-h Outlook for which the forecast is valid for the entire 4-h period. The c-SIGMET is assumed to be a 1-h forecast with the Outlook a 4-h and 6-h forecast. (Statistics generated for other forecast lengths are available on the web, and will be presented in a subsequent document.) The observations within the 10-minute window prior to the valid times were used to verify most of the forecasts, and a 4-h time window was used for the 4-h Outlooks. The results here focus on the variations in the statistical values with forecast length, issue time, product type, and with unfiltered or filtered observations, rather than on the absolute values of the statistics.

6.1. Overall Results

Overall results for the CCFP and the FG forecasts, verified using the NCWDP observations for all heights and coverage areas, are shown in Table 3. The c-SIGMET and 4- and 6-h c-SIGMET Outlook results with corresponding valid times and forecast lengths are also included in Table. The forecasts in Table 3 are grouped by issue time. Note that the FG forecasts issued at 1400 and 1715 UTC were used to generate the CCFP forecasts issued at 1500 and 1900 UTC, respectively.

To evaluate the CCFP characteristics in light of decisions made by traffic managers, specific factors were analyzed: whether convective activity occurred within the forecast region, the portion of the forecast that was incorrect, and the size of the forecast region. First, if the forecast correctly identifies the convective activity, decision-makers can confidently reroute traffic around the forecast. As measured through the PODy and PODn, the range of PODy values for the CCFP and the FG forecast is between 0.21 and 0.35. These low PODy values immediately indicate the difficult nature of forecasting convection. The PODn, as compared to the PODy, is large for both forecast types, reaching values of nearly 1.00. These results suggest that areas with no convective activity are easier to identify than areas with convection. However, if the size of the forecast area were increased, as shown by the c-SIGMET Outlook results in Table 3, an improvement in PODy could occur, but other statistics, such as PODn, FAR, Bias, and % Area would degrade.

Second, if the forecast is incorrect, naturally it is difficult for traffic managers to make rerouting decisions. For the CCFP, FG forecast, and the 6-h Outlook, the FAR (a measure of the proportion of *Yes* forecasts that were incorrect) was quite high, with values ranging from 0.80 to nearly 0.92, except for the c-SIGMET, where the FAR values ranged between 0.67 and 0.76. The 4-h Outlook had FAR values approaching a value of 0.66. The bias for all forecasts (Table 3) had values between 1 and 2, with larger values computed for the c-SIGMET Outlooks. These results indicate a tendency to overforecast convective activity by all of the different types of forecasts. However, to some degree, the overforecasting may be inherent in the nature of convective forecasts, since

convective activity usually is short-lived and difficult to predict at forecast lengths greater than 1 or 2 h.

Finally, traffic managers use area forecasts to divert aircraft traffic between convectively active regions in which a small opening between convective complexes may make the difference between opening or closing an airport. As a result, the smaller the area and higher the PODy, the better the forecast. The results in Table 3 indicate the percentage of area covered by either the FG forecast or the CCFP is relatively small, ranging from 2 - 7%. The longer forecasts generated for the c-SIGMET Outlooks (the 4- and 6-h forecasts) have areas that are 2 or 3 times larger than those for the shorter-term forecasts. The average % Area for the 1-h c-SIGMETs is about half as large as the areas for the CCFP or FG forecast. However, the c-SIGMET forecasts do not allow for much long-term planning by traffic managers.

Table 3. Contingency table for the First Guess (FG) forecast, the Collaborative Convective Forecast Product (CCFP), c-SIGMETs, and c-SIGMET Outlooks for 92 days from 1 June - 31 August 1999 with observations based on the National Convective Weather Detection Product (NCWDP) for all height and coverage categories combined.

Forecast	Issuance Time	Forecast Length	Valid Time	PODy	PODn	FAR	Bias	%Area
	Z	hrs	Z					
FG	1400	2	16	0.25	0.98	0.82	1.41	2.32
FG	1400	4	18	0.25	0.96	0.86	1.74	4.16
FG	1400	6	20	0.25	0.95	0.86	1.78	5.53
FG	1715	5	22	0.32	0.94	0.85	2.12	6.77
FG	1715	7	00	0.27	0.95	0.88	2.23	6.02
FG	1715	9	02	0.21	0.96	0.90	2.06	4.41
CCFP	1500	1	16	0.27	0.98	0.80	1.38	2.25
CCFP	1500	3	18	0.26	0.96	0.85	1.75	4.18
CCFP	1500	5	20	0.27	0.95	0.85	1.88	5.82
CCFP	1900	3	22	0.35	0.94	0.84	2.27	7.28
CCFP	1900	5	00	0.31	0.94	0.87	2.44	6.59
CCFP	1900	7	02	0.24	0.96	0.89	2.25	4.81
c-SIGMETs	1400	2	16	0.15	0.99	0.76	1.00	0.61
c-SIGMETs	1500	1	16	0.21	0.99	0.67	1.01	0.62
c-SIGMETs	1500	2	17	0.14	0.99	0.73	1.03	0.53
Outlooks	1500	6	21	0.42	0.88	0.90	4.02	12.50
Outlooks	1500	4	21	0.41	0.91	0.66	1.20	12.36
Outlooks	1900	6	01	0.52	0.84	0.92	6.68	16.76
Outlooks	1900	4	01	0.51	0.88	0.65	1.46	16.57

6.2. Daily Results

Daily statistics for the CCFP for the 1-, 3-, and 5-h forecasts issued at 1500 UTC are shown in Fig. 3. The time series plots in Figs. 3a - c show values of PODy, PODn, and FAR for each day of the evaluation. Figures. 3d - f are scatterplots of TSS vs % Area, PODy vs 1-PODn, and PODy vs % Area, respectively, where each symbol on the plot represents one forecast per day. Symbols clustered near the upper left-hand corner on the scatter plots would indicate the *ideal* forecast. Large variability in PODy from day to day is noted in Fig. 3a, with large peaks in PODy present for some days and not for others. However, the highest peaks generally occur for the 1-h forecast. For instance, the peak on 25 June for the 1-h forecast reaches a maximum PODy of 0.70 and remains high for both the 3- and 5-h forecasts. In this particular case, the convective activity (Fig. 4) was contained along the Gulf Coast and was nearly covered by the CCFP forecast. The % Area, however, was nearly 13%, somewhat larger than the average value of 7%. Nevertheless, the convective activity warrants the larger area. The FAR for this case was only 80%, which was 2% smaller than the average FAR computed for all FG cases at a valid time of 1600 UTC (Table 3).

Large variability also is evident in the daily FAR values as shown in Fig. 3c. The values for the shorter forecast lengths are quite small on some days, reaching as low as 0.44, and very large on other days, with values up to 1.0. For instance, Fig. 5 displays forecast areas and observations for 24 August, in which case the areas of the CCFP tightly enclosed the convective activity, contributing to the low FAR value. However, the confined nature of the convection provided the opportunity for this situation to occur. In contrast, the convective activity that occurred on 30 June (Fig. 6) was difficult to capture with one or even two small convective forecast areas, because of the scattered isolated convection over the upper mid-western states. The FAR for the CCFP on this day reached 0.86.

In contrast to PODy and FAR, the daily values of PODn in Fig. 3b are nearly consistent from day to day with values remaining at or above 0.90. This result suggests that forecasts of *no convection* are often more accurately defined than areas of convective activity. Figures 4, 5, and 6 show that the majority of the domain is free from convection, with the absence of lightning or NCWDP areas, except for isolated cells apparent in Fig. 6, which are practically omitted when a filter is applied (described in Section 6.4).

The % Area covered by the forecasts is smaller for shorter forecast lengths (Fig. 3d and f), as shown by the triangles, which extend from 4 - 18%, as compared to the asterisks (*), which remain clustered around 1 - 4%. The % of area covered by the 5-h forecast is larger than for the shorter forecast lengths, as shown in Fig. 3f by the 1% to nearly 20% scatter in the 5-h forecasts. This increase in area is a result of the greater uncertainty in convective activity at those longer forecast lengths.

