Quality Assessment Report

National Convective Weather Forecast 2 (NCWF-2)

Quality Assessment Product Development Team

Stacey A. Seseske[1,2], Michael P. Kay[1,2], Sean Madine[1,3], Joan E. Hart[1,2], Jennifer L. Mahoney[1], and Barbara Brown[4]

9 March 2006

Minor modifications added

[1]NOAA Research/Earth System Research Laboratory, Boulder, Colorado

[2]Joint collaboration with the Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado

[3]Joint collaboration with the Collaborative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado

[4]Research Applications Laboratory, National Center for Atmospheric Research, Boulder, Colorado

Contact: Jennifer Mahoney (NOAA/GSD 325 Broadway, R/GSD5, Boulder, Colorado 80305; Jennifer.Mahoney@noaa.gov)

Contents

Section Page

Quality Assessment Report for the

National Convective Weather Forecast (NCWF-2) Product

Quality Assessment Product Development Team

9 March 2006

Executive Summary

This report summarizes a statistical evaluation of the quality of the National Convective Weather Forecast (NCWF-2) product, developed by the Convective Weather Product Development Team under the Federal Aviation Administration's (FAA) Aviation Weather Research Program. The document provides an assessment of the NCWF-2 forecasting capability, including quantitative verification of 1-h and 2-h probabilistic forecasts. The results of this evaluation will be provided to the FAA and National Weather Service (NWS) Aviation Weather Steering Group (AWSG) for its consideration of whether NCWF-2 is ready for transition to operations.

The components of the NCWF-2 considered in this evaluation include all forecast lead times and the probabilistic capability. An evaluation of the NCWF-2 motion vectors is presented in a supplemental report. The NCWF-2 is evaluated using the operational National Convective Weather Detection product (NCWD). For a standard of comparison, the quality of the NCWF-2 is compared to the quality of the operational 1-h NCWF, the 1-h NCWD Persistence, the 1- and 2-h C-SIGMET, and the 2-h Collaborative Convective Forecast Product (CCFP). Overall statistics for the NCWF-2 were computed from 1 April – 30 September 2005. The statistical methodology is consistent with the approach used in previous evaluations of the NCWF such as Brown and Mahoney (2000).

The results indicate that:

- Overall, the NCWF-2 shows modest reliability and resolution. The reliability and resolution decrease with lead time, with the statistics for the 30-min forecast being distinctly better than the statistics for the longer lead times.
- The 1-h probabilistic comparison shows that NCWF-2 performs better than the operational NCWF and C-SIGMET with respect to reliability and resolution.
- The 1-h deterministic comparison shows that NCWF-2 at the 0.4 probability threshold performs similarly to the operational NCWF and better than the C-SIGMET, as indicated by CSI statistics.
- The 2-h probabilistic comparison shows that NCWF-2 performs better than the CCFP and C-SIGMET with respect to reliability and resolution. The reliability of the CCFP approaches that of the NCWF-2 at higher probabilities.
- The 2-h deterministic comparison shows that NCWF-2 at the 0.4 probability threshold performs slightly better than the CCFP and C-SIGMET, as indicated by CSI statistics.

1. Introduction

The second generation National Convective Weather Forecast (NCWF-2), a probabilistic convective forecast developed by the Federal Aviation Administration Aviation Weather Research Program's Convective Weather Product Development Team (FAA/AWRP/CW PDT; Wolfson and Mueller 2006) is being considered for transition from experimental status to National Weather Service (NWS) operations. As part of this transition process, a detailed objective assessment of the forecast quality and performance is required. Therefore, the purpose of this evaluation is to determine whether the NCWF-2 has an acceptable level of forecast capability when compared to other operational convective forecast products.

Previous evaluations of the NCWF-2, carried out to advance the algorithm from test to experimental status, were conducted during the summer of 2003 (Mahoney et al. 2004). Since that time, minor modifications to the NCWF-2 algorithm have occurred and will be discussed in later sections of this report.

The forecast performance of the NCWF-2 was evaluated from 1 April – 30 September 2005 and compared to the performance of the operational version of the NCWF, the Convective Significant Meteorological Advisories (C-SIGMET), the Collaborative Convective Forecast Product (CCFP), and persistence of the National Convective Weather Detection product (NCWD). The probabilistic attribute of the NCWF-2 is the main focus of the results presented in this report. However, to be consistent with previous evaluations of the NCWF and NCWF-2, results are also presented from a deterministic point of view.

The Real-Time Verification System (RTVS; Mahoney et al. 2002) was used to collect the forecast and observation data, and to produce the statistical results used in this evaluation. In addition to the statistical results presented in this report, a full suite of statistical information can be found for the NCWF-2, the NCWF, the C-SIGMETs, and the CCFP at the RTVS web site (http://www-ad.fsl.noaa.gov/fvb/rtvs/).

This report is organized as follows. The overall approach for the evaluation is presented in Section 2. Section 3 briefly describes data, which includes the algorithms, forecasts, and observations. The verification methods are summarized in Section 4, and the results of the evaluation are presented in Section 5. Finally, conclusions are presented in Section 6.

2. Approach

Although the NCWF-2 is a probabilistic forecast rather than a deterministic forecast like its predecessor, the verification approach applied in this study is similar to the approach undertaken in previous studies of the deterministic NCWF (e.g. Brown and Mahoney 2000). However, since the probabilistic characteristics of the NCWF-2 are of particular interest, additional techniques and statistical approaches are utilized to assess the forecast performance from a probabilistic point of view.

The convective forecasts were verified using the operational National Convective Weather Detection field (NCWD), a convective observation product that is based on a combination of radar reflectivity and lightning observations (Orville 1991). For many analyses, the forecast probabilities were transformed into dichotomous fields by determining if the probability at a grid point exceeded or was less than a pre-specified threshold. Numerous thresholds were utilized to examine the full range of performance for the NCWF-2.

In order to determine the suitability of NCWF-2 for transition to operations, the forecast quality of the NCWF-2 is compared to the operational NCWF, the C-SIGMET, and the CCFP. However, it is important to emphasize that the NCWF-2 is a very different type of forecast than the C-SIGMET and CCFP. For instance, the NCWF-2 is an automated, frequently updated, gridded forecast product that produces probabilistic forecasts. The C-SIGMET and the CCFP, on the other hand, are human-generated deterministic forecasts. To account for the differences in these forecast definitions, the forecasts in this study were compared both probabilistically and deterministically. Descriptions of the forecast conversions from probabilistic to deterministic and vice versa are described in Section 4. Users of these statistics should keep these assumptions in mind when evaluating the strengths and weaknesses of each type of forecast.

3. Data

3.1. Forecasts

3.1.1. National Convective Weather Forecast-2 (NCWF-2)

The predecessor to the NCWF-2 is the operational NCWF, shown in Fig. 1a (Mueller et al. 1999; Megenhardt et al. 2000). The NCWF is an automated system that forecasts the location of convective storm areas every five minutes with a lead time of one hour. In contrast, the NCWF-2 (Mueller et al. 2004), shown in Fig. 1b, provides a probabilistic forecast, every 30 minutes out to 2 hours, for convection that will occur at a specific grid point at a specific time.

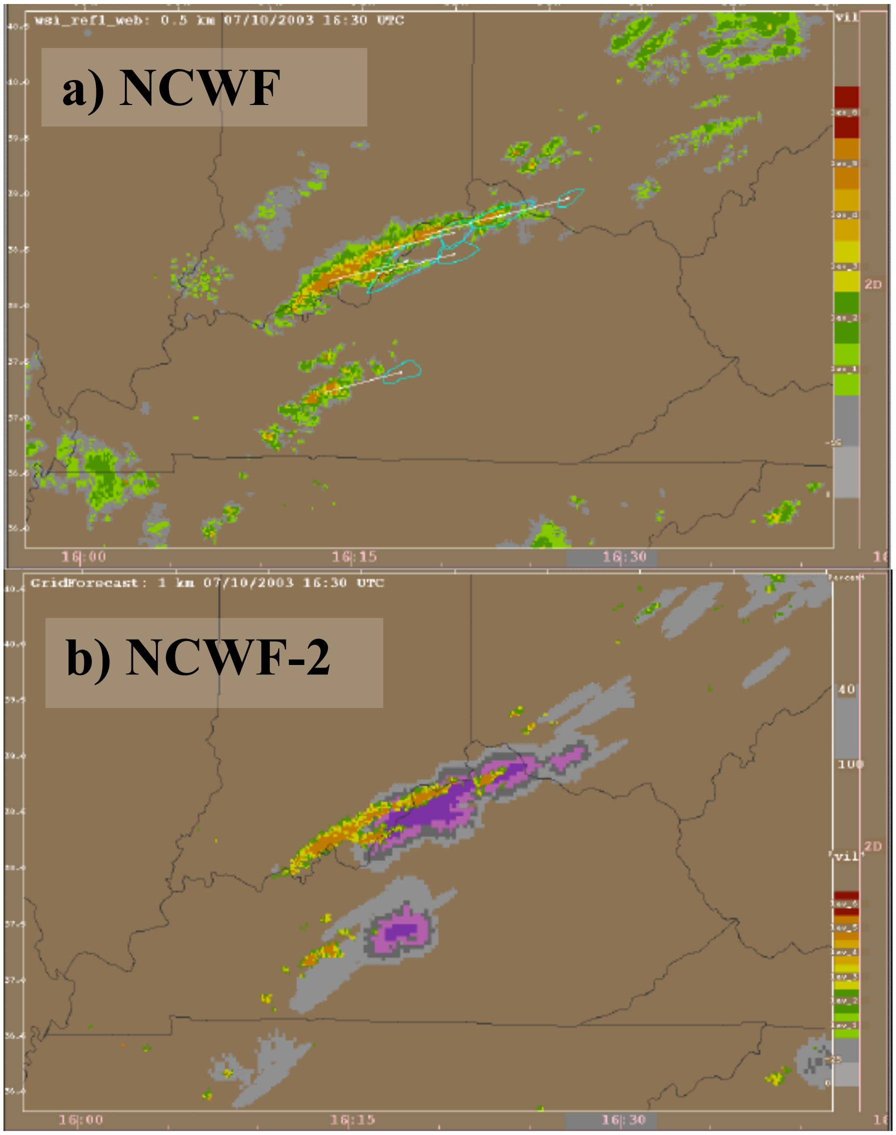

Figure 1. Examples of the 1-h forecast produced by (a) NCWF and (b) NCWF-2. In Fig. 1a, the cyan polygons indicate the 1-h NCWF forecast position overlaid on the observed NCWD reflectivity field. Fig. 1b shows the NCWF-2 probabilities along with a thresholded NCWD reflectivity field.

