16th Conference on Probability and Statistics in the Atmospheric Sciences, 14-18 January 2002, Orlando, FL.

1.5 USING VERIFICATION TECHNIQUES TO EVALUATE DIFFERENCS AMONG CONVECTIVE FORECASTS

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

1. Introduction

With the implementation of National Convective Weather Forecast Product (NCWF) and the Collaborative Convective Forecast Product (CCFP) into National Weather Service (NWS) operations, users are striving to understand the similarities and differences between the various products, their application to aviation operations, and how each forecast fits within the context of on-going forecast products, such as the Convective SIGMET. As stated in Brooks and Doswell (1996), verification of weather forecasts is an essential part of any forecasting system. It can provide a mechanism for identifying the strengths and weakness in forecasting systems and provide a method for choosing appropriate forecasting procedures for measuring improvement.

Following Brooks and Doswell (1996), an intercomparison exercise was conducted from 1 April – 30 September 2001 to obtain an understanding of these strengths and weaknesses. In order to do so, the forecasts were compared using verification methods that were consistently applied. In this paper, we will demonstrate that by comparing various convective forecasts the differences and similarities of the forecasts are brought to light and that this approach is useful for evaluating how the forecasts can be used.

[1] Corresponding author: NOAA Forecast Systems Laboratory, 325 Broadway, Boulder, CO 30303; mahoney@fsl.noaa.gov

[2] Research Application 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.

2. Data

2.1 Forecasts

All forecasts used in this evaluation are operational forecasts supported by the NWS and are described in this Section.

Collaborative Convective Forecast Product (CCFP) – The CCFP consists of two parts: 1) a Preliminary forecast that is developed and issued by a forecaster at the NWS Aviation Weather Center (AWC) as a precursor to the Final, and 2) a Final forecast that is developed through a collaborative process that takes place between the AWC forecasters and the airline and other meteorologists. The CCFP product is generated as a graphic depicting predicted areas of convective activity valid at specific times. The CCFP is ultimately used by decision-makers for routing traffic around convective areas (Phaneuf and Nestoros 1999).

Convective SIGMET (C-SIGMET) - This product, generated by AWC forecasters, is a text forecast of convective activity. The forecast is issued hourly and is valid for up to 2 h (NWS 1991). The forecasts are issued to capture severe or embedded thunderstorms and their hazards (*e.g.,* hail, high winds) that are either occurring or forecasted to occur within 30 minutes of the valid period. C-SIGMETs are also issued for thunderstorm lines and areas of active thunderstorms affecting at least 3,000 square miles. For this evaluation, the C-SIGMETs are treated in a variety of ways; a 0-h forecast or now-cast of convective activity, a forecast of 1 h duration that is valid at the *end* of the period; a forecast of 1 h duration that is valid at the end of the period, but is corrected by a speed and direction component; a forecast of 2 h duration, valid at the *end* of the period corrected by speed and direction; and 2 h forecast, valid *throughout* the entire 2-h period.

National Convective Weather Forecast (NCWF) – The NCWF, developed by the National Center for Atmospheric Research (NCAR; 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 5 minutes. The NCWF targets airline dispatchers, general aviation users, and FAA Traffic Management Units (TMU). As of September 20, 2001 the NCWF has become an operational forecast product that is supported by the NWS.

2.2 Observations

The National Convective Weather Detection Product (NCWD; Mueller et al. 1999) was used to verify the forecasts. The NCWD combines a two-dimensional mosaic of radar reflectivity with radar-derived cloud top data and a grid of lightning detections from the National Lightning Data Network (NLDN; Orville 1991). The cloud top data are primarily used to remove anomalous propagation and ground clutter. The lightning data help to keep the NCWD current, since lightning data have a lower latency than radar data. The NCWD fields were made available on a 4-km grid, with convective storms delineated by a threshold of 40 dBZ, or more than 3 lightning strokes in 10 minutes.