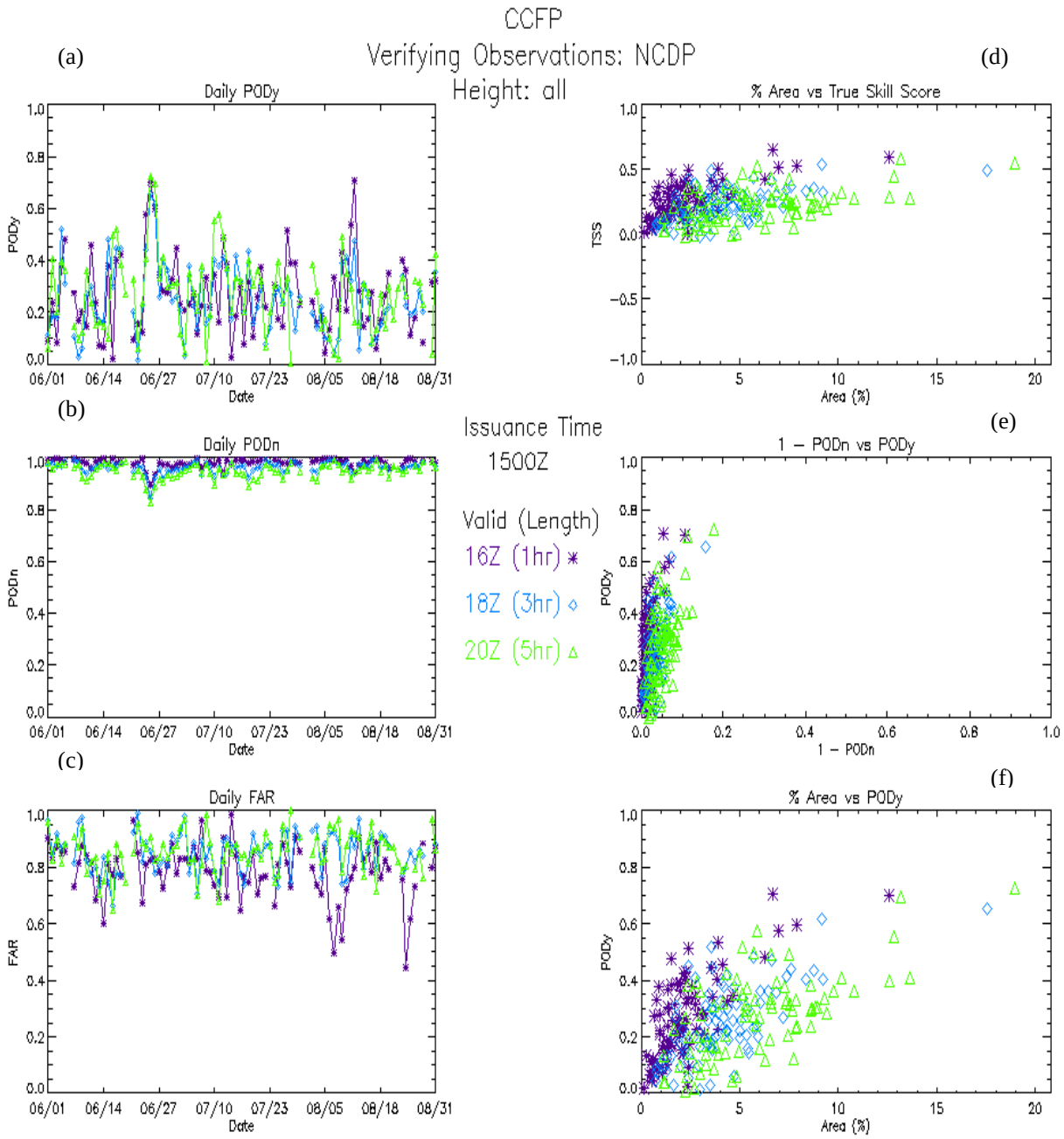

Figure 3. Daily results shown through time series and scatterplots for 92 days from 1 June - 31 August 1999 for the 1- (asterisk), 3- (diamond), and 5-h (triangle) CCFP forecasts issued at 1500 UTC: (a) PODy; (b) PODn; (c) FAR; (d) TSS vs % Area; (e) PODy vs 1-PODn; and (f) PODy vs % Area. Each dot on the scatterplots represents one forecast period per day.

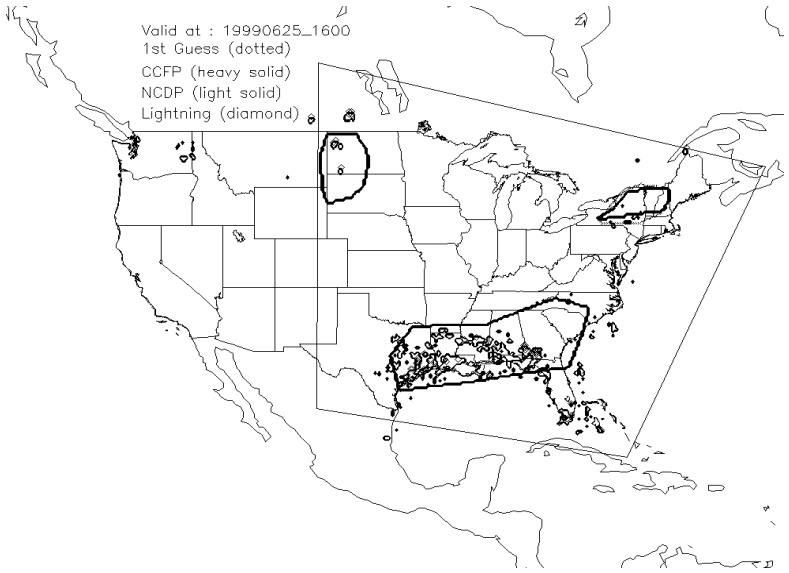

Figure 4. Display of FG forecast (dotted) and CCFP (heavy solid) with the NCWDP (light solid) and lightning (diamond) observations for 25 June valid at 1600 UTC. Large thin-lined box is CCFP domain.

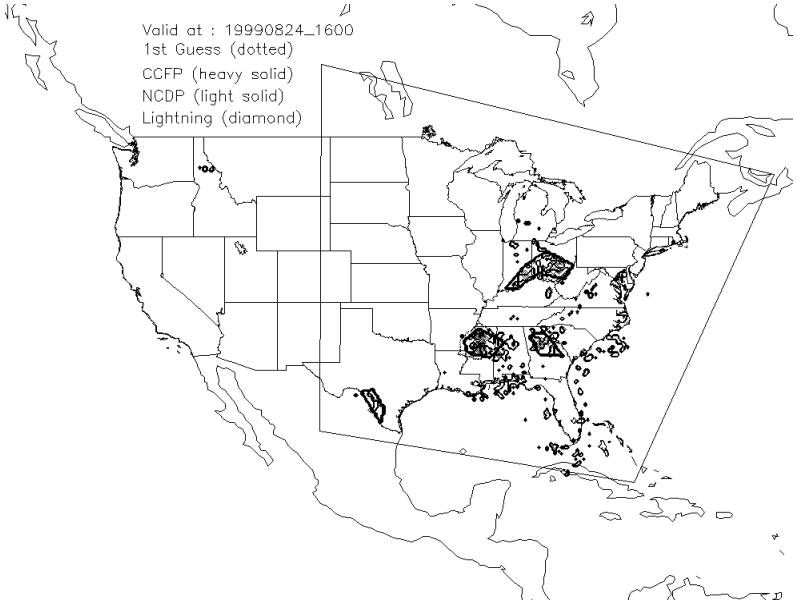

Figure 5. As in Fig. 4, except for 24 August forecasts valid at 1600 UTC.

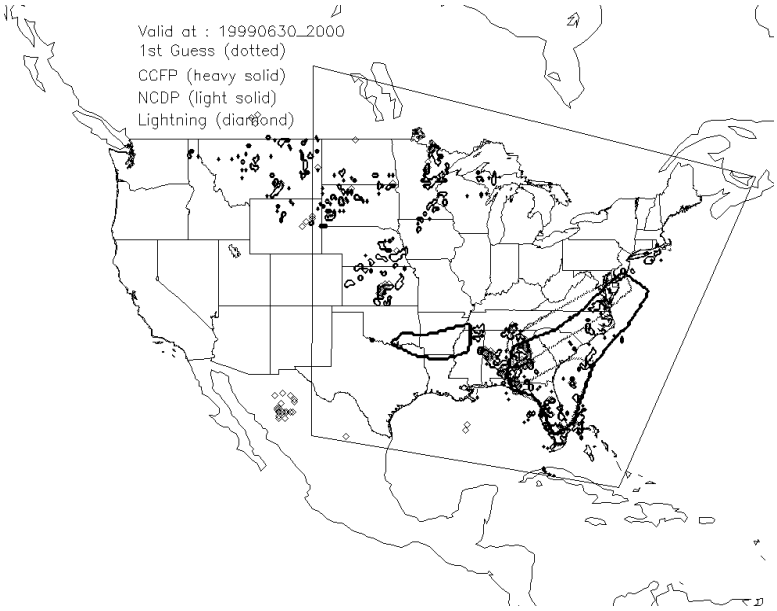

Figure 6. As in Fig. 4, except for 30 June forecasts valid at 2000 UTC.