The probabilistic forecasts produced by NCWF-2 are based on work by Turner et al. (2004). The probabilities are calculated by determining the areal coverage of convection within an elliptical filter region. The 60-km elliptical filter is rotated at 10-degree intervals to determine the orientation with the maximum area coverage. The maximum area coverage calculated within the elliptical filter is mapped as the probability level. The NCWF-2 forecast probabilities are extrapolated using storm motion vectors. In addition to extrapolating the convection, the NCWF-2 also attempts to capture regions of growth using output from the Rapid Update Cycle numerical weather prediction model (RUC; Benjamin et al. 1998), along with radar trending and climatological considerations.

3.1.2. NCWF

The NCWF, developed by researchers in the Convective Weather Product Development Team (Wolfson and Mueller, 2006; Mueller et al. 1999), provides a depiction of current convective hazards and 1-h extrapolation forecasts of thunderstorm hazard locations. The hazard field and forecasts are updated every five minutes. The NCWF is primarily designed to be used by aviation weather forecasters, airline dispatchers, general aviation users, and FAA Traffic Management Units (TMU). The NCWF became an official NWS convective forecast product on 20 September 2001.

3.1.3. NCWD Persistence

Persistence is used as a baseline for measuring forecast skill. In this evaluation, persistence is derived from the NCWD observations by projecting the observations to the appropriate valid time. Verification statistics for persistence provide a measure of performance attained by a naïve forecast. Clearly, the convective forecasts should perform better than the persistence, showing that they add skill over a simply derived forecast.

3.1.4. CCFP

The CCFP is a convective forecast that is prepared through a collaborative process (Weather Applications Workgroup 2003; Hudson and Foss 2002; Phaneuf and Nestoros 1999) that begins with AWC forecasters, but includes participation from airline meteorologists and dispatchers, as well as meteorologists from the Center Weather Service Units (CWSUs) at the Air Route Traffic Control Centers (ARTCCs). The CCFP is used as a strategic decision aid by decision-makers at the airlines, and at the Air Traffic Control System Command Center (ATCSCC) for rerouting air traffic around convective weather. It is issued every 2 hours, with lead times of 2-, 4-, and 6-h. Minimum requirements for the issuance of a CCFP forecast polygon include an area of at least 3,000 mi^2 with convective coverage (40 dBZ or greater reflectivities) of at least 25%, and also coverage of at least 25% with echo tops of 25,000 ft and higher. There are three possible coverage thresholds for CCFP forecasts: sparse (25- 49% coverage within a polygon), moderate (50-74%), and solid (75-100%) (Weather Applications Workgroup, 2005).

3.1.5. C-SIGMET

The C-SIGMET is a polygon-based text forecast that is generated by AWC forecasters for areas of convective activity. The forecast is issued hourly and is valid for up to 2 h (National Weather Service 1991). The forecasts are issued to capture severe or embedded thunderstorms and their hazards (*e.g.*, hail, high winds) that are either occurring or forecast to occur within 30 min. of the valid period. C-SIGMETs are also issued for thunderstorm lines and areas of active thunderstorms affecting at least 40% of the forecast area that is at least 3,000 mi^2.

3.2. Observations

The NCWD is a convective hazard field depicting areas of convective weather that are deemed to be hazardous to aviation. The hazard field is based on WSR-88D National Radar Mosaics and National Lightning Detection Network cloud-to-ground lightning data (Orville 1991). Echo top data are used to threshold radar-derived Vertically Integrated Liquid (VIL) observations. The VIL data are provided in the WSR-88D product stream and are mapped to a national mosaic by the UNISYS Corporation. The VIL observations provide information about the intensity of a storm throughout its vertical extent, and serves as a proxy for vertical storm development. VIL values are translated to a VIP (Video Integrator and Processor) scale. VIP values of 3 and greater are used to assess the quality of the convective forecasts evaluated in this study.

4. Methodology

4.1. Spatial Matching of Forecasts to Observations

The convective forecasts and the verifying observations were compared on a perfectly matched, 4-km grid. Since the NCWD observations and the NCWF-2 had native grid resolutions of 4-km, no further re-mapping was necessary for these products. The NCWF, C-SIGMET, and CCFP products, which consist of forecast polygons, were mapped to the 4-km grid by the following approach. If any part of a forecast polygon intersected a box on the 4-km grid, that box was labeled with a *Yes* forecast. If a forecast polygon did not intersect the grid box, then a *No* forecast was assigned to that box. All forecasts were verified using the observation file that is closest in time to the valid time, which is typically within five minutes of the forecast valid time.

4.2. Statistical Verification Methods

By overlapping the *Yes/No* forecast and observation grids for a particular forecast, each grid point is assigned a forecast/observation designation (*i.e.*, YY, YN, NY, NN). The *Yes/No* forecast/observation pairs are then counted to fill in a 2x2 contingency table (Table 1). For instance, for a given forecast, all of the grid boxes with a *Yes* forecast and a *Yes* observation are counted to obtain the YY count (a hit); all of the grid boxes with a *Yes* forecast and a *No* observation are counted to obtain the YN count (a false alarm); and so on. Individual forecast contingency tables are accumulated to obtain tables representing particular days, months, or other period, including the entire forecast period. Additional statistical information is available via the RTVS web site (www-ad.fsl.noaa.gov/fvb/rtvs/; link to convection).

Table 1. Basic 2x2 contingency table for evaluation of dichotomous (*i.e.*, *Yes/No*) forecasts. Elements in the cells are the counts of forecast-observation pairs.

Forecast	***Observation***		***Total***
	Yes	***No***	
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 that are included in the evaluation. In this report, the forecast accuracy is expressed mainly by the CSI, PODy, and bias. General descriptions of these statistics include the following:

- PODy and PODn are estimates of the proportions of *Yes* and *No* observations, respectively, that were correctly forecast (*e.g.*, Brown et al. 1997).
- FAR is the proportion of *Yes* forecasts that were incorrect.
- Bias is the ratio of the number of *Yes* forecasts to the number of *Yes* observations, and is a measure of over- or underforecasting.
- The Critical Success Index (CSI), also known as the Threat Score, is the fraction of forecast or observed events that were correctly forecast (Wilks 1995). The CSI is a function of both PODy and FAR.
- The True Skill Statistic (TSS) (*e.g.*, Doswell et al. 1990) is a measure of the ability of the forecast to discriminate between *Yes* and *No* observations; TSS is also known as the Hanssen-Kuipers discrimination statistic (Wilks 1995).
- The Heidke Skill Score (HSS) is the percent correct, corrected for the number expected to be correct solely by chance.

Table 2. Dichotomous 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
HSS	[(YY+NN)-C1]/(N-C1), where N=YY+YN+NY+NN C1=[(YY+YN)(YY+NY) + (NY+NN)(YN+NN)] / N	Heidke Skill Score

4.3. Deterministic and Probabilistic Conversions

Probabilistic forecasts express the probability of occurrence of an event. In contrast, deterministic forecasts predict whether the event will occur or will not occur, without providing any information that describes the degree of uncertainty associated with the forecast. In this study, the forecasts are evaluated both probabilistically and deterministically, since the NCWF-2 is a probabilistic forecast, and the NCWF, C-SIGMET, and CCPF are deterministic forecasts. Methods for how this is done are described in this section.

4.3.1. Converting deterministic forecasts to probabilistic forecasts

The deterministic C-SIGMET, CCFP, and NCWF forecasts were converted into probabilistic forecasts by interpreting the coverage attributes of each deterministic forecast as a probabilistic threshold. In the case of the C-SIGMET, the forecast threshold for convective probability is 0.4. For the CCFP, the probability thresholds are 0.37, 0.62, and 0.875, which represent the mid-points of the forecast coverage ranges defined for the CCFP. NCWF forecast polygons indicate 100% convective coverage.

4.3.2. Converting probabilistic forecasts to deterministic forecasts

The probabilistic NCWF-2 forecast was converted into a deterministic forecast by selecting a representative threshold(s) from the NCWF-2. However, each choice of threshold probability produces different statistical information, as shown in Figs. 2 and 3. In general, lower thresholds result in larger PODy (Fig. 2) and bias (Fig. 3) scores, while higher thresholds result in smaller PODy and bias scores. Users may choose to select different thresholds depending upon the particular forecast application (*e.g.* for interest in severe events only, higher thresholds may be selected), but for this study, a threshold of 0.4 was chosen because the scores produced at a 0.4 threshold maximize the overall performance statistic (*i.e.*, CSI), which combines information about both PODy and bias. The NCWF-2 statistics at the 0.4 threshold were most similar to those produced for the operational NCWF, and this produced the best overall skill for the NCWF-2.