3. Verification Approach and Statistics

3.1 Approach

Verification of the convective forecasts was provided in near real time by the Real-Time Verification System (RTVS; Mahoney et al. 1997). The RTVS is a verification tool, being developed by the Forecast Systems Laboratory with funds provided by the FAA Aviation Weather Research Program (AWRP). RTVS uses the most advanced verification techniques and allows users to easily compare forecasts through a Web-based graphical user interface. Only a subset of results provided by RTVS are presented in this report. Users are encouraged to access RTVS for further analyses (http://www-ad.fsl.noaa.gov/afra/rtvs; link CCFP or convection).

In general, the verification methods were applied to all convective forecasts. However, differences in time windows and verification grids were modified for each forecast to account for specific forecast attributes (Mahoney et al. 2000).

3.2 Matching methods

Before the forecasts were matched to the observations, a grid (*e.g*., 20 x 20 km grid) was overlaid on the observation field. Each box on the overlay grid was assigned a *Yes* or *No* value depending on whether a positive observation (i.e., one 4-km NCWD observation with reflectivity greater than 40 dBZ) fell within the defined grid box. The same procedures were applied to the forecasts, with a grid box labeled as *Yes* when any part of the forecast polygon intersected that box. If a forecast polygon did not intersect the grid box, then a *No* forecast was assigned to the box.

Once the matching process was completed, each grid box on the observation grid was matched to each grid box on the forecast grid. This technique produced the pairs used to generate the verification statistics. For example, a *Yes* forecast box overlapping a *Yes* observation box produced a *Yes-Yes* pair. Similarly, a *Yes* forecast and *No* observation produced a *Yes-No* pair, and so on, filling the two-by-two contingency table (e.g., Wilks 1995).

For this evaluation, the sizes of the grid boxes used to verify the forecasts were modified according to Mahoney et al. (2000) to account for scaling dependencies inherent in the forecasts. For instance, a 4-km grid was used to verify the NCWF, a 20-km grid for the C-SIGMETs, and 40-km grid for the CCFP. Similarly, the time window with which the forecasts and observations were matched was also modified to meet the needs of the individual forecasts. The time window used to verify the NCWF was the valid time, and a 20-minute time window surrounding the valid time was used to verify the C-SIGMETs, and CCFP.

3.3 Statistics

The verification methods used in this study are based on standard verification concepts that take into account the underlying statistical basis for verification, as well as the associated high dimensionality of the verification problem (*e.g*., Murphy and Winkler 1987; Brown *et al.* 1997). The primary verification statistics used in this analysis include the following:

- PODy and PODn are estimates of the proportion of *Yes* and *No* observations, respectively, that were correctly forecast.
- FAR is the proportion of *Yes* forecasts that were incorrect.
- The Bias represents the ratio of the number of *Yes* forecasts to the number of *Ye* observations and is a measure of over ar underforecasting.
- The Critical Success Index (CSI; Schaef 1990), also known as the Threat Score, is th proportion of hits that were either forecast observed.
- % Area is the percentage of the total area the forecast domain where convection forecast to occur (*e.g.*, Brown *et al*. 1997).

4. Results

The results presented in this section illustra how verification information can be used to ga understanding of the strengths and weaknesse of the forecasting system, which ultimately ca lead to forecast improvements and corre application of the forecast to a particular weath situation.

4.1 Comparisons between the similar forecasts

Figures 1 and 2 demonstrate the difference between the two stages of the CCFP forecast (i.e., the Final and Preliminary). The Preliminar is issued by one AWC forecaster as a precurso to the Final forecast. The Final is issue through a collaborative process between AW forecasters and airline and other meteorologists The statistical results for the Final an Preliminary forecasts are shown for PODy (Fig 1) and FAR (Fig. 2). Each dot on the lin represents a statistic computed from a 7 - da accumulation of forecast/observation pair covering the period from 1 April – 30 Septembe 2001.

