1.6 IMPROVED APPROACHES FOR MEASURING THE QUALITY OF CONVECTIVE WEATHER FORECASTS

Barbara G. Brown[1], Jennifer L. Mahoney[2], Christopher A. Davis[3],
Randy Bullock[3], and Cynthia K. Mueller[3]

1. INTRODUCTION

Verification is a critical component of the development and use of forecasting systems. Ideally, verification should play a role in monitoring the quality of forecasts, provide feedback to developers and forecasters to help improve forecasts, and provide meaningful information to forecast users to apply in their decision-making processes. In addition, as noted by Mahoney et al. (2002) forecast verification can help to identify differences among forecasts. Finally, because forecast quality is intimately related to forecast value, albeit through sometimes complex relationships, verification has an important role to play in assessments of the value of particular types of forecasts (Murphy 1993).

Recently, verification of convective and quantitative precipitation forecasts has received a great deal of attention. For example, a recent (May 2001) World Meteorological Organization, World Weather Research Program workshop focused on the verification of quantitative precipitation forecasts (see http://www.chmi.cz/meteo/ov/wmo/). Unfortunately, traditional approaches for the verification of convective forecasts are inadequate, for a number of reasons. These methods generally rely on overlaying a forecast grid on an observation grid and computing standard statistics based on the 2x2 verification table. A major flaw with this type of verification is that it is insensitive to the size of the forecast error in terms of intensity, areal coverage, location, and timing. For purposes of this paper, we will not directly address an additional important issue, concerning characteristics of the observations used for these types of verification analyses. These observations generally are based on rain gages (which poorly sample the regions of interest) or radar observations (which provide an indirect measure of the phenomena of interest). Finally, spatial and temporal scale issues are of critical importance for precipitation/convective forecasts, and can have especially large impacts on verification results based on this simple approach.

This paper reviews some of the issues mentioned above and some alternative approaches for verification of convective forecasts. Further background is provided in Section 2, with a discussion of specific verification issues in Section 3. Some alternative approaches are described in Section 4, with an example of a simple diagnostic approach presented in Section 5. Future work in this area is considered in Section 6.

2. BACKGROUND

Two general types of convective forecasts are of particular interest here. These could be classified as “area” forecasts and “gridded” forecasts. In general, area forecasts include human-generated “outlook” or warning areas. Gridded forecasts typically are model- or algorithm-generated. Examples of human-generated forecasts are convective outlooks produced by the NWS’s Storm Prediction Center (SPC); the Collaborative Convective Forecast Product (CCFP), which is a 2-to-6-hour forecast of convection expected to impact air traffic, produced by the NWS’s Aviation Weather Center (AWC) in collaboration with airline and other meteorologists (Phaneuf and Nestoros 1999); and convective SIGMETs, also produced by AWC forecasters. Gridded forecasts include national scale forecasts, such as the National Convective Weather Forecast (NCWF), which extrapolates the position of detected convective regions; the output of mesoscale convective forecasting systems [e.g., the National Severe Storms Laboratory's Warning Decision Support System (WDSS)], or the output of numerical weather prediction systems such as the Rapid Update Cycle (RUC).

As noted in the previous section, standard approaches for verification of convective forecasts of these types are based on simple grid overlays. Note that since the observations commonly are gridded, the area-type forecasts generally are also mapped to a grid. From these overlays, counts of forecast/observation (Yes/No) pairs can be computed, to complete the standard 2x2 verification contingency table. The counts in this table can be used to compute a variety of verification measures and skill scores, such as the Probability of Detection (POD), False Alarm Ratio (FAR), Critical Success Index (CSI), and Bias (e.g., Doswell et al. 1990; Stanski et al. 1989; Wilks 1995).