6.3. Forecast Length Comparisons

6.3.1. Daily Plots

Forecast quality is important to traffic managers, particularly at longer forecast lengths when plans are being developed for routing traffic. In this section, results for the 5- and 7-h forecasts are compared.

Figure 7 shows statistical results for the 5-h CCFP forecasts issued at 1500 and 1900 UTC and the FG forecasts issued at 1715 UTC. The panels are similar to those described for Fig. 3. Immediately apparent in Fig. 7a is the absence of peaks greater than 0.5 in the PODy for forecasts issued late in the convective season, extending from 15 July - 31 August. This feature may suggest that the nature of the convection changes during the summer season from defined lines of thunderstorms to smaller areas of isolated convection, which are more difficult to identify and capture in a forecast (e.g., as shown by the isolated convection in Fig. 4). Overall values of PODy for the 5-h forecasts varied only slightly among the three types of 5-h forecasts, as shown in Table 3, while the scatter in the distributions are nearly identical, as shown in Fig. 7f.

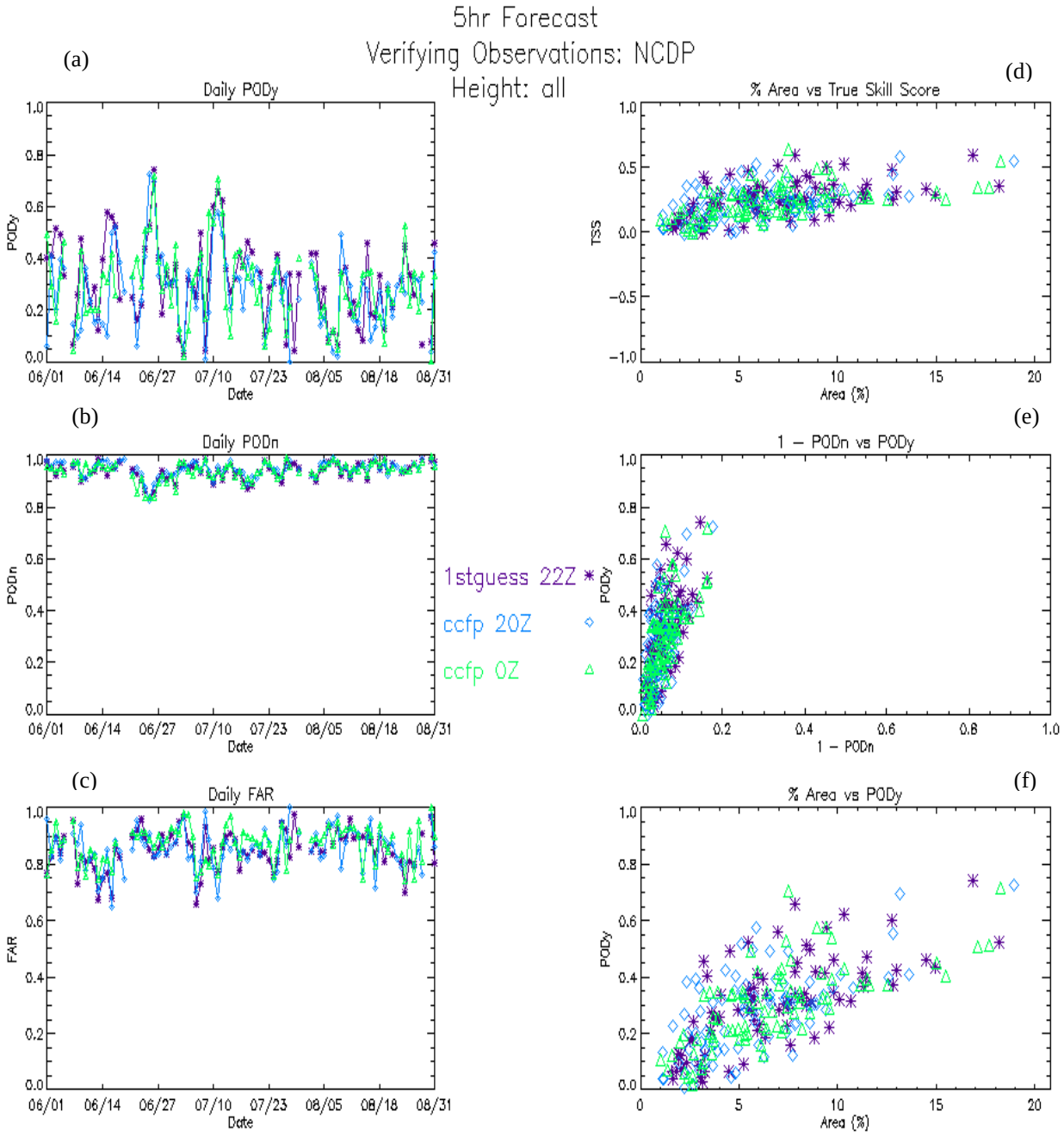

Figure 7. As in Fig. 3, except for 5-h FG forecasts issued at 1715 UTC valid at 2200 UTC, and for the CCFP issued at 1500 and 1900 UTC valid at 2000 UTC and 0000 UTC, respectively.

Little difference is also noted in the PODn (Fig. 7b), and FAR (Fig. 7c) values as shown by the over-lapping lines on the plots. This similarity is also evident in the overall statistics (Table 3) where the PODn and FAR values for the 3 forecasts were nearly identical.

Results for the 7-h CCFP forecasts issued at 1900 UTC and the FG forecasts issued at 1715 UTC are shown in Fig. 8. The peaks in PODy that are apparent early in the season (Figs. 3a and 7a) for 25 June and 7 July are also peaks for the 7-h forecasts (Fig. 8a). This result may suggest that convection correctly identified at the 1-, 5-, and 7-h forecast lengths is sustainable long-lived activity, perhaps associated with a front or large-scale meteorological feature, as opposed to isolated small-scale convection, which would make it easier to identify and track. As was the case for the 5-h forecasts, the PODy values for the 7-h forecasts (Fig. 8a) also decreased during the second half of the season. However, the overall PODy value was reduced 5% when 2-h was added to the forecast length. In contrast, the overall PODn (Fig. 8b) and FAR (Fig. 8c) values increased only slightly (Table1) when the forecast length increased from 5 to 7-h. These results are considered further in Section 6.3.2 through box plot diagrams of the distributions of daily statistics.

6.3.2. Daily Distributions

Day-to-day variations of the verification statistics are considered using box plots. These plots show the distributions of the daily values of the various statistics. For example, Figure 9 shows the distributions of the various statistics for all 5-h FG and CCFP forecasts, with the FG forecasts issued at 1715 UTC (FG17), and the CCFP issued at 1500 and 1900 UTC (CCFP15 and CCFP19, respectively). The plots in Figure 9 show various quantiles of the daily verification statistics for these forecasts.

The box portion of a box plot encloses the region between the 0.25^{th} and the 0.75^{th} quantiles (i.e., the middle 50% of the distribution), and the line inside the box represents the median value, for which 50% of the values is larger and 50% is smaller. The ends of the "whiskers" extending above and below the box are the 0.95^{th} and 0.05^{th} quantiles (i.e., the values for which 5% of the daily statistics is larger and smaller, respectively). Finally, the open point inside the box represents the mean value of the statistic and the points above and below the ends of the whiskers are the extreme large and small values. The notches on the sides of the box represent an approximate 95% confidence interval for the median. The differences between two medians can be assumed to be statistically significant if their confidence intervals (i.e., notched regions) do not overlap.

As an example, consider the PODy plot for FG17 in Figure 9a. This plot indicates that the lower quartile (0.25^{th} quantile) value for PODy is about 20%, the median is about 30%, and the upper quartile (i.e., 0.75^{th} quantile) is about 42%. The 0.05^{th} quantile is somewhat smaller than 10%, and the 0.95^{th} quantile is around 60%.