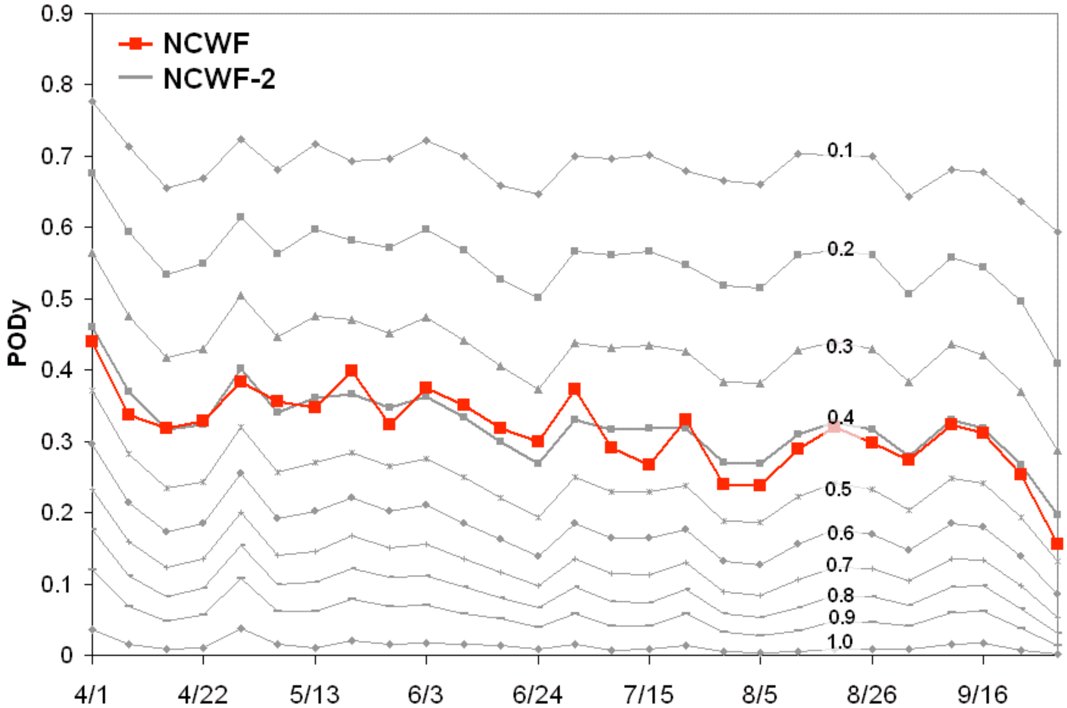

Figure 2. Weekly time series of PODy for the 1-h NCWF (bold square) and all threshold probabilities of NCWF-2 for the period 1 April – 30 September 2005.

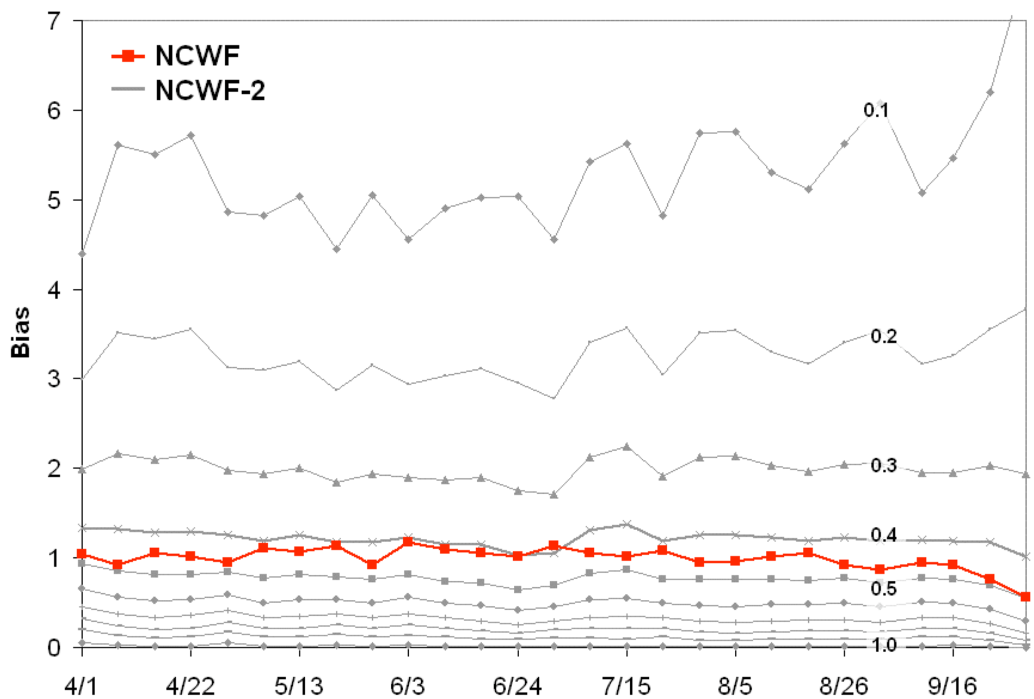

Figure 3. Weekly time series of bias for the 1-h NCWF (bold square) and all threshold probabilities of NCWF-2 for the period 1 April – 30 September 2005.

5. Results

In this section, the NCWF-2 is first assessed for all lead times and threshold probabilities to give a basic understanding of how the probabilistic forecast performs. Next, NCWF-2 is compared to other operational convective forecasts both probabilistically and deterministically at 1-h and 2-h lead times. At the 1-h lead time, the NCWF-2 is compared to the NCWF, persistence, and the C-SIGMETs. At the 2-h lead time, the NCWF-2 is compared to the CCFP and the C-SIGMETs.

5.1. Assessment of NCWF-2

In this section, a probabilistic assessment of NCWF-2 is presented and stratified by forecast lead time. A reliability diagram is used to describe the "calibration" of the forecast. A probabilistic forecast is well-calibrated, or reliable, when the observed coverage values match the forecast probability. For example, 30% probabilities should verify 30% of the time, 60% probabilities should verify 60% of the time, etc. The diagram shows the observed coverage as a function of the forecast probability. Perfect reliability is delineated by the y=x diagonal line on the graph. Forecast points plotted above the y=x diagonal line indicate underforecasting, while those plotted below the diagonal indicate overforecasting. The forecast resolution refers to whether the forecast can resolve the event better than the climatological observed average, which is estimated to be on the order of one percent for the evaluation period. The forecast resolution is graphically represented as the vertical difference between the forecast points and the "no resolution" line, which may be plotted as a horizontal line representing the climatological observed average. The forecast frequency is displayed as a histogram inset, showing the forecast distribution that is independent of the observations (Wilks 1995). This inset provides important information needed in addition to the reliability diagram, indicating the frequency of forecast probability values issued most and least often.

The reliability diagram for the NCWF-2 for the four 30-min lead times is shown in Fig. 4. The 30-min forecast ('diamonds') appears to be the most reliable of the four forecast lead times as indicated by the closer proximity of the 30-min line to the y=x diagonal line. Overall, the forecast reliability decreases as the lead time increases, as suggested by the increase in vertical distance for the other forecast leads from the diagonal line. At the lowest probabilities, the 30-min forecast is distinctly more reliable than the other forecasts. Overforecasting increases with lead time, as indicated by the increasing distance from the diagonal as lead time increases. The forecast resolution is best for the 30-min lead time and is notably better for the higher probabilities, as indicated by the increased distance from the horizontal "no resolution" line, which falls approximately at y=0.01. As lead time increases, the forecast resolution becomes poorer. The difference in reliability between the 30-min forecast and those with lead times greater than 60 minutes is evident for all probabilities. Thus, the 30-min forecast is more reliable than the forecasts with longer lead times.

The inset shows a histogram of the issuance frequency for all probabilities. The diagram shows that lower probabilities are forecast more frequently than higher probabilities, especially at the 120-min lead time.

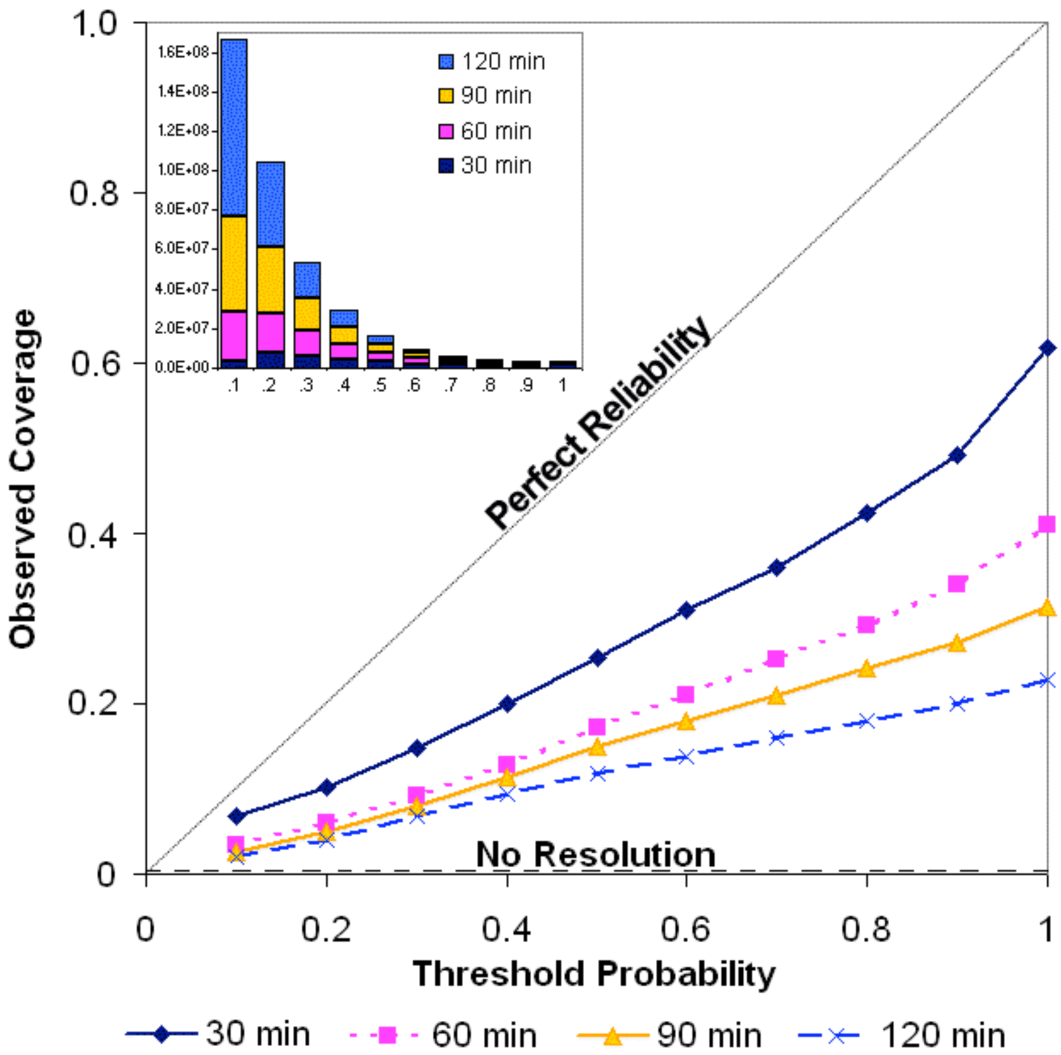

Figure 4. Reliability diagram for NCWF-2 for lead times of 30, 60, 90, and 120 min. The histogram inset shows the frequency of issuance of probabilities, which were binned in 0.1 increments from 0.1 to 1.0. No resolution line was qualitatively estimated.