Some of the issues that arise with use of this approach are (a) it is non-diagnostic; (b) it is insensitive to the size of location and timing errors; (c) it is highly sensitive to scaling of observations and forecasts; and (d) the statistics computed have

[1] Corresponding author address: NCAR, P.O. Box 3000, Boulder, CO 80307; e-mail bgb@ucar.edu
[2] Forecast Systems Laboratory, Boulder, CO
[3] National Center for Atmospheric Research, Boulder, CO

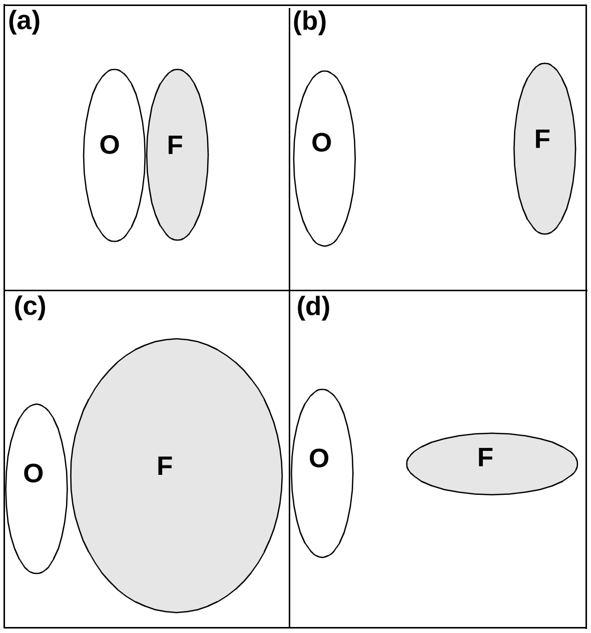

FIGURE 1. Simple example of problems with standard "object-oriented" verification approaches applied to convective forecasts. For each example, the "O" shape represents an observed region; each "F" shape represents a forecast region. In all cases POD=0, FAR=1, CSI=0.

some inherent limitations which cloud their interpretation. In addition, characteristics of the observations are of critical importance.

3. SOME VERIFICATION ISSUES FOR CONVECTIVE FORECASTS

3.1 *Diagnostic verification*

The value of utilizing diagnostic verification approaches has been demonstrated in numerous recent studies (e.g., Brooks and Doswell 1996; Murphy et al. 1989). Diagnostic approaches are based on the principle that verification should help lead to improvements in forecasts and should provide information that is useful to decision makers. For example, verification results should be able to help forecasters identify systematic errors or other factors leading to poor forecasts, or they should lead forecast developers to identify particular problems with a forecasting system.

One important aspect of diagnostic verification is the desirability of examining a variety of measures, to evaluate several attributes of forecast performance. This aspect clearly is achievable with statistics based on the 2x2 table. However, many of the measures associated with the 2x2 table (e.g., POD, FAR) are difficult for forecasters and developers to translate into information regarding needed improvements. Moreover, single measures (e.g., CSI) frequently are relied upon in practice, both for simplicity and because these measures have historical precedence.

3.2 *Sensitivity to timing and location errors*

Another unfortunate characteristic of the standard verification approach for convective forecasts is illustrated in Figure 1. This figure shows four examples of forecast/observation pairs, with the forecasts and observations represented as areas. For a forecast user, the four cases clearly demonstrate four different types or levels of goodness: (a) appears to be a fairly good forecast, just offset somewhat to the right; (b) is a poorer forecast since the location error is much larger than for (a); (c) is a case where the forecast area is much too large and is offset to the right; (d) shows a situation where the forecast is both offset and has the wrong shape. Of the four examples, it appears that case (a) is the "best". Given the perceived differences in performance, it is somewhat dismaying to note that all four examples have identical basic verification statistics: POD=0, FAR=1, CSI=0. Thus, the approach is insensitive to differences in location and shape errors. Similar insensitivity could be shown to be associated with timing errors.

3.3 *Scaling sensitivity*

Scaling issues arise in determining how to match forecasts to observations, as was recently discussed by Tustison et al. (2001). In addition, scaling issues are related to the way forecasts and observations are defined and depicted. Thus, they are related to both the forecasts and the observations, as well as their combination. When verification statistics are based on the 2x2 verification table, the scaling choices can have huge impacts on the values computed.

Figure 2 and Table 1 illustrate some of these impacts. The example shown in Figure 2 is for verification of 1-h CCFP forecasts issued at 1500 UTC on 18 July 1999. The observations used to verify the forecasts are based on the National Convective Weather Detection (NCWD), which combines radar and lightning observations on a 4-km scale (Mueller et al. 1999). In particular, a threshold of vertically integrated liquid of 3.5 kg m^{-2} and/or 3-5 lightning strikes within 10 minutes are used to define a 4-km Yes observation. Three different scales of observations are shown in the figure: in (a), the observations are presented on their "native" 4-km grid; in (b) they are mapped to a 20-km grid (i.e., assigning a Yes observation to a 20-km area if at least one of the 4-km areas embedded in it is a Yes); and in (c) the observations are mapped in the same way to a 40-km grid. It is evident that the observations, which are hardly visible in Figure 1a, cover much larger areas when the observation grid scale is increased to 40 km (Figure 2c).