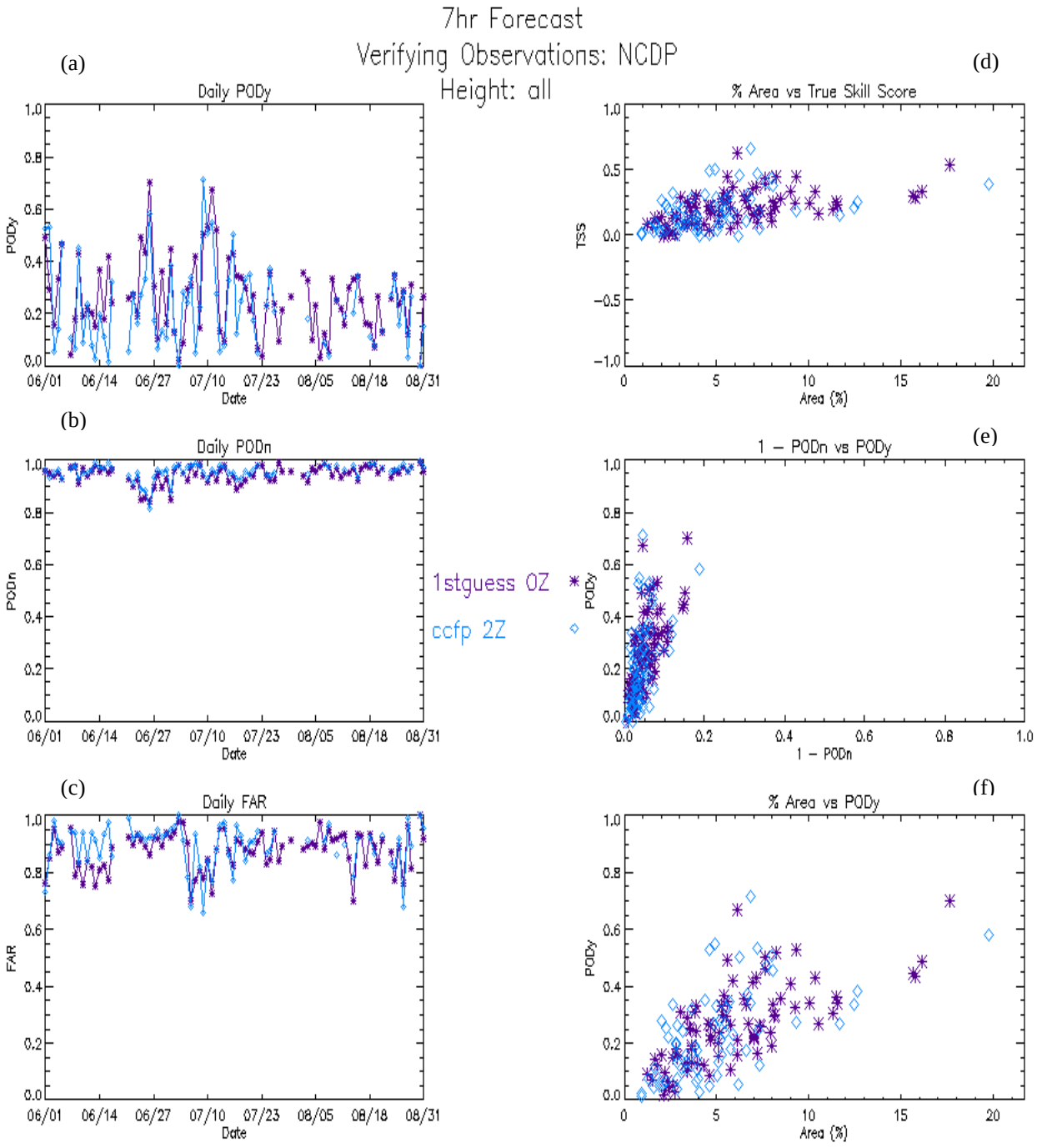

Figure 8. As in Fig. 7, except for 7-h forecasts for the FG issued at 1715 UTC valid at 0000 UTC, and the CCFP issued at 1900 UTC valid at 0200 UTC.

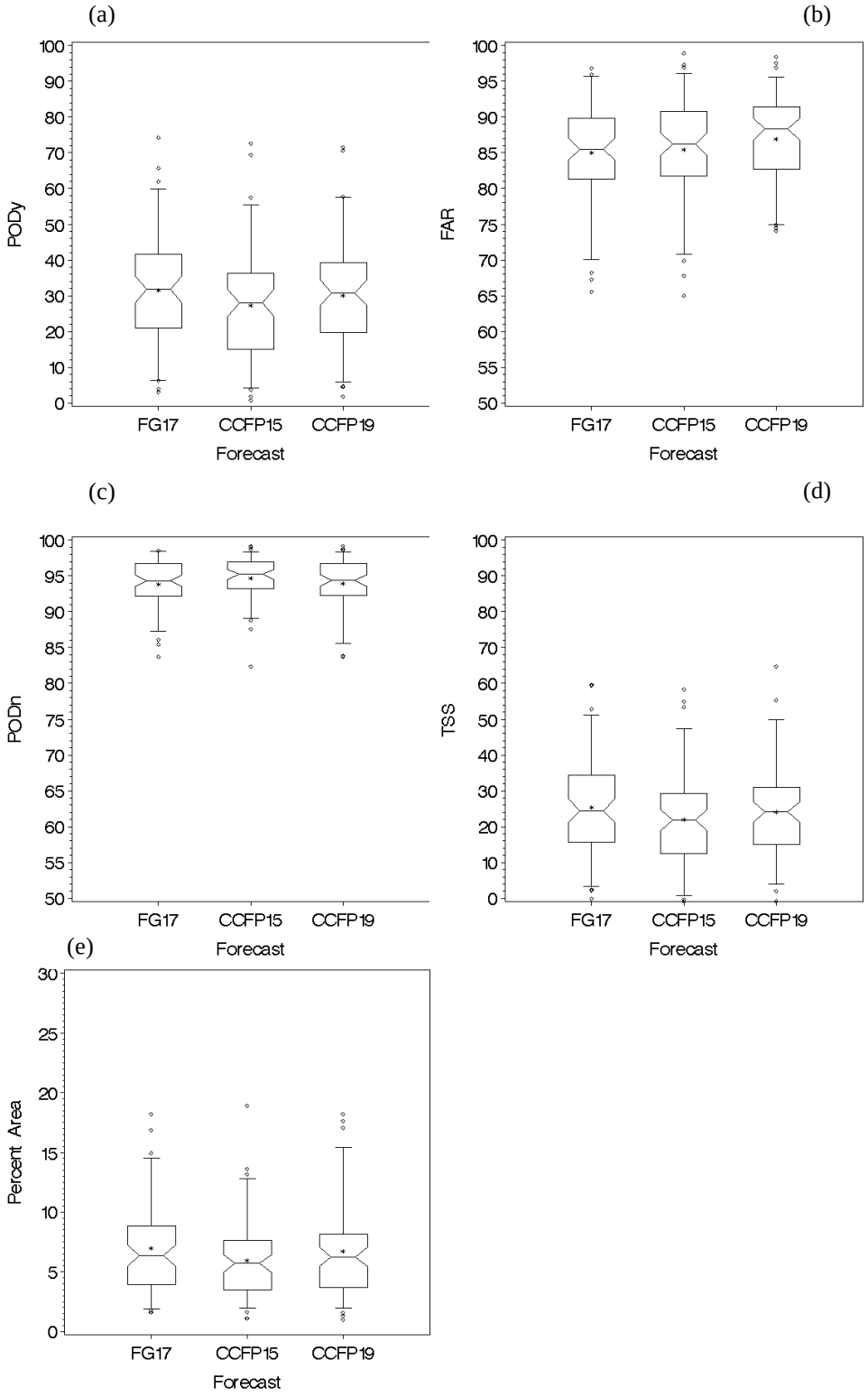

Figure 9. Distributions of daily verification statistics for 5-h forecasts, including FG forecasts issued at 1715 (FG17) and CCFP forecasts issued at 1500 and 1900 UTC (CCFP15 and CCFP19, respectively). Note that PODy, PODn, FAR, and TSS values have been converted to percentages by multiplying by 100.

Box plots are especially useful for comparing two or more distributions. The plots in Fig. 9 indicate that the PODy values for the CCFP19 forecasts are about the same as the values for the FG17 forecasts, but the PODy distribution for the CCFP15 forecasts are slightly lower than the others. Distributions of the other statistics also have similar characteristics among the three types of forecasts, with slightly larger TSS values and slightly smaller FAR values for the FG17 forecasts. However, none of the differences in the median values are statistically significant.