Figure 5 shows the behavior of dichotomous verification statistics for the 1-h NCWF-2 for all probability thresholds. The largest bias, PODy, TSS and FAR scores occur at the lowest NCWF-2 probability threshold (0.1). The largest bias indicates overforecasting, which contributes to the large PODy and FAR. As the threshold increases, the bias, PODy, TSS and FAR scores decrease. The bias is closest to one near the 0.4 probability threshold, which indicates a minimum of over- and underforecasting. The CSI and HSS curves show a slight peak at the 0.4 probability threshold. Since many of the scores for the NCWF-2 are optimized at the 0.4 probability threshold, this value is mainly used to compare forecast quality to the other dichotomous forecasts, which will be summarized in later sections of this report.

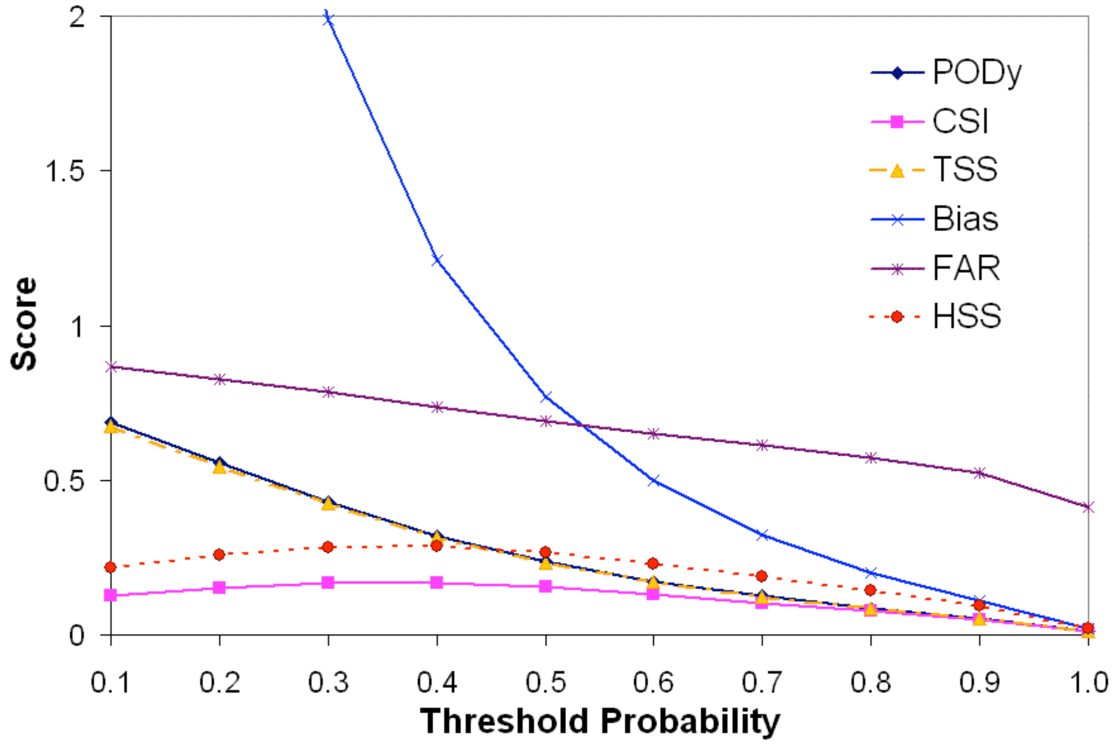

Figure 5. Diagram showing several dichotomous statistics for the 1-h NCWF-2 for all threshold probabilities.

5.2 Comparison of 1-h operational forecasts

5.2.1 Probabilistic assessment

This section includes a probabilistic assessment of the NCWF-2, as compared to probabilistic assessments of the NCWF and the C-SIGMET. The reliability diagram, shown in Fig. 6, is used to summarize the calibration of the 1-h forecasts from the NCWF-2, the NCWF, and the C-SIGMET. The fundamental characteristics of the diagram are described in detail in Section 5.1. The 25th and 75th percentiles for each forecast are also plotted to illustrate the sharpness of the distribution, as indicated by the light blue band surrounding the reliability line for the NCWF-2, and the bars that are associated with the C-SIGMET and the NCWF. Since the NCWF and C-SIGMET are inherently deterministic forecasts, only one threshold is plotted for each of these forecasts.

The results shown in Fig. 6 indicate that the NCWF-2 probabilistic forecasts are more reliable than both the probabilistic C-SIGMET and the NCWF, as suggested by the closer proximity of the NCWF-2 line to the diagonal line. All products overforecast, as indicated by the curves and points falling below the “perfect reliability” line. The C-SIGMET and NCWF both overforecast to a greater degree than does the NCWF-2. The plot indicates that the NCWF-2 shows modest resolution and reliability as suggested by its position relative to the ‘perfect reliability’ and ‘no resolution’ lines. Also, note that the sharpness of the NCWF-2 forecast decreases as the probability increases, as indicated by the width of the distribution. The C-SIGMET sharpness is similar to the NCWF-2, but the NCWF is distinctly sharper than the NCWF-2 because it has a more narrow distribution. Although the 1-h probabilistic forecast from the NCWF-2 is not entirely reliable, these results suggest that the NCWF-2 performs better than the 1-h forecasts from the C-SIGMET and the NCWF.

The inset shows a histogram of the issuance frequency for all probabilities. It shows that NCWF-2 forecasts are issued most frequently at the lowest probabilities and are issued less frequency as probability increases.

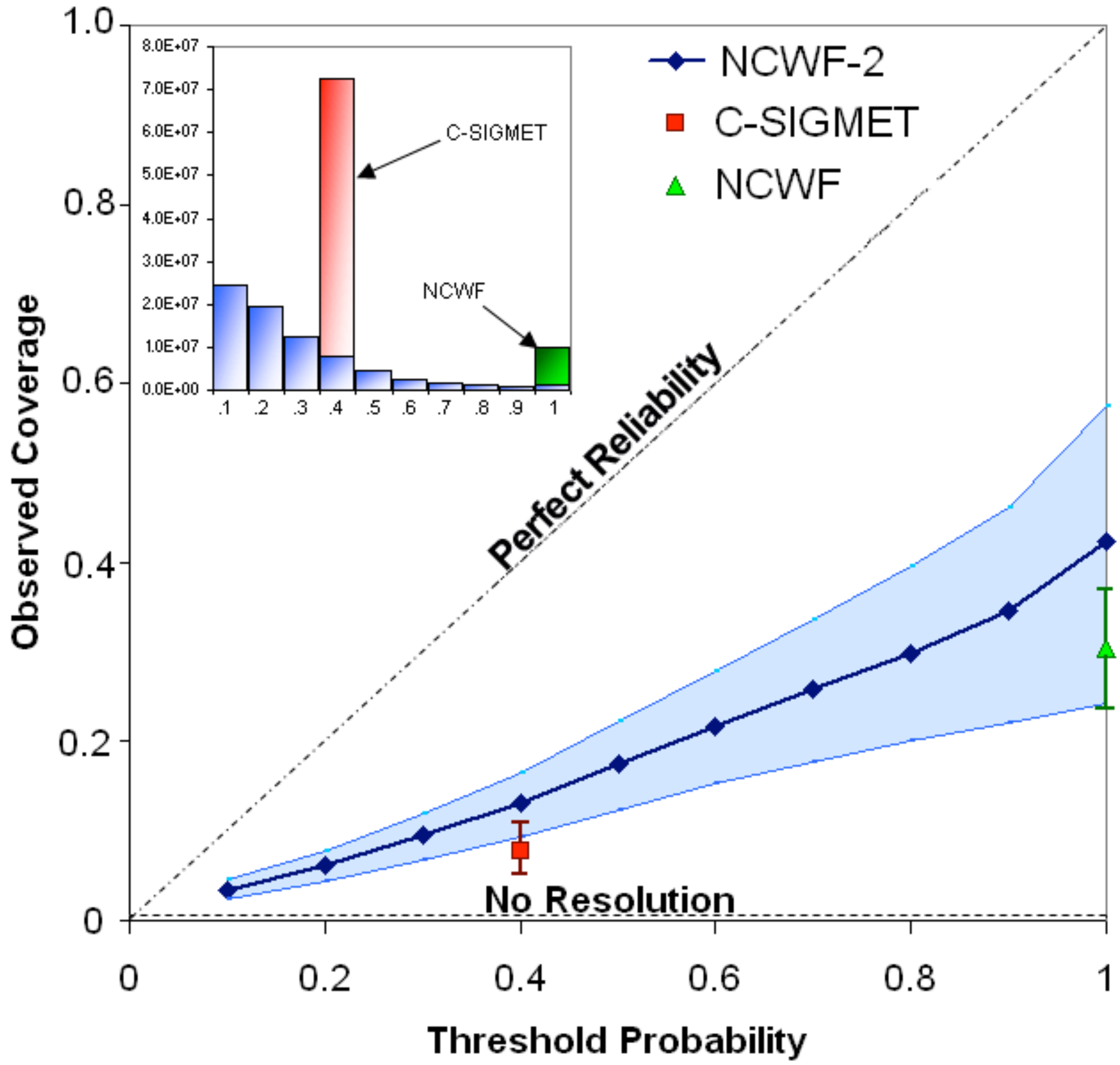

Figure 6. Reliability diagram for 1-h NCWF-2, NCWF, and C-SIGMET forecasts for the period 1 April – 30 September 2005. The histogram inset shows the frequency of issuance of NCWF-2 forecast probabilities along with those of C-SIGMET and NCWF, which were binned in 0.1 increments from 0.1 to 1.0. The "no resolution" line was qualitatively estimated.