Table 1 shows how these grid differences impact the verification statistics for this case. Not surprisingly, POD, FAR, and Bias decrease, and CSI increases as the grid spacing increases. Overall, the changes in the statistics are quite large, simply due to changes in the grid spacing. In fact, the change in CSI (which is related to the increased relative frequency of Yes observations; see Mason 1989) is larger than might be expected to occur with an actual improvement in the forecasts.

This example raises the question, What is the "correct" grid spacing? Should it be consistent with the granularity of the forecasts? Or the observations? This issue becomes particularly critical when two different types of forecasts are being compared, such as when a new forecasting system that produces automated, gridded forecasts is being compared to the operational standard, which may be a human-generated areal forecast limited to much less granularity (e.g., Mahoney et al. 2002). These issues are much more critical when the verification is based on the 2x2 verification table than when a more diagnostic approach is applied, as is discussed in Sections 4 and 5.

TABLE 1. Verification statistics for CCFP, 1-h forecast valid at 1600 UTC on 18 July 1999, with different grid sizes applied to the observations.

Observation Grid Size (km)	POD	FAR	CSI	Bias
4	0.44	0.96	0.04	10.38
20	0.39	0.84	0.12	2.51
40	0.35	0.71	0.19	0.12

3.4 *Statistical measures*

It is well known that the measures computed from the 2x2 verification table have some undesirable properties, such as a lack of equitability (e.g., Marzban 1998). Moreover, some of the measures are quite dependent on the climatological probability and/or other measures. For example, FAR and CSI are strongly dependent on the sample climatological probability (Brown and Young 2000; Mason 1989). These characteristics may in some cases lead to misinterpretation of results (e.g., when comparing results for cases with different climatological probabilities). While these aspects of the verification statistics are pervasive (i.e., they impact all verification analyses based on the 2x2 table, not just the convective forecast verification problem), they represent another concern, in addition to those just mentioned, that must be kept in mind when verifying convective forecasts using standard approaches.