Figure 10 shows distributions of the differences in the statistics for all of the individual days. The box plots in this figure confirm the differences noted in Figure 9. In particular, on most days the PODy value is larger for the FG17 forecasts than for CCFP15 forecasts, as is the TSS value. Statistics for the FG17 and CCFP19 forecasts are about the same (with distributions centered on zero), as shown in Fig. 10b.

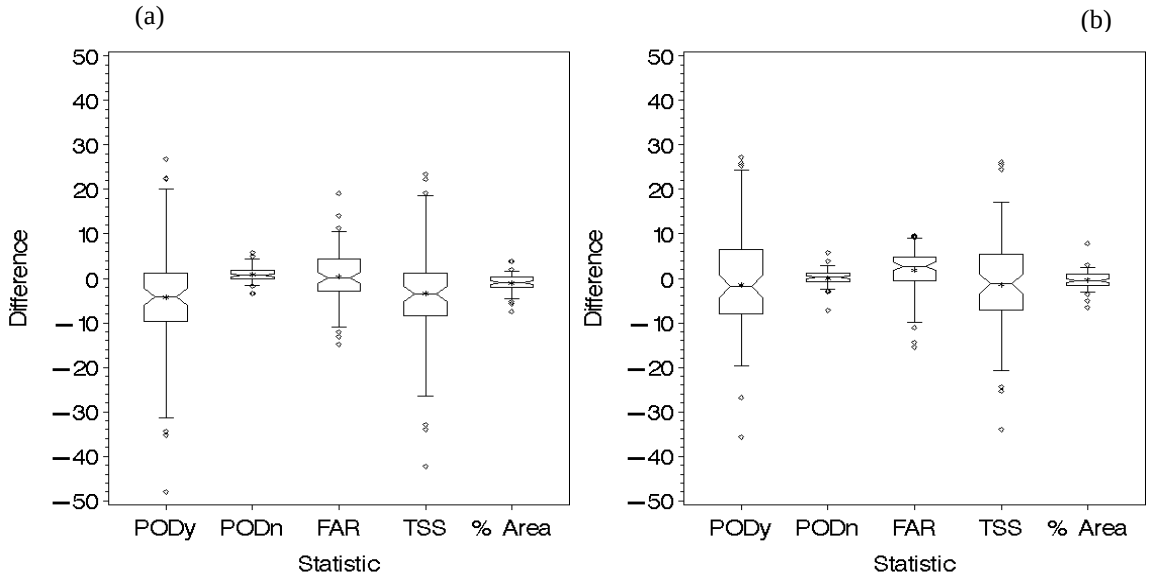

Figure 10. Distributions of differences in daily verification statistics for 5-h forecasts: (a) 1500 UTC CCFP forecasts vs. 1715 UTC FG forecasts; (b) 1900 UTC CCFP forecasts vs. 1715 UTC FG forecasts.

Figure 11 shows distributions of the verification statistics for FG and CCFP forecasts valid at 2200 UTC, in particular the 5-h FG forecasts issued at 1715 UTC and the following 3-h CCFP forecasts issued at 1900 UTC. The box plots in this figure indicate that the PODy distribution for the CCFP forecasts is somewhat less variable than the PODy distribution for the FG forecasts, with a slightly larger median PODy value for the CCFP forecasts. The PODn and FAR distributions are very similar for the two sets of forecasts, but the TSS and % Area distributions for the CCFP forecasts are somewhat higher than the corresponding distributions for the FG forecasts. These differences also are shown in Figure 12, which presents distributions of the differences in the daily statistics. This figure shows that the daily PODy values for the CCFP can be as much as 15% larger than the values for the FG forecasts, and the TSS values can be as much as 10 larger for the CCFP in comparison to the FG. For about 25% of the CCFP forecasts, the PODy value was increased by more than 5% over the FG forecast; similarly, for about 25% of the CCFP forecasts, the TSS was increased by at least 5% over the comparable FG forecast. This increase in the statistics for the CCFP compared to the FG is likely due to a combination of two factors: 1) the collaborative decision-making process, and 2) the decreased forecast lead time associated with the CCFP, and the associated availability of more recent data sources for formulating the CCFP.

Finally, Figure 13 shows the distributions of the verification statistics for the 7-h FG and CCFP forecasts, with the FG forecasts issued at 1715 UTC and the CCFP forecasts issued at 1900 UTC. These figures suggest that, in this case, the FG forecasts have slightly better performance characteristics than the CCFP forecasts, although the differences in the medians are not statistically significant. In particular, the box plots in Figure 13 indicate the FG forecasts have somewhat larger PODy and TSS values, and somewhat smaller FAR values. However, the FG forecasts also cover larger areas. These results also are demonstrated in Figure 14, which shows that 75% of the FG forecasts had larger PODy values than the corresponding CCFP forecast, and more than 50% had larger TSS values.

6.4. Filtered Results

The NCWDP was filtered (Section 5.1) in an attempt to screen out isolated convection, and leave long-lived traceable convective areas. This filter had the effect of allowing better detection of convection, as indicated by the PODy. For example (Fig. 15), when the data for the 5-h forecasts are filtered, the PODy values increase by nearly 10%. The increase in PODy suggests that if significant convection occurred, it was essentially captured by the forecast. This capability is hidden in Fig. 7 due to the inclusion of isolated convection in the verification data. Interestingly, the daily PODn trace is nearly the same for both the unfiltered (Fig. 7b) and the filtered (Fig. 15b) data.