Figure 7 shows box plots of the observed percent coverage contained within a forecast area for probability thresholds 0.4 and 1.0. From top to bottom, the whiskers on the box plots indicate the 95th, 75th, 50th (*i.e.*, median), 25th and 5th percentiles. The notches represent approximate 95% confidence intervals for the median values, which provides a way to measure the uncertainty in the median values. A very accurate forecast would have observed coverage values close to the respective forecast probabilities. The median value for the NCWF-2 is greater than the median value for the C-SIGMET, indicating that it is more accurate than the operational product. For a threshold of 1.0, the median of the NCWF-2 is much greater than that of the NCWF, but is still significantly lower than the expected coverage. The difference between the 25th and 75th percentiles of the NCWF-2 indicates a wider distribution of observed coverages for NCWF-2 forecasts than for the operational NCWF forecasts; that is, the observed coverage values for NCWF-2 are more variable.

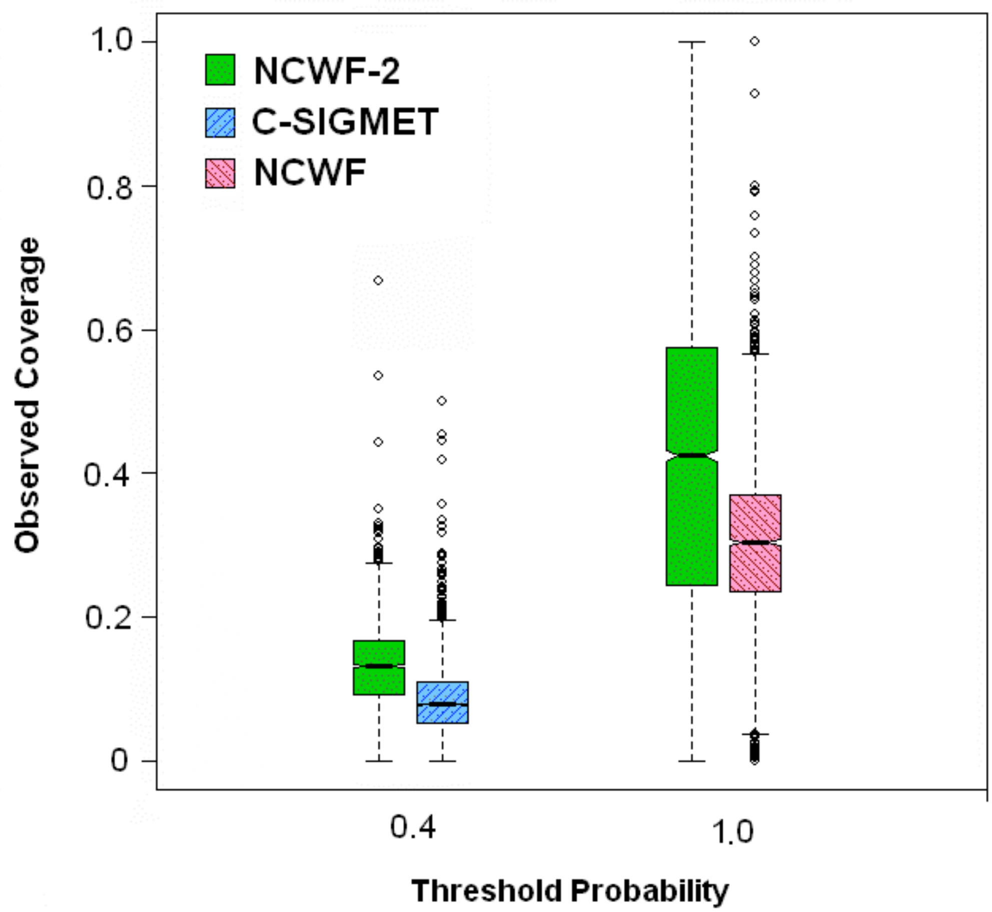

Figure 7. Box plots of the 1-h NCWF-2, NCWF, and C-SIGMET observed coverage values for two probabilities, for the period 1 April – 30 September 2005.

5.2.2 Deterministic assessment

This section summarizes the deterministic assessment of the NCWF-2, the NCWF, persistence, and the C-SIGMET for the 1-h forecasts. NCWF-2 was assessed deterministically for this evaluation using only one forecast threshold. Results for all thresholds can be obtained from the RTVS web site. Forecast accuracy, which measures how similar the forecasts are to the observations, is expressed primarily in terms of the Critical Success Index (CSI). The CSI describes the percentage of forecast or observed events that are correctly forecast, with higher scores representing more skill. Scores for

PODy and bias are also presented. It is important to remember that the NCWF-2 is inherently a probabilistic forecast, but for this section of the evaluation, the NCWF-2 is interpreted as a deterministic forecast.

Weekly CSI scores for the analysis period are shown in Fig. 8 for the 1-h forecasts. Results are quite similar for the NCWF and the NCWF-2, with typical values between 0.15 and 0.25. The CSI scores for the C-SIGMET and 1-h persistence are nearly 50% lower than those computed for the NCWF and NCWF-2. These results suggest that both NCWF products improve upon the persistence and C-SIGMET forecasts.

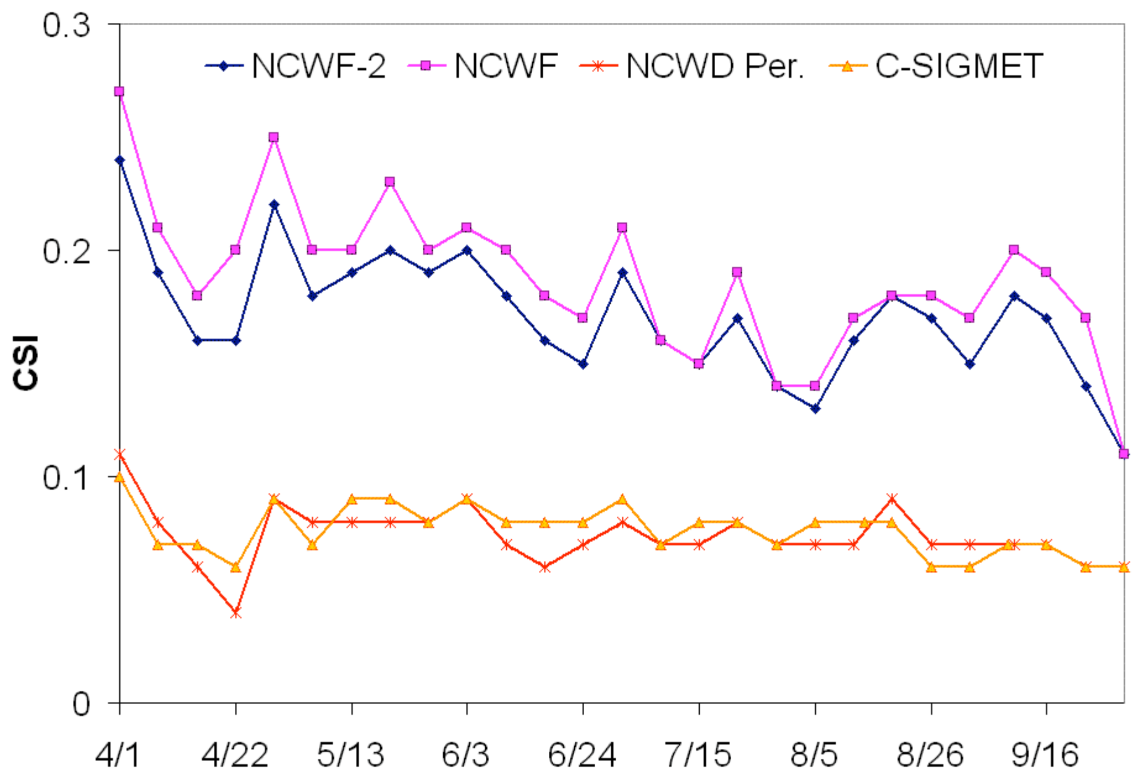

Figure 8. Weekly time series of CSI for the 1-h NCWF-2, NCWF, NCWD Persistence, and C-SIGMET for the period 1 April – 30 September 2005.

Figure 9 shows a weekly time series of PODy for the 1-h NCWF-2, the NCWF, persistence, and the C-SIGMET for the 2005 convective season. The NCWF-2 and NCWF perform similarly for the entire convective season, with fairly consistent scores that range between 0.3 and 0.5. Small PODy scores for persistence are evident and are typically at or below 0.2. The C-SIGMET has considerably larger PODy scores than the other forecasts, which range between 0.5 and 0.7, but those scores are often achieved at the expense of large bias values.

The bias statistic is used to describe the degree of over- or underforecasting of the convective forecasts. Figure 10 shows weekly bias scores for all forecasts except the C-SIGMET, which had such large biases (ranging from six to ten) that it was not included on the plot. The 0.4 threshold of NCWF-2 shows slightly larger bias values than the NCWF, indicating that it overforecasts more than the operational product. The persistence shows a bias of nearly 1.0 over the entire season.