The statistical results shown in Figs. 1 and show little improvement in the quality of the Fina forecast as compared to the Preliminary forecas as a result of the collaboration process. Thi result is very important when users ar evaluating the impact of the "collaboration process on forecast quality. Although th statistics exclude outside benefits on th aviation community, this result can only be obtained by comparing the two forecasts. It also is important to note that the Final forecasts are issued approximately one hour after the Preliminary, so the forecasters have the benefit of additional information.

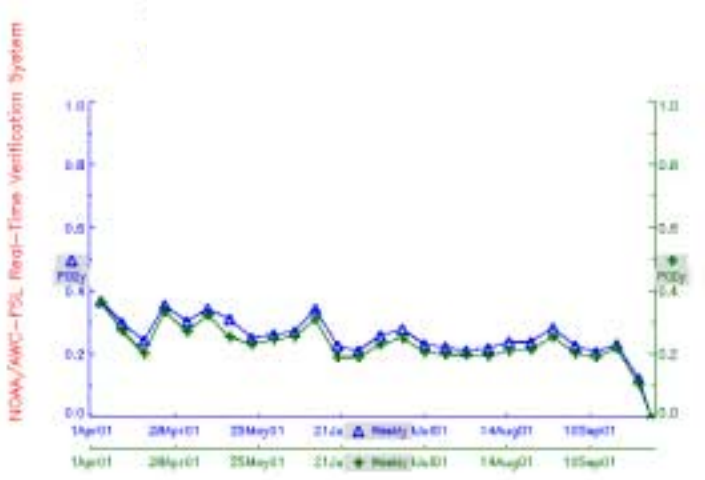


Figure 1. Time series of PODy for Final ('+') and Preliminary ('triangle) CCFP forecasts computed from 1 April – 30 September 2001.

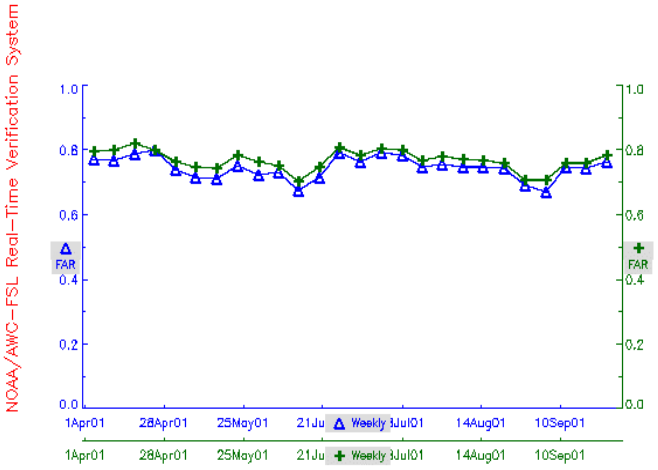


Figure 2. Same as Fig. 1, except for FAR.

Another example of information that can be obtained by comparing forecasts is demonstrated in Figs. 3 (PODy) and 4 (FAR). In this case, the statistics generated from 1 April – 30 September 2000 and 2001 for the 2-h CCFP forecast are compared to provide insight into the seasonal trends associated with the convective forecasts.

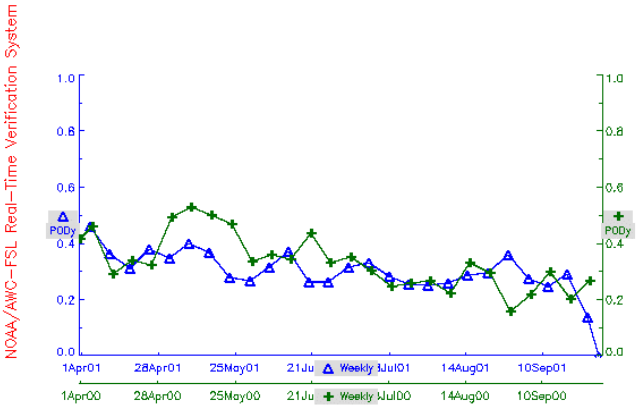


Figure 3. Time series of PODy for the 2-h CCFP forecast from 1 April – 30 September 2000 (;+;) and 2001 ('triangle').