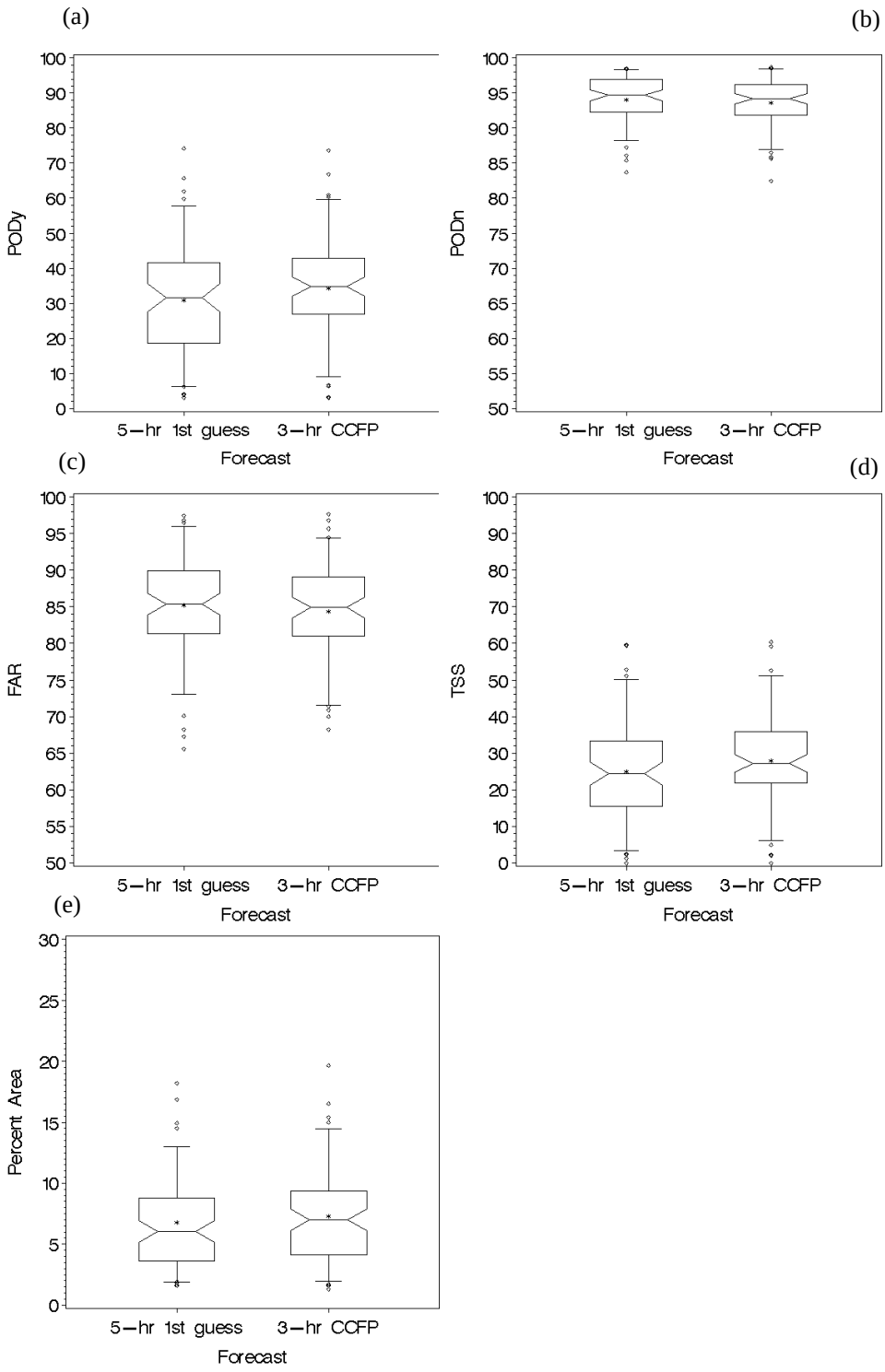

Figure 11. As in Fig. 9, forecasts valid at 2200 UTC, including 5-h FG forecasts issued at 1715 UTC and 3-h CCFP forecasts issued at 1900 UTC.

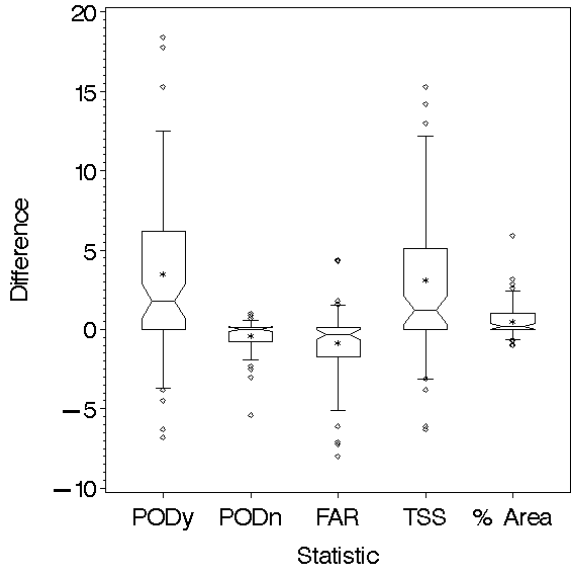

Figure 12. As in Fig. 10, forecasts valid at 2200 UTC (5-h FG and 3-h CCFP forecasts).

On the other hand, the FAR values are nearly 9% larger on average for the filtered data (Fig. 15c) than for the unfiltered data (Fig. 7c), mainly because of the reduced number of *Yes* observations when isolated convection is omitted. For example, small pockets of isolated convection are deleted from the observations (Fig. 16) for 25 June as compared to Fig. 4, leaving "white space" within the forecast area. This white space becomes part of the region of false alarms and thus contributes to the FAR. This effect also is evident in Figs. 7f and 15f, in which the scatter in PODy with % Area is visibly larger for the filtered data (Fig. 15f) than for the unfiltered data (Fig. 7f). This result suggests that the forecast areas do approximately capture the significant convection, but that the areas become too large when the isolated convection is filtered out. Moreover, the 5-h FG forecast, valid at 2200 UTC improves over the CCFP when the observations are filtered, as shown by the movement of the star-shaped symbols toward the upper left-hand corner of Fig. 15e as compared to Fig. 7e.

When the observations for the 7-h forecasts were filtered (Fig. 17), the PODy dropped and the FAR rose dramatically, suggesting that although the significant convective activity may be identified at a forecast length of 5 h, it is not appropriately identified at 7 h. This result may be linked to the absence of convection at this later valid time or it may solely be due to the increased forecast length.

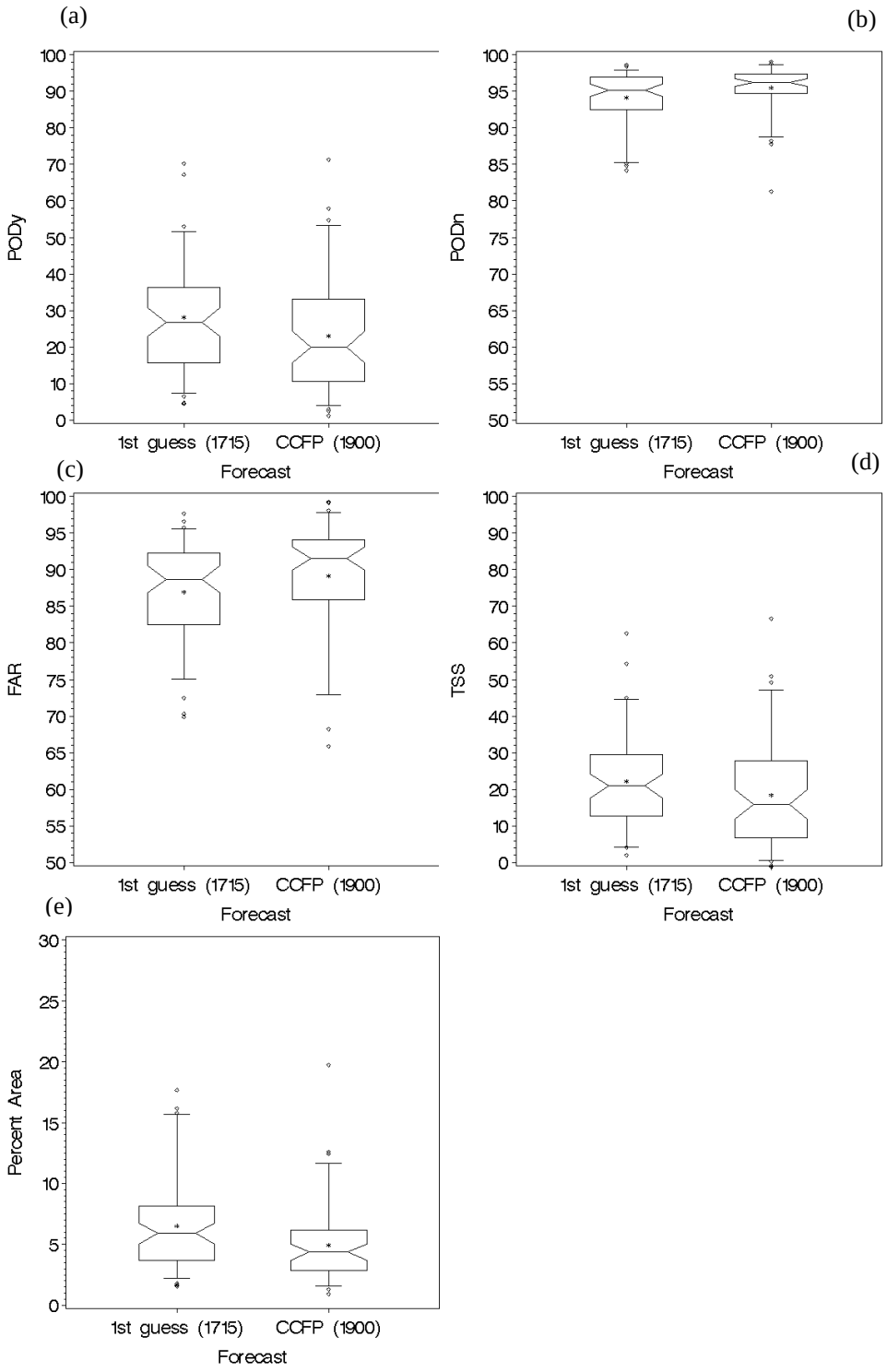

Figure 13. As in Fig. 9, for 7-h forecasts, including FG forecasts issued at 1715 UTC and CCFP forecasts issued at 1900 UTC.