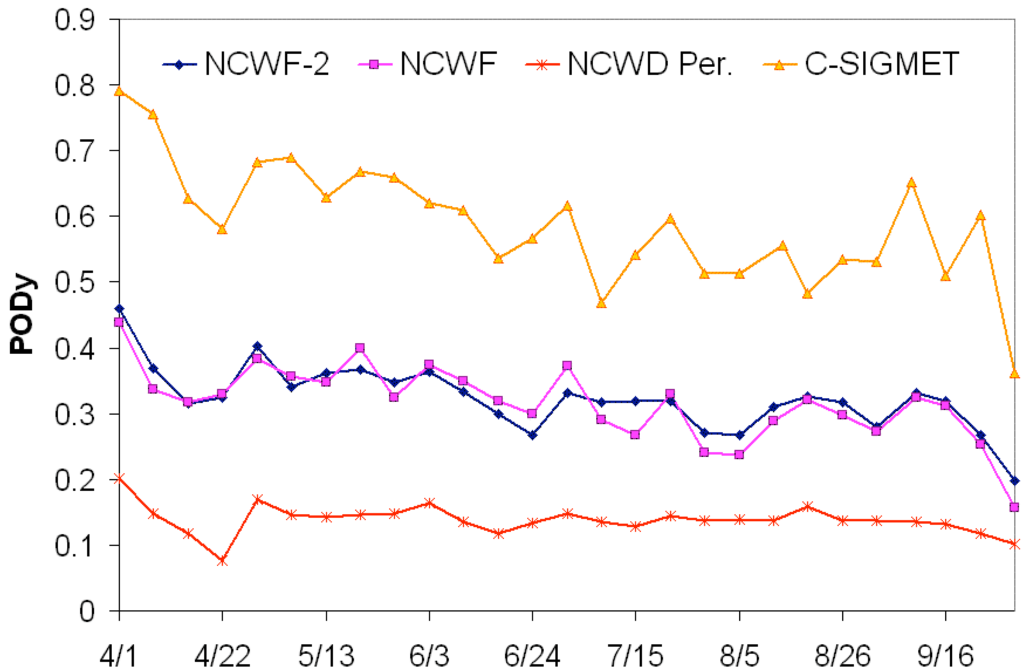

Figure 9. Weekly time series of PODy for the 1-h NCWF-2, NCWF, NCWD Persistence, and C-SIGMET for the period 1 April – 30 September 2005.

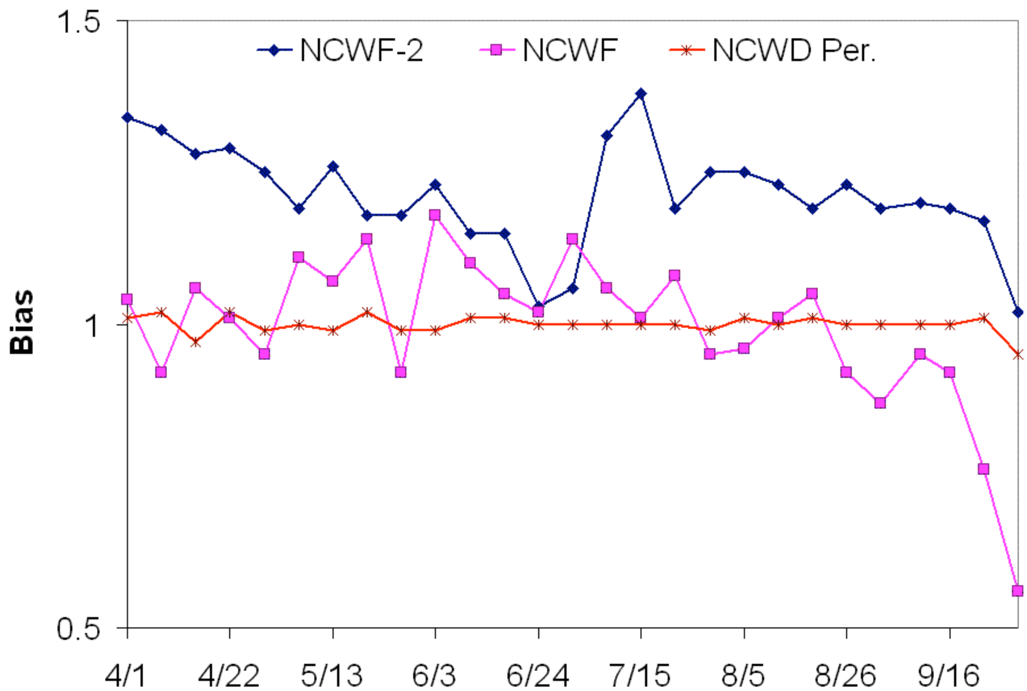

Figure 10. Weekly time series of bias for the 1-h NCWF-2, NCWF, NCWD Persistence, and C-SIGMET for the period 1 April – 30 September.

5.3 Comparison of 2-h operational forecasts

5.3.1 Probabilistic assessment

This section includes a probabilistic assessment of the 2-h NCWF-2, the C-SIGMET, and the CCFP forecasts. In order to compare these forecasts, the C-SIGMET and the CCFP were converted to probabilistic forecasts as described in Section 4.3.

Figure 11 shows a reliability diagram for the 2-h NCWF-2, the C-SIGMET, and the CCFP. The 25th and 75th percentiles for each forecast are also plotted to illustrate the sharpness of the distribution, as indicated by the light blue band surrounding the reliability line for the NCWF-2, and the bars that are associated with the C-SIGMET and the CCFP. As described in Section 4, the CCFP and the C-SIGMET are inherently deterministic forecasts and therefore their derived probabilistic attributes are not shown over the full range of thresholds used to categorize the NCWF-2. As a reminder, the probabilities assigned to CCFP coverage categories are 0.37, 0.62 and 0.875, which do not exactly match those defined for the NCWF-2.

The line plotted for NCWF-2 on the diagram shown in Fig. 11 is closest to the y=x diagonal (*i.e.*, perfectly reliable forecast), which indicates that the NCWF-2 is more reliable than both the C-SIGMET and the CCFP at producing probabilistic convective forecasts. The largest difference in reliability between the CCFP and NCWF-2 occurs at the lowest probability of 0.4 (NCWF-2) and 0.37 (CCFP). As the probability increases, the reliability of the CCFP approaches that of the NCWF-2. All forecast points lie well below the 'perfect reliability' line, indicating that all products overforecast. The C-SIGMET and the CCFP show larger biases than the NCWF-2. The results presented in the plot suggest that the NCWF-2 shows modest resolution and reliability, as suggested by its line's position relative to the 'perfect reliability' and 'no resolution' lines. The CCFP has slightly lower reliability and resolution than the NCWF-2. While the resolution and reliability of the C-SIGMET cannot be fully determined in this analysis, the results do appear to indicate that the NCWF-2 performs slightly better than both operational products. Also, note that the sharpness of the NCWF-2 decreases as the probability increases, as indicated by the width of the distribution. The C-SIGMET and CCFP are somewhat sharper forecasts than the NCWF-2, particularly at the higher probabilities.

The histogram inset shows that NCWF-2 forecasts are issued most frequently at the lowest probabilities. Relative to the lower probabilities, the higher probabilities are not forecast as often.

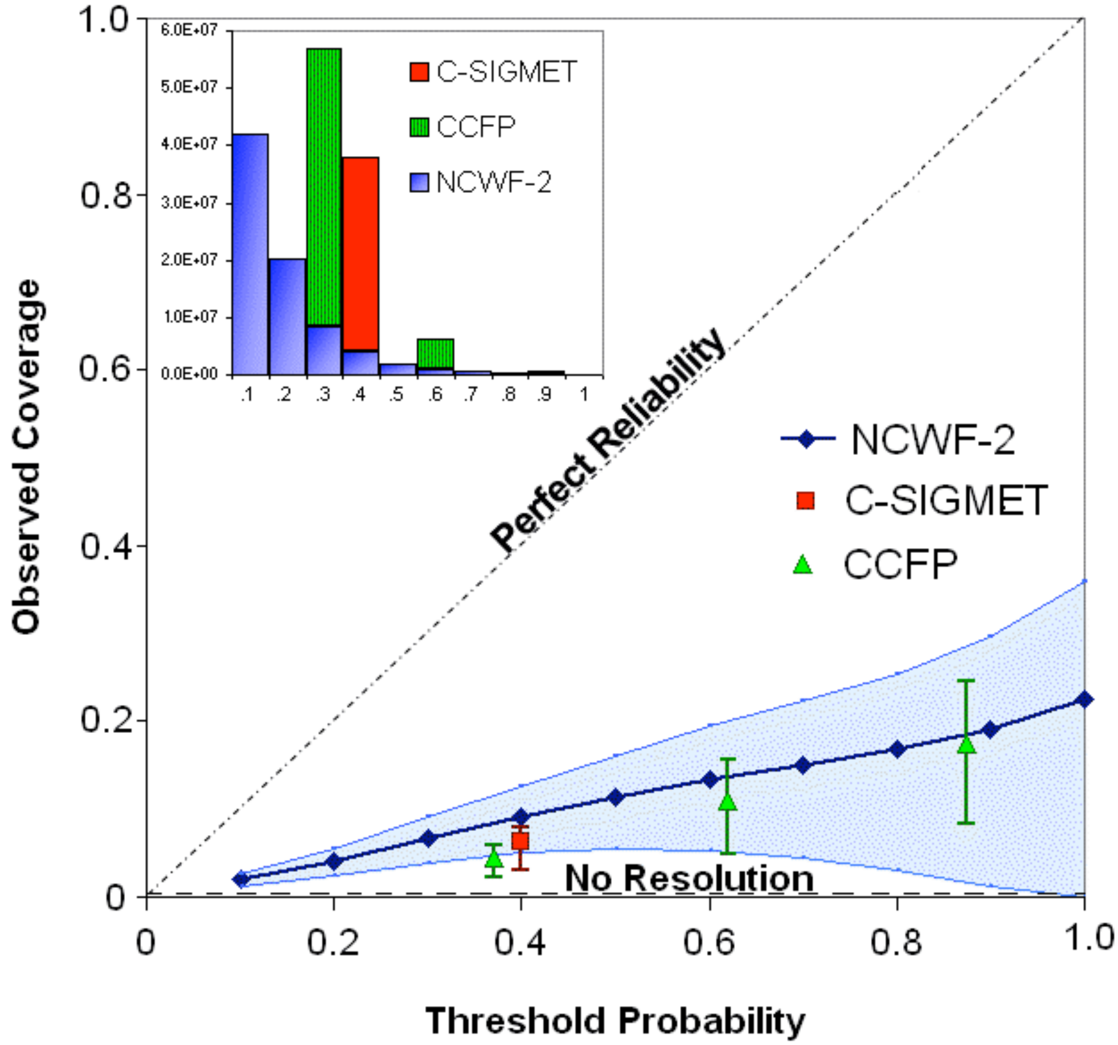

Figure 11. Reliability diagram for 2-h NCWF-2, CCFP, and C-SIGMET forecasts for the period 1 April – 30 September 2005. The histogram inset shows the frequency of issuance of forecast probabilities, which were binned in 0.1 increments from 0.1 to 1.0. The "no resolution" line was qualitatively estimated.

