Tornado-Warning Performance in the Past and Future—Another Perspective

—Bob Glahn
NOAA, National Weather Service,
Meteorological Development Laboratory,
Silver Spring, Maryland

Brooks (2004) presents a provocative study of the quality of National Weather Service (NWS) tornado warnings over the past 18 yr through the use of the relative operating characteristic (ROC) from signal detection theory (SDT; Swets 1973; Mason 1980). Primary conclusions, stated as a possible interpretation of plotted results, are that 1) "performance improved from the late 1980s . . . into the 1990s," 2) "the change from the early 1990s to the late 1990s is consistent with a change in the threshold at which decisions are made," and 3) "the primary effect was to improve POD with a small increase in FAR." Brooks sums up his analysis with "Performance in most of the 1990s represents a period where the quality was relatively constant, with only changes in the decision threshold."

How to assess the quality of tornado warnings has enchanted meteorologists at least as far back as the famous Finley (1884)–Gilbert (1884) discussion.[1] In some respects, this is a very simple problem in that all the non–time dependent information is contained in a 2×2 contingency table—the forecasts and events are both dichotomous. Many scores and ways of diagnosing such a contingency table have been developed over the years, the critical success index (CSI; Donaldson et al. 1975), being one that can be calculated from the probability of detection (POD) and false-alarm ratio (FAR).

The contingency table for tornado warnings presents an interesting challenge—it is not complete. In Fig. 1 of Brooks (2004), *d* is not known. That is, how many times was a forecast of no tornado made and no tornado occurred? However this is viewed, it is likely to be quite large compared to the other cell values (*a*, *b*, and *c*) and if considered in a summary measure or score, such as the Heidke Skill Score, would dominate the results. In deriving the CSI, Donaldson et al. (1975) bypassed this problem by ignoring the missing cell; CSI, POD, and FAR do not need the value.

What Brooks (2004) has done is to postulate what this might be by resorting to stratification into easy forecasts and harder ones (the ones for which numbers would be put into the table). This follows, as he states, Murphy's (1995) suggestion of stratification of the full forecast–observed table into strata and looking at each stratum separately. Murphy suggests that instead of formulating the probability of an event directly (e.g., the probability of precipitation occurrence), a forecaster might make a conditional forecast (a forecast conditioned on something such as a weather pattern that "describes completely and unambiguously the meteorological conditions of interest") and then estimate the probability of that condition occurring. The desired probability is, then, the product of the two, and is embodied in the familiar Bayes formula. This is not a new concept in meteorology and was discussed in detail in the 1960s (e.g., Epstein 1962; Olsen 1965).

Murphy (1995) goes on to state that if *comparative* verification is being done, then the condition for stratification should be the same for both (all) sets of forecasts. That is, if one is only looking at the goodness of forecasts when a particular weather or radar pattern occurs, then that pattern should be used for all sets of forecasts for firm conclusions to be drawn. But Murphy only deals with the situation where a full set of data is available, not a partial table.

So to put the tornado warning situation into that stratification framework, one has to define a condition so that the large number of possible "no"

[1] An excellent discussion of dialogue of this era occasioned by the Finley forecasts is given by Murphy (1996).

DOI:10.1175/BAMS-86-8-1135

forecasts is whittled down to a number that seems plausible. Brooks (2004) does this by postulating that warning situations can be divided into difficult and trivially easy ones and that the relative frequency, *f*, of tornadoes in the difficult cases is 0.1. I interpret this to mean that for every "yes" tornado that occurs in the difficult cases, there are nine cases when a tornado does not occur. He does not specify how the easy and difficult cases are distinguished; that is, there are no objective criteria stated—it is left up to the forecaster to decide.

A warning can be issued at any time; a "no warning" is implied when a warning is not made (or is not in effect). But at what frequency is the latter, even in the difficult cases? It has been suggested to me the difficult cases might be those in which a watch is in effect for the same location; however, the same problem of frequency still exists. Schaefer (1990) deals with this by postulating that a forecast of either a tornado or no tornado is made every 10 min when the radar shows echo tops greater that 40 000 ft. Doswell et al. (1990), in their elegant discussion of various scores, have dealt with a grid covering the area and treat each hour separately, resulting in values in all cells.

Given that the conclusions depend on an assumed *f*, which is related to the distinction between difficult and easy forecasts, I am left with the following questions: 1) How sensitive are the results to the value of *f*, and 2) rather than the forecasters' threshold for making a tornado forecast changing over time, might it be possible that *f* changes because some of the difficult situations become trivially easy or vice versa. In other words, there are really two decisions (thresholds) involved: one to differentiate an easy and difficult forecast situation and another to decide whether to make a tornado forecast for the difficult cases. In a sense, Brooks (2004) is comparing forecasts from different years, but has not followed Murphy's (1995) presumption that the same method for stratification is used for all sets (here, years) of forecasts.[2] Since it was not defined, it may be that the former of these thresholds changed over the period of study as well as, or instead of, the latter. It is possible the answers to these questions are embedded in the curves in Brooks' Fig. 5, but they are not apparent to me.