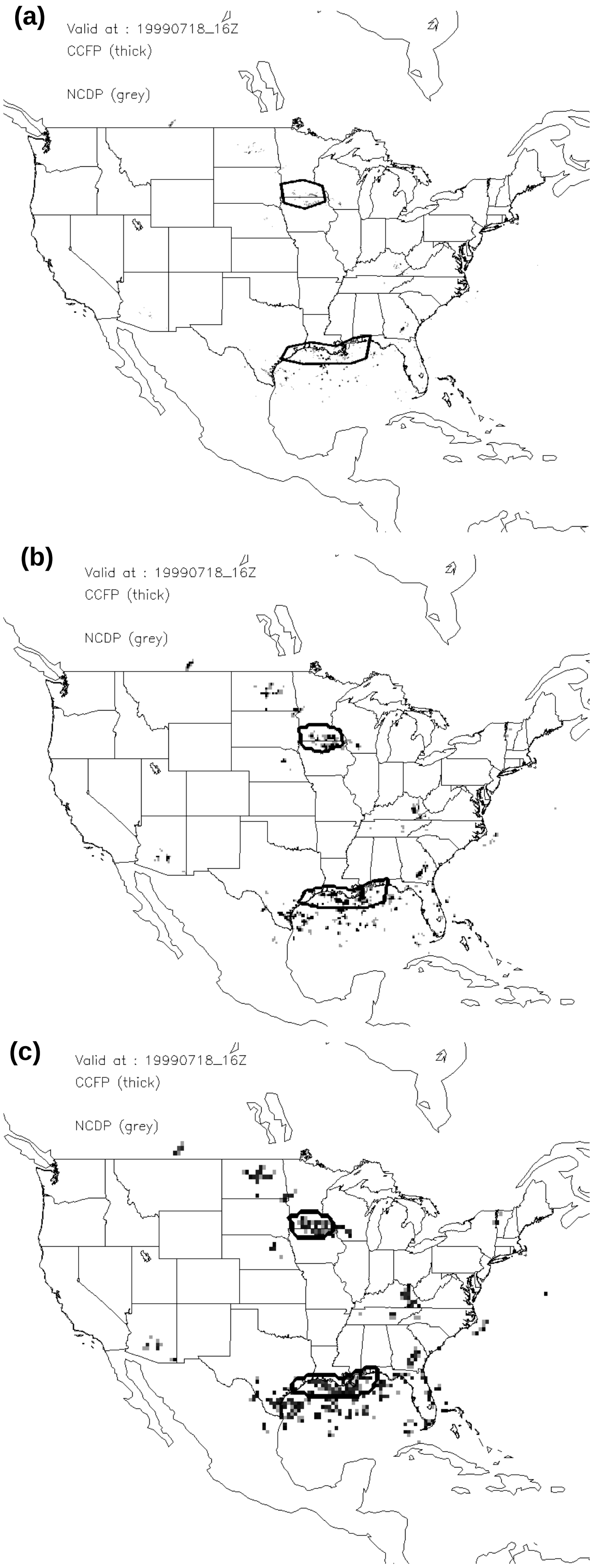

FIGURE 2. Verification maps for 1-h CCFP valid at 1600 UTC on 18 July 1999. Dark shape outlines are the CCFP areas; dark shaded areas are the observations. Three levels of scaling of the observations are shown: (a) 4 km; (b) 20 km; and (c) 40 km.

4. SOME ALTERNATIVE APPROACHES

The previous section has outlined a variety of difficulties associated with standard approaches for verification of convective forecasts. These problems include the non-diagnostic nature of the approach and the difficulty of interpreting some of the measures; the insensitivity of the results to the sizes of the errors; and scaling issues. Fortunately, a number of efforts are already underway toward improving verification approaches for convective and quantitative precipitation forecasts, and for coping with some of the issues mentioned in the previous section. These approaches specifically attempt to evaluate errors in location, intensity, and sometimes the shape of convective or precipitation areas.

Hoffman et al, (1995) developed an approach of interest for application to the output of numerical weather prediction models, with respect to modeled features (e.g., low pressure systems). In this approach, forecast errors are decomposed into errors related to the location, shape, and size of the “objects”.

Ebert and McBride (2000) have extended some of these ideas to verification of precipitation field forecasts. In particular, contiguous rain areas are defined and the root mean squared error (RMSE) of the forecast is decomposed into various sources, including location and intensity errors. In both the Hoffman et al. and Ebert and McBride approaches, the optimal match between forecasts and observations is determined by moving the forecast field around in a systematic manner.

Other approaches associated with the observations also may be beneficial. For example, Briggs and Levine (1997) suggest the use of wavelet transform approaches, which would separate out the less predictable parts of the observation field prior to forecast verification. Another approach involves definition of “practically perfect” forecasts, based on observations, as a standard of comparison for the actual forecasts. Brooks et al. (1998) demonstrate the use of kernel density estimation approaches for this purpose, with application to the SPC’s convective outlooks.

The methods outlined here do not completely solve all of the problems that have been identified. However, they do provide a good initial step in that direction. In addition, they should lead, with further development and testing, to approaches that are more robust and less impacted, for example, by questions of scale and other issues.

5. AN EXAMPLE

Some of the ideas presented in the previous section have recently been applied in the development of a simple diagnostic verification approach for the CCFP. This approach involves systematically moving and rotating the CCFP shapes until an optimal match is achieved with the observations.

Figure 3 shows an example of the application of this approach to a particular case. This example is based on a very simple initial implementation of the approach, and does not incorporate many attributes that we hope to include in the future. In particular, each CCFP shape was moved individually (but objectively) to identify an optimal match with NCWD observations in its local region. The forecast objects were only allowed to translate and to rotate; they were not allowed to change in shape or size.

The original three CCFP shapes for this case are shown in Figure 3a, along with the verifying observations on a 40-km scale. Figure 3b shows the same shapes after translation and rotation. Impacts of the approach are shown in Table 2, which presents the overall verification statistics associated with the two plots (for all three forecast areas together). Table 3 shows the optimal translations and rotations that were applied to the CCFP shapes.