Figure 12 shows box plots of observed percent coverage for the 0.4, 0.7, and 0.9 probabilities for NCWF-2, the 0.4 probability assigned to C-SIGMETs, and the 0.37, 0.67 and 0.875 probabilities assigned to the CCFP coverage categories. Perfectly reliable forecasts would have percent coverage values that correspond to the probability. For example, at a probability of 0.4, the NCWF-2 should contain 40% of the convection within the forecast area. However, none of the three forecasts captured the correct amount of the observed coverage. Nevertheless, some indications of improved forecast coverage accuracy can be identified for the NCWF-2. For instance, at the 0.4 probability, the NCWF-2 median is higher than the C-SIGMET and CCFP medians. These results suggest that the NCWF-2 is more reliable at the lower probability than the other forecast products. At the 0.7 probability, the difference between the medians of the NCWF-2 and the CCFP is smaller than the difference at the lower probability, but the NCWF-2 median remains higher than the CCFP median. However, at the 0.9 value, the CCFP median moves above that of the NCWF-2. A closer inspection of the NCWF-2 at probabilities 0.7 and 0.9 reveals that the median values are the same for both values, but that the distribution of the percent coverage widens at the 0.9 probability. This result suggests that although the NCWF-2 is more efficient at capturing the correct coverage, it does lose some precision at the highest probabilities as is indicated by the wider distribution of percent coverage.

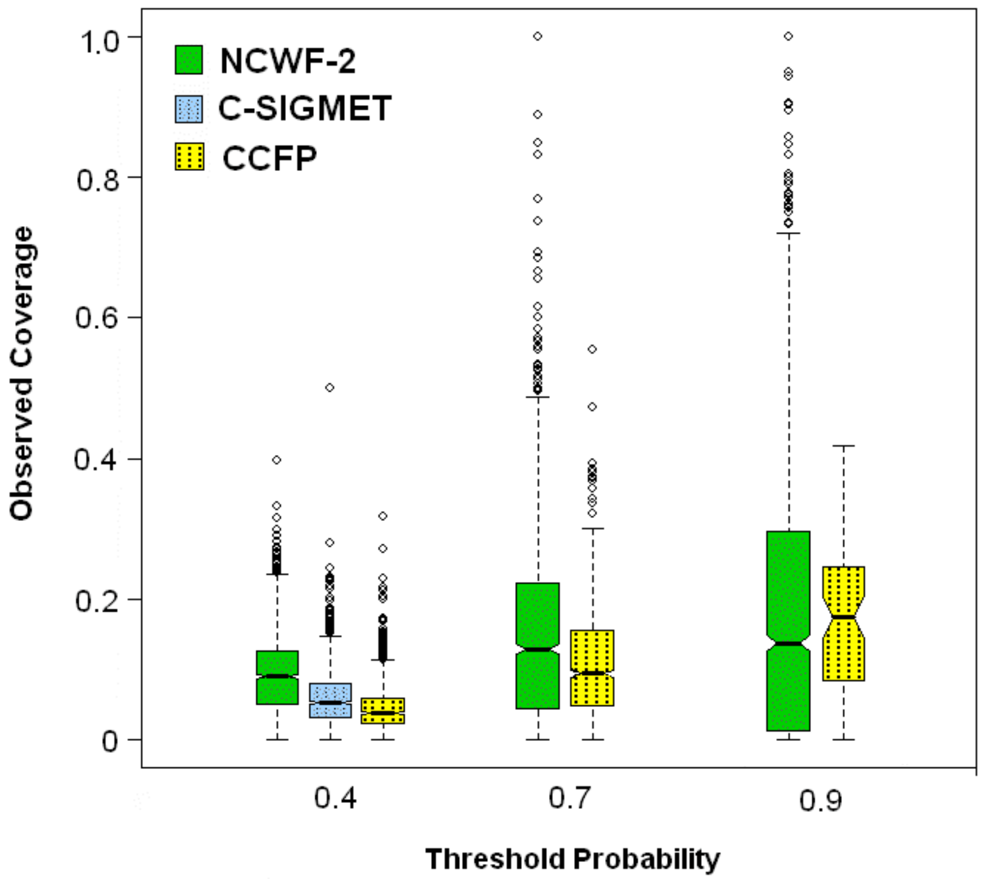

Figure 12. Box plots for 2-h NCWF-2, CCFP, and C-SIGMET observed coverage values for three probabilities, for the period 1 April – 30 September 2005.

5.3.2 Deterministic assessment

This section summarizes the deterministic assessment of the NCWF-2, the CCFP, and the C-SIGMET for the 2-h forecast. NCWF-2 was assessed deterministically for this evaluation using only one forecast threshold, although results for all thresholds can be obtained from the RTVS web site. Forecast accuracy, which measures how similar the forecasts are to the observations, is expressed primarily in terms of the Critical Success Index (CSI), but scores for PODy and bias are also presented. It is important to remember that the NCWF-2 is inherently a probabilistic forecast, but for this section of the evaluation, the NCWF-2 is interpreted as a 2-h deterministic forecast.

Figure 13 shows the weekly CSI for the 2-h forecasts. Recall from Section 5.2.2 that the CSI represents the fraction of forecast or observed events that are correctly forecast, with higher scores showing more skill. The NCWF-2 performs better than the other two forecasts throughout the entire convective season, with CSI scores ranging from 0.006 to 0.16. However, from late June through early August, the CSI scores for the NCWF-2 are closer to those of the C-SIGMET and CCFP. Nevertheless, the CSI scores for the CCFP are lower than either the NCWF-2 or the C-SIGMET for the majority of the time period.

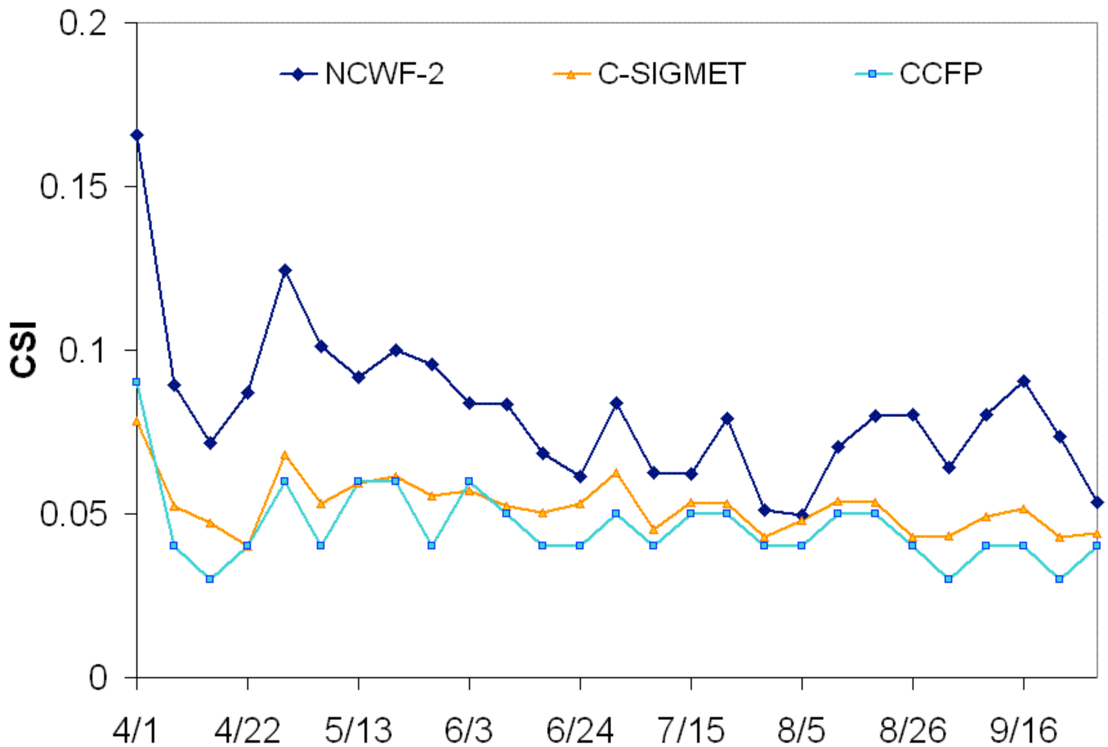

Figure 13. Weekly time series of CSI for the 2-h NCWF-2 (0.4 threshold), CCFP, and C-SIGMET for the period 1 April – 30 September 2005.

Figure 14 shows a weekly time series of PODy for the 2-h NCWF-2, CCFP and C-SIGMET for the 2005 convective season. It is immediately apparent from the results in Fig. 14 that the PODy scores for the NCWF-2 are substantially lower than those of the C-SIGMET and the CCFP. For instance, for NCWF-2, the PODy values range between nearly 0.3 in early April to values of 0.1 in late September. The PODy values for the CCFP and the C-SIGMETs are typically between 0.4 and 0.8. The elevated PODy scores for the C-SIGMET and the CCFP are achieved at the expense of a large bias as shown in Fig. 15. Bias values for the C-SIGMET are generally 4 to 10 times the value of the bias of the NCWF-2. The CCFP biases are even larger, with values ranging from approximately 7 to 17.