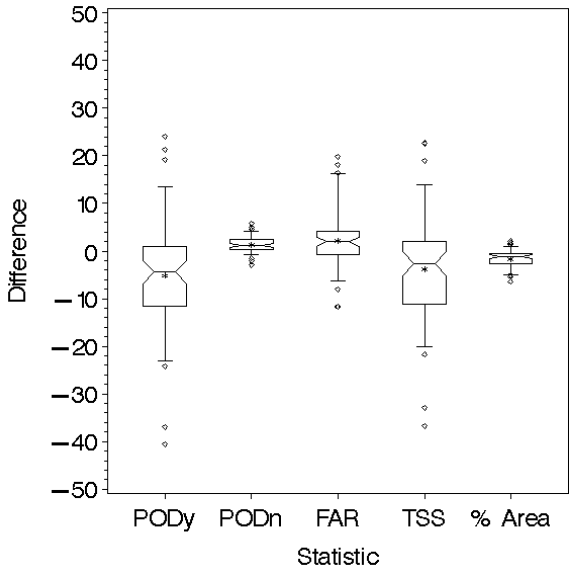

Figure 14. As in Fig. 10, for 7-h forecasts (1715 UTC FG and 1900 UTC CCFP forecasts).

7. Summary and Conclusions

In general, forecasting for convective activity is difficult, as illustrated by low values of PODy for all forecast products. PODy values increased, however, when convective activity was associated with longer-lived convection rather than short-lived isolated cells. This result is somewhat intuitive, since long-lived convection usually is associated with larger-scale traceable meteorological features. Forecasts also were able to discriminate quite well between convective and nonconvective areas, as indicated by the large values of PODn. The FAR for all forecasts was large, with values usually around 80-90%. However, the nature of the convection contributed somewhat to this high FAR. For instance, the forecast areas generally included *white space* (on the displays) between individual convective cells. This white space decreased when the convection was clustered or grouped, and increased when activity was widespread and isolated. The FAR values responded by decreasing with clustered convection and increasing with widespread convection.

It also is important to note that the verification approach utilized in this study is quite demanding. In particular, the forecasts are penalized for errors in both the timing and spatial location of convection. Thus, it isn't surprising that the PODy values appear to be small and the FAR values appear to be large. Verification statistics with a similar magnitude often are found in studies of this type, using this verification approach (e.g., Brown and Brandes 1997).

As compared to the CCFP, larger PODy and smaller FAR values were recorded for the 4-h and 6-h c-SIGMET Outlooks. However, the areas of these outlooks covered nearly one - third to one - half of the forecast domain and included moving and develop-

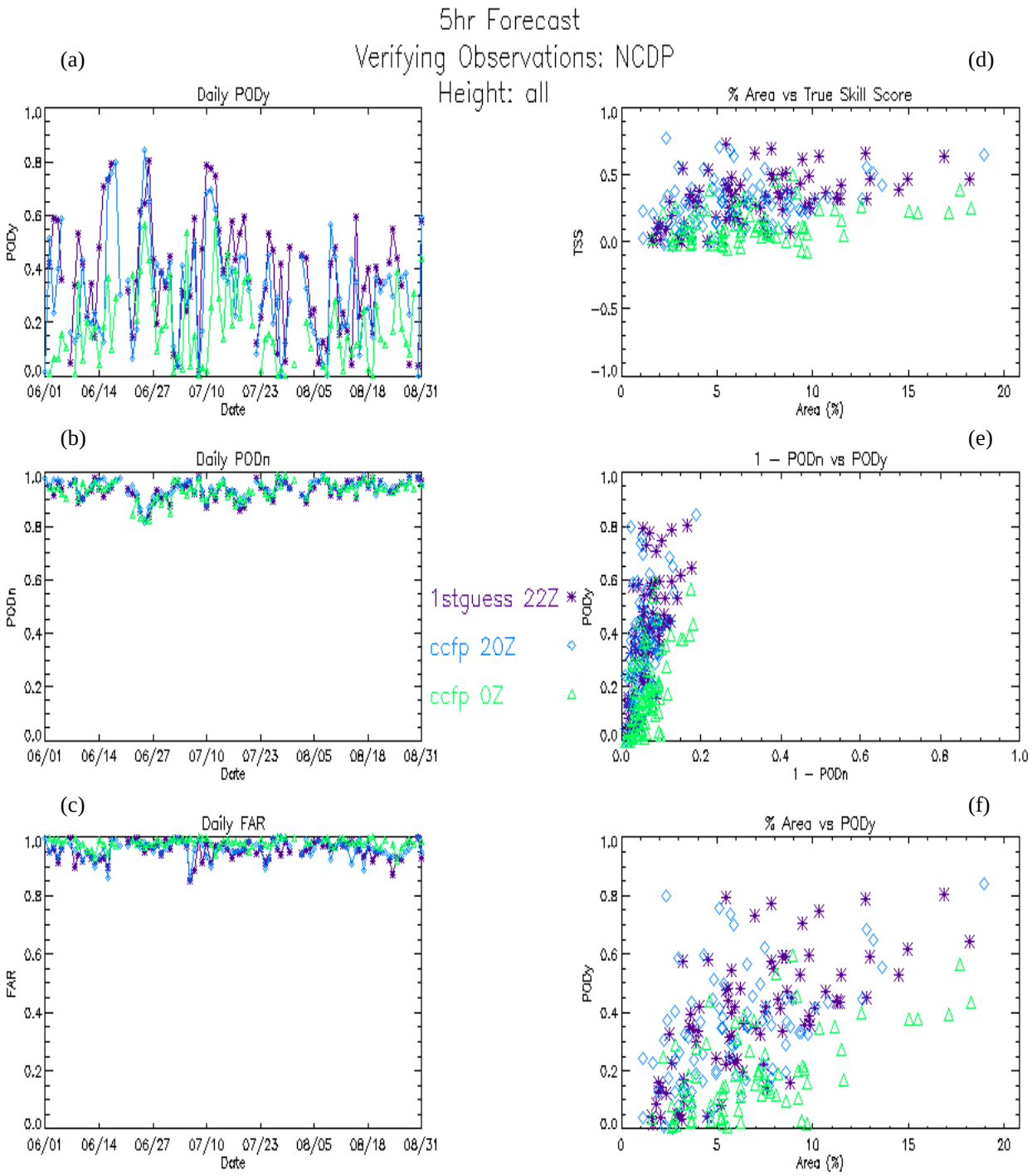

Figure 15. As in Fig. 7, except for filtered NCWDP.

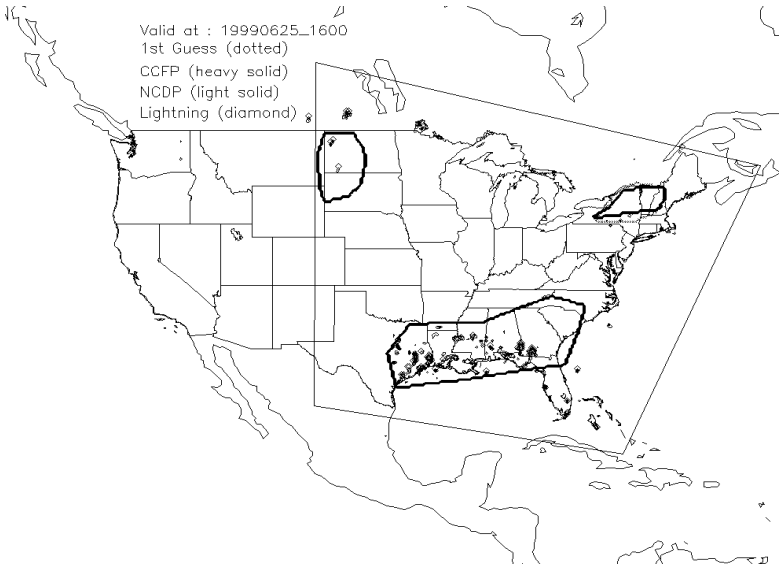

Figure 16. As in Fig. 4, except for filtered NCWDP.

ing convective activity over the entire length of the forecast. This information is difficult to interpret when trying to direct aircraft traffic around convective areas at a specific time. The CCFP, on the other hand, was valid at one specific time and covered only 2-7% of the forecast domain, providing specific detailed information to decision-makers. FAR values recorded for the c-SIGMET were relatively small, as were the c-SIGMET forecast areas. However, the 1- or 2-h forecast length for the c-SIGMET is inadequate for longer-term flight planning and decision-making.

The daily results, as presented by the box plots, indicate that the CCFP forecasts improved overall on the FG forecasts that were used as guidance in their creation. However, this improvement may result in part from the decreased forecast lead time associated with the CCFP, as well as the collaborative process used to formulate the CCFP. Comparisons of FG and CCFP forecasts with the same lead times suggest that the two sets of forecasts have similar performance characteristics, with the FG forecasts improving over the CCFP forecasts in a number of cases. This difference raises the issue of whether the FG forecast might sufficient for this forecasting process.