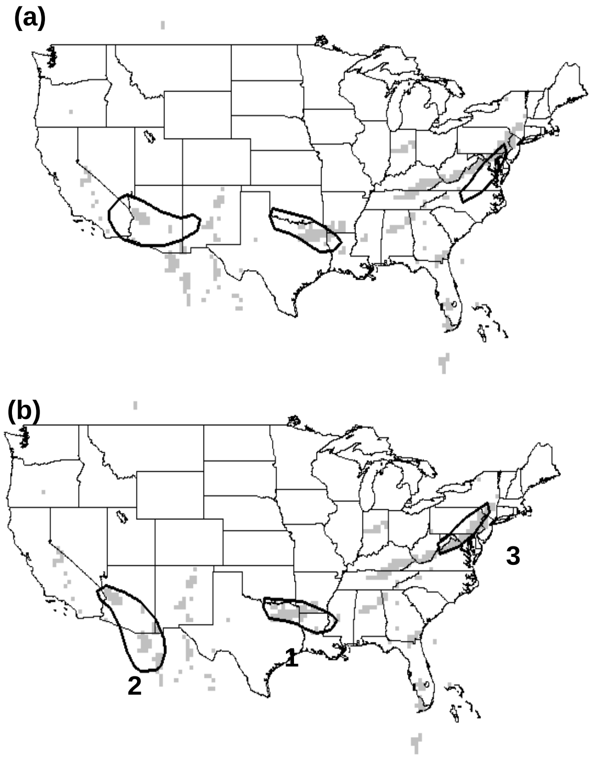

FIGURE 3. Six-hour CCFP forecast, valid at 1900 UTC on 21 June 2000: (a) original forecast areas; (b) optimally translated and rotated shapes. Observations, on a 40-km scale are show in gray.

TABLE 2. Verification statistics for example case shown in Figure 3.

Shapes	POD	FAR	CSI	Bias
Original	0.26	0.86	0.08	1.1
Translated	0.42	0.63	0.24	1.1

TABLE 3. Translation and rotation statistics for example case shown in Figure 3.

Shape	Translation	Rotation
1	218 km	9°
2	279 km	135°
3	249 km	-8°

The statistics in Table 2 indicate that the translated and rotated shapes match the observations much more closely than the original shapes. In particular, POD is nearly doubled, FAR is about two-thirds as large, and the CSI is tripled. The Bias (the ratio of the forecasted area to the observed area) is unchanged, since the sizes of the forecast shapes were not altered. The statistics in Table 3 indicate that the optimal shapes involved a translation of 200 to 280 km. The optimal rotation required to achieve the “best” statistics was quite small for shapes 1 and 3, whereas a large rotation was applied to shape 2. These results provide an indication of the sensitivity of standard verification statistics to relatively small location and orientation errors. In addition, the diagnostic information that is provided clearly delineates the sizes of the true errors in the forecasts.

Examining the shapes in Figure 3b suggests that the forecasts could easily be improved further. For example, forecast shape 3 would provide a better indication of the convection along the frontal region if it extended further along the line to the southwest. Hence, an additional evaluation would consider changes in the shape and/or size of forecast areas that are needed to improve the quality of the forecasts. A hierarchical evaluation approach is envisioned, starting with evaluation of translation and rotation errors, and proceeding to evaluation of size, shape, and intensity errors. Timing errors could be considered in a similar manner. Hence, this simple approach could easily be enhanced and improved a great deal. This proof-of-concept suggests that such enhancements are worth pursuing.

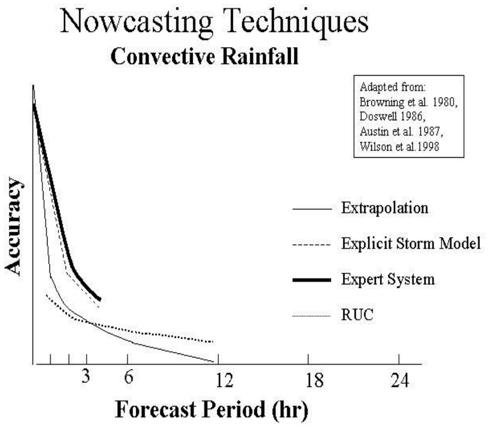

FIGURE 4. Schematic diagram illustrating the idealized relationship between forecast accuracy and temporal scale, based on similar diagrams by Browning et al. (1980), Doswell (1986), Austin et al. (1987), and Wilson et al. (1998)

6. CONCLUSIONS AND FUTURE WORK

This paper has examined some of the problems associated with current methods used to verify convective weather forecasts. In addition, a few approaches for coping with some of these problems have been described.