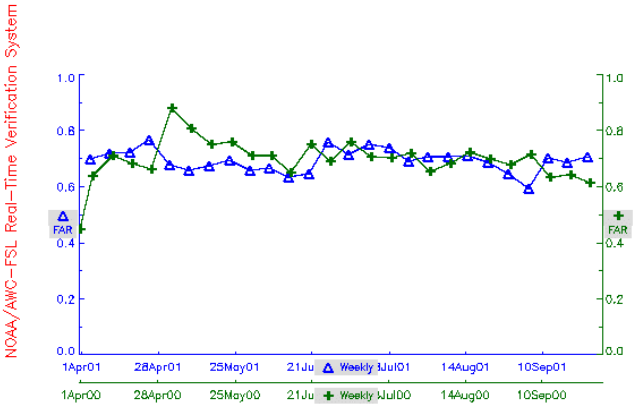


Figure 4. Same as Fig. 3, except for FAR.

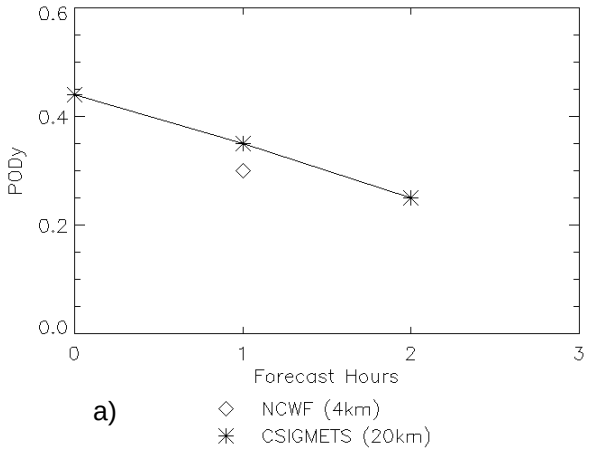
Overall, the quality of the forecasts between the 2000 and 2001 results for the 2-h CCFP are fairly consistent. The trend in PODy values is to reach a maximum early in the convective season (April – June) then decrease to a minimum in September. Excluding a few peaks during May, the FAR remains consistent across the months for both the 2000 and 2001 over the entire convective season. These results suggest that the forecasts are best at capturing convection during the spring and early summer months, possibly due to the more predictable nature of the convection during these periods because of its association with synoptic scale forcing mechanisms.

4.2 Comparisons between different forecasts

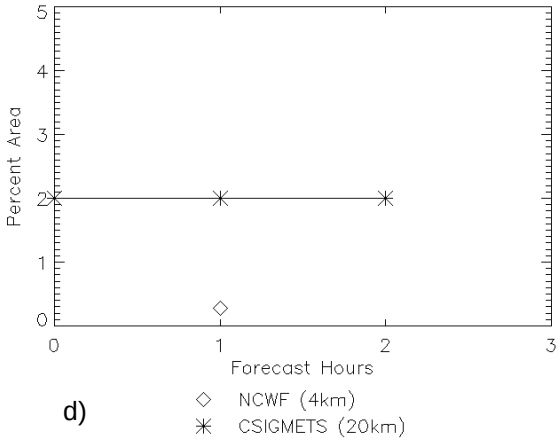
A more difficult comparison than those described in Section 4.1 occurs when forecasts with differing spatial and temporal characteristics are evaluated. Although the comparisons can be difficult, it is necessary, for example, to develop an 'operational' baseline, using the standard operational forecasts, for which new forecast products, such as the NCWF, can be evaluated against. This comparison not only allows calibration of the new forecast, but also provides a mechanism by which the new forecast can considered for operational implementation at the NWS. To illustrate this point, we compare the newly implemented NCWF with the standard operational C-SIGMETs.

Overall statistical results for the NCWF and the C-SIGMETs are summarized in Fig. 5 a-e. The NCWF results are presented for the 4-km grid resolution and the C-SIGMET results are based on the 20-km resolution. The statistics that are associated with each symbol on the plots were generated by accumulating all the counts over all the days and issue times for which data were available from 1 April – 30 September 2001.