Filtering the observations improved the PODy for all forecast periods except the 7- and 9-h forecasts, in which the values decreased considerably. This effect may be a result of the time of day at which the 7- and 9-h forecasts are valid or it may be due to the long forecast length, which is uncharacteristic of short-lived convective weather. Rarely does convection maintain its form 7- to 9-h into the future. The FAR, on the other hand, increased nearly 10% for the CCFP and FG forecasts when the observations were filtered.

Results from this evaluation will be analyzed further to include specific examples of the c-SIGMETs, c-SIGMET Outlooks, and the National Convective Weather Forecast Product (not presented here). In addition, detailed analyses will be undertaken to

investigate the usefulness of the verification methods, observation types, and assumptions used to evaluate these forecast products.

Plans are underway to continue this intercomparison exercise through the summer of 2000. The intercomparison will again involve providing real-time statistical displays

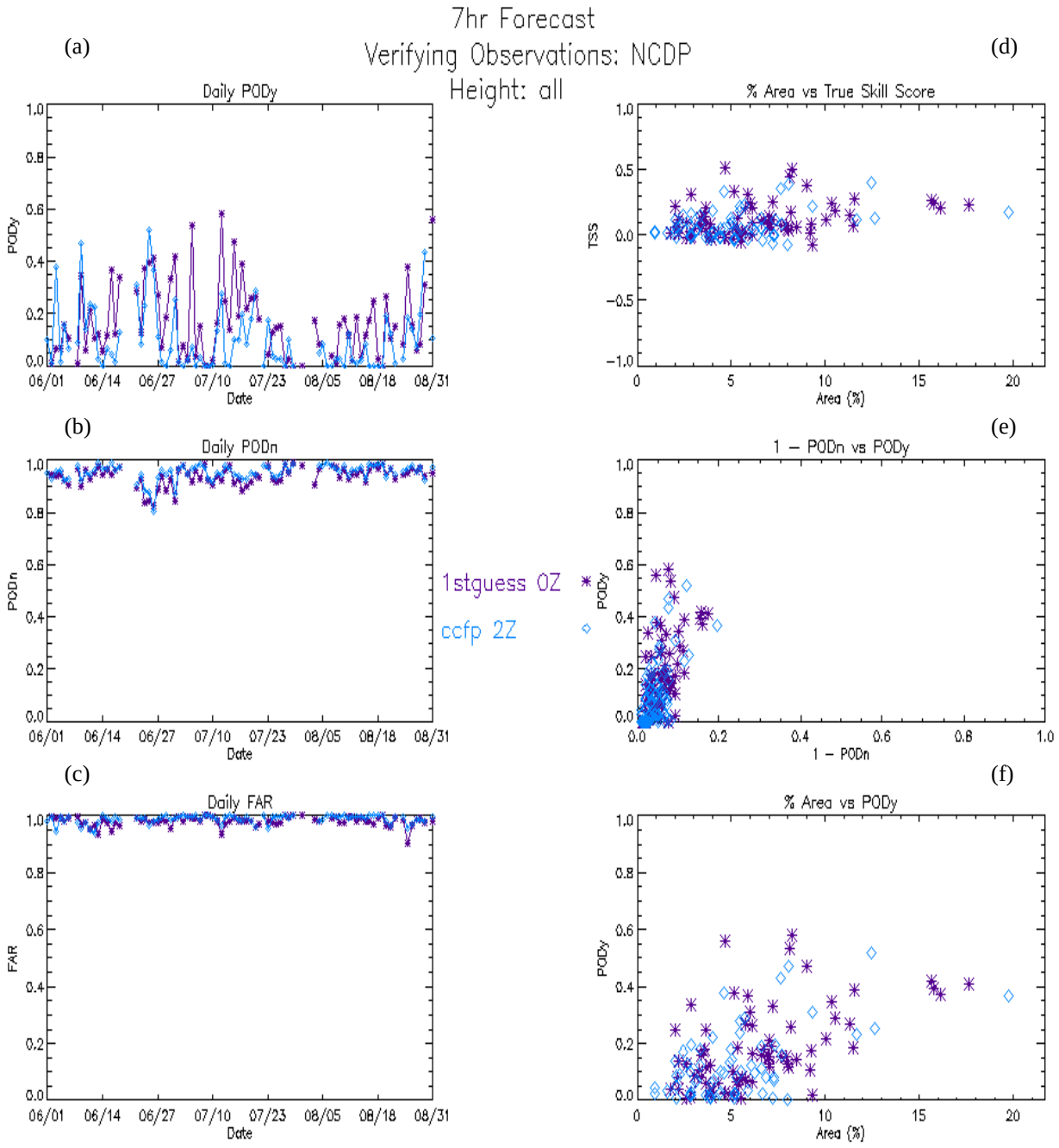

Figure 17. As in Fig. 8, except for filtered NCWDP.

generated by the Real-Time Verification System through the World Wide Web, as well as a detailed statistical analysis of the verification results presented in a written document. We also hope to develop methods that will allow a more thorough evaluation of the coverage and probability forecasts. Several questions that should be answered prior to the start of the next exercise include: 1) How should the assumptions used to classify the observations change to account for the variations in the probability of occurrence? 2) Does the combined lightning and radar detection field (NCWDP) adequately represent convection so that individual lightning and radar data sets can be eliminated from the exercise? and 3) Do the filtered observations represent long-lived convective activity?

Acknowledgements

This research is in response to requirements and funding by the Federal Aviation Administration. The views expressed are those of the authors and do not necessarily represent the official policy and position of the U.S. Government.

We would like to thank Cindy Mueller (NCAR) for her effort in helping to develop verification methods, Nancy Rehak (NCAR) for providing the NCWDP, and Don Frank (AWC) for providing the FG and CCFP forecasts to the RTVS. We also would like to thank Judy Henderson (FSL) for her work on the RTVS and Jamie Riggs and Nita Fullerton (FSL) for their helpful reviews of this work.

References

Brown, B.G., and E. Brandes, 1997: An intercomparison of 2-D storm extrapolation algorithms. *Preprints, 28th Conference on Radar Meteorology,* Austin, TX, 7-12 Sept., American Meteorological Society, 495-496.

Brown, B.G., J.L. Mahoney, R. Bullock, J. Henderson, and T.L. Kane, 1999: Turbulence Algorithm Intercomparison: 1998-1999 Initial Results. FAA Turbulence Product Development Team Report to FAA Aviation Weather Research Program (Available from B. Brown, Research Applications Program, National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO 80307-3000).

Brown, B.G., G. Thompson, R.T. Bruintjes, R. Bullock, and T. Kane, 1997: Intercomparison of in-flight icing algorithms. Part II: Statistical verification results. *Wea. and Forec*., **12,** 890-914.

Doswell, C.A., R.Davies Jones, and David L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. and Forec.*, **5**, 576-585.

FAA, 1999: The Program Plan for Operational Test Program for the Collaborative Convective Forecast Product.

Mahoney, J.L., J.K. Henderson, and P.A. Miller, 1997: A Description of the Forecast Systems Laboratory's Real-Time Verification System (RTVS). Preprints, *7th Conference on Aviation, Range, and Aerospace Meteorology*, Long Beach, American Meteorological Society, J26-J31.

Mueller, C.K., C.B. Fidalego, D.W. McCann, D. Meganhart, N. Rehak, and T. Carty, 1999: National Convective Weather Forecast Product. *Preprints, 8th Conference on Aviation Range, and Aerospace Meteorology,* American Meteorological Society (Boston), 230-234.

Murphy, A.H. and R.L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330-1338.

NWS, 1991: National Weather Service Operations Manual, D-22. National Weather Service. (Available at Website http://www.nws.noaa.gov).

Orville, R.E., 1991: Lightning ground flash density in the contiguous United States-1989. *Mon. Wea. Rev.*, **119**, 573-577.

Phaneuf, M. W. and D. Nestoros, 1999: Collaborative convective forecast product: Evaluation for 1999. (Available from the author at CygnaCom Solution, Inc.)

Sankey, D., K.M. Leonard, W. Fellner, D.J., Pace, K.L. Van Sickle, 1997: Strategy and Direction of the Federal Aviation Administration's Aviation Weather Research Program. Preprints, *7th Conference on Aviation, Range, and Aerospace Meteorology*, Long Beach, American Meteorological Society, 7-10.

Schaefer, J.T., 1990: The Critical Success Index as an indicator of warning skill. *Wea. and Forec.*, **5**, 570-575.

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