In the future, we intend to test and extend some of the methods described in Sections 4 and 5, for several different types of convective and quantitative precipitation forecasts. For example, the approach described in Section 5 will be enhanced to allow changes in the sizes of forecast shapes and to limit the rotation of shapes to reasonable values. In addition, more sophisticated optimization and search routines will be implemented to facilitate expanded use of this approach.

Once more appropriate methods have been developed, it will be possible to cope more directly with issues of scale. For example, it will be possible to examine and compare the predictability of convective rainfall as a function of scale, which has commonly been illustrated schematically as in Figure 4. Current standard verification methods limit our ability to quantify diagrams like this, but we anticipate being able to do so with the advent of enhanced verification approaches. Moreover, the verification results associated with these approaches will provide much more meaningful representation and understanding of particular forecast errors.

ACKNOWLEDGMENTS

This research is in response to requirements and funding by the Federal Aviation Administration (FAA) and is also partially funded by the National Science Foundation (NSF) through the U.S. Weather Research Program. NCAR is sponsored by the NSF. The views expressed are those of the authors and do not necessarily represent the official policy or position of the FAA.

REFERENCES

Austin G. L., A. Bellon, P. Dionne and M. Roch, 1987: On the interaction between radar and satellite image nowcasting systems and mesoscale numerical models. *Proceedings, Mesoscale Analysis & Forecasting*, European Space Agency SP-282, Vancouver, Canada, 225-228.

Briggs, W.M., and R.L. Levine, 1997: Wavelets and field forecast verification. *Monthly Weather Review*, **125**, 1329-1341.

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

Brooks, H. E., M. Kay, and J. A. Hart, 1998: Objective limits on forecasting skill of rare events. *Preprints, 19th Conf. On Severe Local Storms*, Minneapolis, MN, Amer. Meteor. Soc., 552-555.

Brown, B.G., and G.S. Young, 2000: Verification of icing and turbulence forecasts: Why some verification statistics can't be computed using PIREPs. *Preprints, 9th Conference on Aviation, Range, and Aerospace Meteorology*, Orlando, FL, 11-15 September, American Meteorological Society (Boston), 393-398.

Browning, K. A., 1980: Local weather forecasting, *Proc. R. Soc. London Ser.*, **A371**, 179-211.

Doswell,C.A., 1986: **Short-range forecasting. Mesoscale Meteorology and Forecasting**, P.Ray, Ed. Amer. Meteor. Soc., Boston, 793 pp.

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

Ebert, E., and J.L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *Journal of Hydrology*, **239**, 179-202.

Hoffman, R.N., Z. Liu, J.-F. Louis, and C. Grassotti, 1995: Distortion representation of forecast errors. *Monthly Weather Review*, **123**, 2758-2770.

Mahoney, J.L., B.G. Brown, J E. Hart, and C. Fischer, 2002: Using verification techniques to evaluate differences among convective forecasts. *Preprints, 16th Conference on Probability and Statistics in the Atmospheric Sciences*, 14-18 January, Orlando, FL, American Meteorological Society (Boston).

Marzban, C, 1998: Scalar measures of performance in rare-event situations. *Weather and Forecasting*, **13**, 753-763.

Mason, I., 1989: Dependence of the critical success index on sample climate and threshold probability. *Australian Meteorological Journal*, **37**, 75-81.

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*, American Meteorological Society (Boston), 230-234.

Murphy, A.H., 1993: What Is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, **8**, 281-293.

Murphy, A.H., B.G. Brown and Y.-S. Chen, 1989: Diagnostic verification of temperature forecasts. *Weather and Forecasting*, **4**, 485-501.

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

Stanski, H., L.J. Wilson, and W.R. Burrows, 1989: Survey Of Common Verification Methods In Meteorology. WMO World Weather Watch Tech. Rep. 8, 114 pp.

Tustison, B., D. Harris, and E. Foufoula-Georgiou, 2001: Scale issues in verification of precipitation forecasts. *Journal of Geophysical Research*, **106**, D, 11775-11784.

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

Wilson, J. W., N. A. Crook, C. K. Mueller, J. Sun and M. Dixon, 1998: Nowcasting Thunderstorms: A Status Report. *Bull. Amer. Meteor. Soc.*, **79**, 2079-2099.