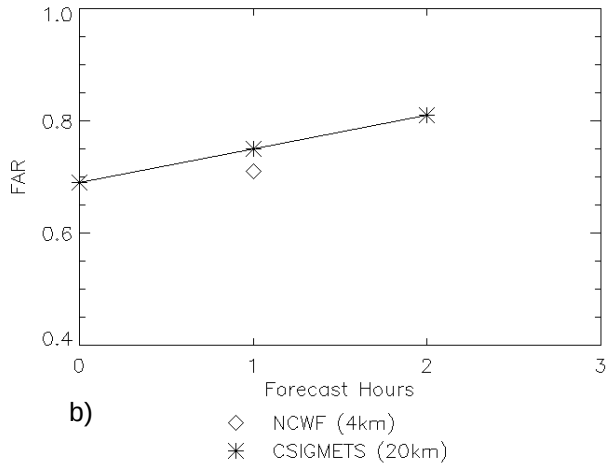
The statistics, shown in Fig. 5 for the NCWF and the C-SIGMETs, are nearly identical with a difference of only 0.02 - 0.03 between the PODy and FAR values for the two types of forecasts. The bias values are nearly equal to 1.0 for both forecasting systems, indicating that the appropriate verification methodology, in space and time, were applied to the forecasts. Large differences in the statistics for the two forecasts were evident in the % Area (Fig. 5d). In particular the % Area value for the NCWF is considerably smaller than the % Area for the



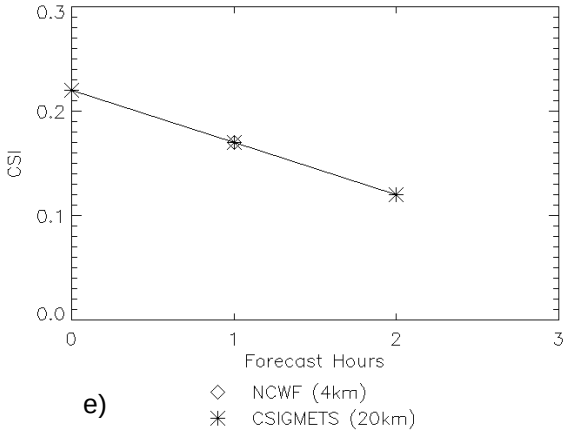
a)



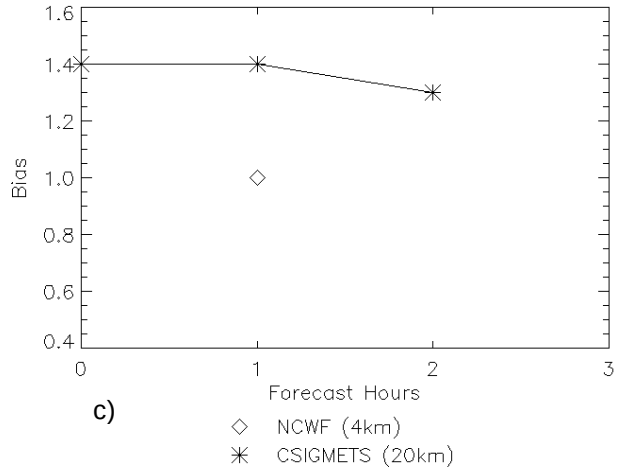
d)



b)



e)



c)

Figure 5 a-e. Verification statistics by forecast lead time for the 1-h NCWF (diamond); and the 0-, 1-, and 2-h C-SIGMETs ("*"; with speed and direction): (a) PODy; (b) FAR; (c) Bias; (d) % Area; and (e) CSI for the period from 1 April – 30 September 2001

C-SIGMETs, at least partially due to the differences in scaling. Finally, the CSI values are also nearly the same for the NCWF and the C-SIGMETs.