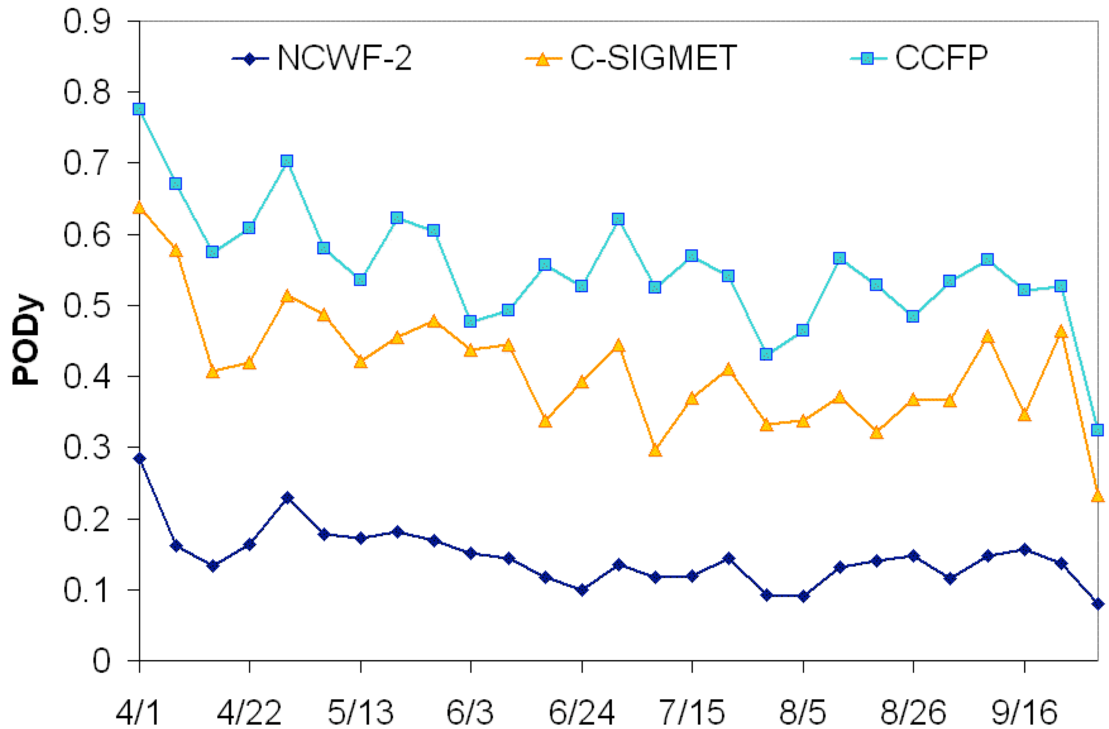

Figure 14. Weekly time series of PODy for the 2-h NCWF-2 (0.4 threshold), CCFP, and C-SIGMET for the period 1 April – 30 September 2005.

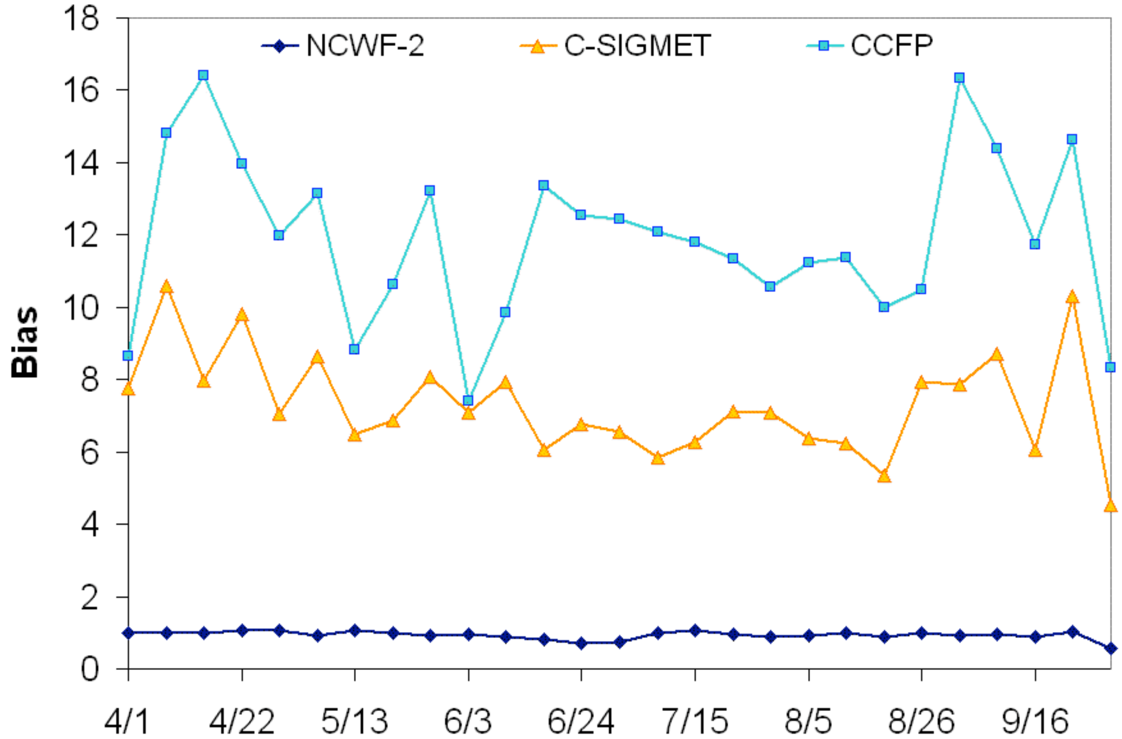

Figure 15. Weekly time series of bias for the 2-h NCWF-2 (0.4 threshold), CCFP, and C-SIGMET for the period 1 April – 30 September 2005.

6. Conclusions

This report has summarized the forecast skill of the NCWF-2 for the period 1 April – 30 September 2005 in comparison to other operational convective forecast products. Results from this assessment indicate that:

- Overall, the NCWF-2 shows modest reliability and resolution. The reliability and resolution decrease with lead time, with distinctly better performance associated with the 30-min forecasts, in comparison with forecasts at longer lead times.
- The 1-h probabilistic comparison shows that NCWF-2 performs better than the operational NCWF and C-SIGMET with respect to reliability and resolution.
- The 1-h deterministic comparison shows that NCWF-2 at the 0.4 threshold performs similarly to the operational NCWF and better than the C-SIGMET, as indicated by CSI statistics.
- The 2-h probabilistic comparison shows that NCWF-2 performs better than the CCFP and C-SIGMET, with respect to reliability and resolution. The reliability of the CCFP approaches that of the NCWF-2 at higher probabilities.

- The 2-h deterministic comparison shows that NCWF-2 at the 0.4 threshold performs better than the CCFP and C-SIGMET, as indicated by CSI statistics.

Acknowledgements

This research is in response to requirements and funding provided 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 the members of the Convective Weather Product Development Team for their support of this independent verification effort over the last several years. We also would like to thank the verification groups at GSD and NCAR for providing outstanding support and many contributions that made this study and report possible. Finally, we would like to thank the AWRP Leadership Team for their support of the independent verification activities undertaken by the Quality Assessment Product Development Team.

References

Benjamin, S.G, J.M. Brown, K.J. Brundage, B.E. Schwartz, T.G Smirnova, T.L. Smith, 1998: The operational RUC-2. *Preprints, 16th Conference on Weather Analysis and Forecasting*, Amer. Meteor. Soc., Phoenix, AZ, 249-252.

Brown, B.G. and J.L. Mahoney, 2000: Quality assessment of the National Convective Forecast Product. Quality Assessment Report provided to the Technical Review Panel of the Aviation Weather Technology Transfer Board.

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. Forecasting*, **12**, 890-914.

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.

Hudson, H. R. and F. P. Foss, 2002: The Collaborative Convective Forecast Product from the Aviation Weather Center's Perspective. *Preprints, 10th Conference on Aviation, Range, and Aerospace Meteorology*, Amer. Meteor. Soc., Portland, OR, 73-76.

Mahoney, J.L, J. K. Henderson, B.G. Brown, J.E. Hart, A. Loughe, C. Fischer, and B. Sigren, 2002: The Real-Time Verification System (RTVS) and its application to aviation weather forecasts. *Preprints, 10th Conference on Aviation, Range, and Aerospace Meteorology*, Amer. Meteor. Soc., Portland, OR, 323-326.

Mahoney, J.L., P. Kucera, J.E. Hart, and B.G. Brown, 2004: Quality Assessment Report: National Convective Weather Forecast (NCWF-2). Submitted to Aviation Weather Technology Transfer (AWTT) Technical Review Panel.

Megenhardt, Dan, C.K. Mueller, N. Rehak, and G. Cunning, 2000: *Preprints, 9th Conference on Aviation, Range, and Aerospace Meteorology*, Amer. Met. Soc., Orlando, FL.

Mueller, C.K., C.B. Fidalgo, 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*, Amer. Met. Soc., Boston, MA, 230-234.

Mueller et al. 2004: NCWF-2. Report to the AWTT Board.

NWS, 1991: National Weather Service Operations Manual, D-22. National Weather Service. (Available at this 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 AvMet Applications, Inc.).

Turner, B. J., U. Germann, and I. Zawadzki, 2004: Scale dependence of the predictability of precipitation from continental radar images. Part II: Probability Forecasts. *Jour. Appl. Met.*, **43**, 74-89.

Weather Applications Workgroup, 2003: Statement of User Needs CCFP/2003. FAA, CDM, CR-Workgroups. February 2003. 26 pp.

Weather Applications Workgroup, 2005: Statement of User needs CCFP/2005. FAA, CDM, CR-Workgroups. February 2005, 33 pp.

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

Wolfson, M.M. and C.K. Mueller, 2006: FAA AWRP Convective Weather Development Team. *Preprints, 12th Conference on Aviation, Range, and Aerospace Meteorology,* Amer. Met. Society, Atlanta, GA.