Although statistically the results for the NCWF and the C-SIGMETs are similar indicating little difference in performance, fundamentally, the NCWF and the C-SIGMETs are quite different. For instance, the frequency at which the forecasts are issued (5 minutes vs. hourly), the period over which the forecasts are valid (at the end of the period vs. over a valid period), the process by which the forecasts are generated (automated vs. human), and the type of convection captured by the forecasts (NCWF focuses on active convection that is expected to persist while the C-SIGMETs are designed to include developing and moving convection) are very different. Since users struggle to make sense of the similarities in the results, the fundamental differences between the forecasting systems are magnified and users become more aware of the details inherent in the forecasts.

4.3 Comparisons used to understand the purpose of the forecast

Forecasts can also be compared to gain further understanding of the purpose and utility of the forecasts. Figures 6-9 illustrate how this was applied to the C-SIGMETs.

Four different verification approaches (i.e., 0-h forecasts, 1-h without speed and direction, 1-h with speed and direction, and 0-2 h valid over the entire 2-h period) were applied to the C-SIGMETs to determine over what period the forecasts were valid. These results, summarizes using PODy (Fig 6), FAR (Fig. 7), CSI (Fig. 8), and Bias (Fig. 9) are shown as time series plots over the period from 1 April – 30 September 2001. These statistics are based on accumulating all forecast/observation pairs over all issue and times per day.

Overall, the best performance for the C-SIGMETs is gained by evaluating the forecast polygons as 0-h forecasts. For instance, the PODy values (Fig. 6) for the 0-h approach are considerably above the scores that are generated using the other verification

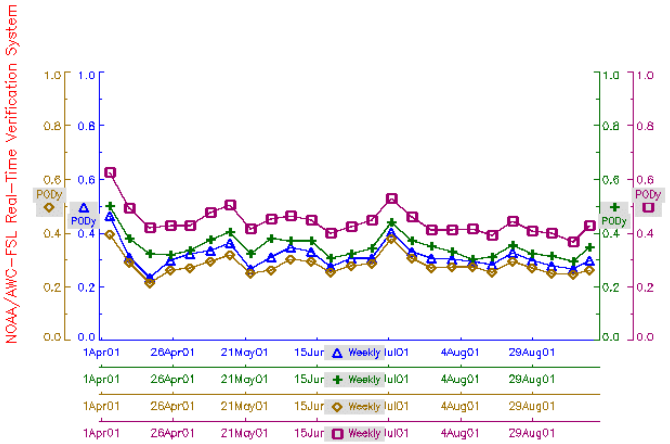


Figure 6. Time series plot of weekly PODy for the C-SIGMET treated as a 1-h forecast with a 20-minute window (triangle), 1-h forecast with speed and direction over a 20-minute window ('+'), 0-2 h forecast (diamond) over a 2-h window, 0-h forecast over a 20-minute window.

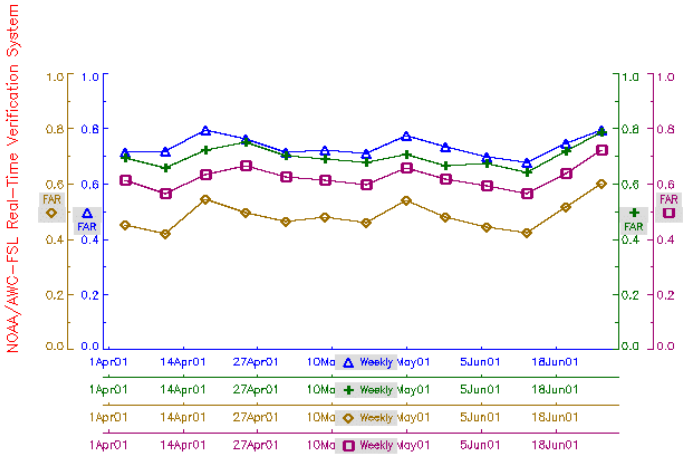


Figure 7. Same as Fig. 6, except for FAR.

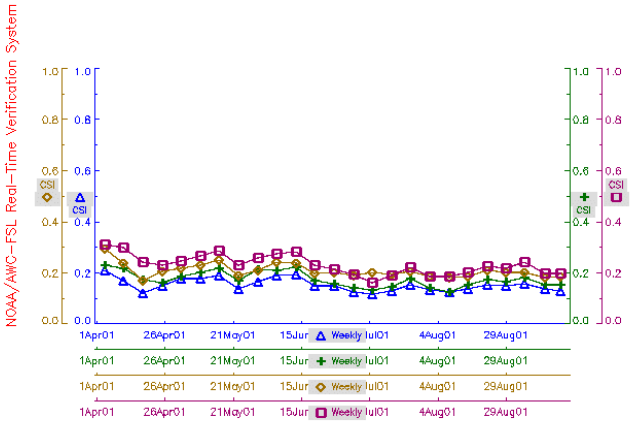


Figure 8. Same as Fig. 6, except for CSI.

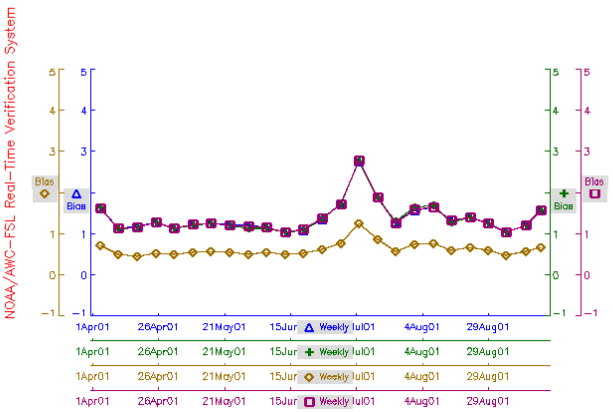


Figure 9. Same as Fig. 6, except for Bias.

approaches. Although the 0-2 h approach has the lowest FAR value (Fig. 7), the scores for the 0-h approach remain lower than the 1-h and 1-h with speed and direction. Large values of CSI (Fig 8) are once again obtained for the 0-h approach. Interestingly, however, all approaches used to verify the C-SIGMETs, with the exception of the 0-2 h approach, slightly overforecast (underforecast) the convective activity as shown by the Bias in Fig. 9.

5. Discussion and conclusions

In this paper, we demonstrated that by comparing the statistical results generated for various convective forecasts, the strengths and weakness of those forecasts became apparent.

When making comparisons between forecasts, it is important to emphasize their differences and develop methods that are appropriate for each forecasting system. For instance, the differences between the NCWF and the C-SIGMETs made it difficult to clearly compare the two forecasts. However, the comparisons showed the quality of the experimental NCWF against the operational standard, which was provided by the C-SIGMETs. This critical information is needed by decision-makers to provide guidance when evaluating whether an experimental forecast should become operationally supported by the NWS.

Forecast comparisons can also be used to gain further understanding of what the forecast is and how it should be used. In this case, the C-SIGMETs were evaluated as if they provided forecasts at 4 different time intervals. The results indicated that the forecast polygons were best at capturing convection at the 0-h time period.

Future work will include continuing evaluations of the NCWF, C-SIGMETs, and CCFP. Users can access results and displays for past and current evaluations at http://www-ad.fsl.noaa.gov/afra/rtvs; link CCFP or convection.

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 express our appreciation to the AWC staff and Cindy Mueller (NCAR) for their assistance with developing the verification methods. We would also like to thank Judy Henderson (FSL), Andy Loughe (FSL/CIRES), and Beth Sigren (FSL/CIRES) for their work on RTVS.

References

Brooks, H.E. and C. A. Doswell III, 1996: A comparisons of measures-oriented and distributions-oriented approaches to forecast verification. *Wea Forecasting.* **11,** 288-302.

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.

Mahoney, J.L., B.G. Brown, C. Mueller, and J.E. Hart, 2000: Convective intercomparison exercise: Baseline statistical results. Preprints, *9th Conference on Aviation, Range, and Aerospace Meteorology*, Orlando, Fl., American Meteorological Society, 403-408.

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.)

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

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