Irrespective of how the forecasters reached their decisions, the improvement in tornado forecasting has been phenomenal over the past 18 yr, the period of record addressed here. This can be deduced from Brooks' Fig. 5, but may be more apparent in Figs. 1–5 presented here. In these graphs, the POD (Fig. 1),[3] FAR (Fig. 2), CSI (Fig. 3), lead time (Fig. 4), and percentage of warnings with lead times greater than zero (Fig. 5) are shown. [A warning with zero lead time occurs when a warning was not issued or was not issued until the tornado was actually reported (Polger et al. 1994). This was a prevalent occurrence before Next Generation Weather Radar (NEXRAD) was implemented, occurring between 60% and 80% of the time.] In these figures, one can consider three time periods—pre-NEXRAD (1986–91), NEXRAD deployment (1992–95), and post-NEXRAD (1996–2003). The deployment of the Weather Surveillance Radar-

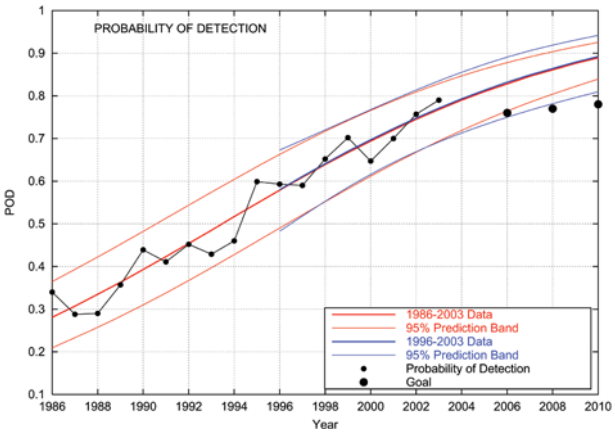

Fig. 1. POD for tornado warnings, 1986–2003. A weighted logit curve fit to the total period and its associated 95% prediction band are shown. Also shown are a weighted logit curve fit to the post-NEXRAD period 1996–2003 and its associated 95% prediction band. The weights used were the number of events, the denominator in POD. NOAA targets for the years 2006, 2008, and 2010 are also shown as heavy dots.

[2] Murphy (1995) states that this method of stratified verification ". . . involves (a) introducing a variable (or covariate) that describes completely and unambiguously the meteorological conditions of interest and (b) conditioning the underlying joint distribution on the basis of the values of this covariate" (p. 1582). On comparative verification, he says "Application of the extended framework in the context of comparative verification appears to be relatively straightforward, since it presumably would involve a common stratification scheme applied to both sets of distributions" (p. 1587).

[3] As footnoted in Brooks (2004), because of the way the NWS computes POD and FAR, two contingency tables are involved, and the CSI is computed from POD and FAR to make the three scores consistent, rather than computing CSI directly from one of the two tables. A slight change on 1 January 2002, (rule 2) in the verification procedure had minimal impact (W. Lerner 2004, personal communication).

1988 Doppler (WSR-88D) [also called NEXRAD (Friday 1994, p. 44)] started in June 1992 and over 85% had been installed by the tornado season in 1996. One might have expected a relatively flat level of performance in the pre-NEXRAD period (1986–91), a gradual, pronounced improvement during the deployment years (1992–95), and again a leveling off in the years 1996–2003; that is not, however, the case. Rather, following 1986–89, there was marked improvement before NEXRAD deployment started in 1992 and, generally, improvement not only throughout deployment but also post-deployment. Since 1986, the average lead time has increased from about 5 min to over 12 min, the POD has increased from 30% to greater than 75%, the number of warnings with greater than 0-min lead time has increased from less than 30% to about 70%, the CSI has increased from about 0.13 to 0.23, and at the same time FAR has decreased, especially in the post-NEXRAD deployment period.

In each of the five figures, an improving trend can be easily discerned (the FAR in Fig. 2 being the most variable and least consistent in that regard), and either a weighted least squares linear regression or a weighted logit is shown and extended to 2010. [Discussions of these methods are given in Montgomery and Peck (1982), p. 99, pp. 239–240; Neter and Wasserman (1974), p. 331; and Glahn (2002).] While one may disagree with the exact weights used in fitting the lines, the results are not much different from unweighted fits. There is considerable variability of yearly scores about the fitted line. Given this variance, prediction lines can be plotted, with reasonable distributional assumptions,

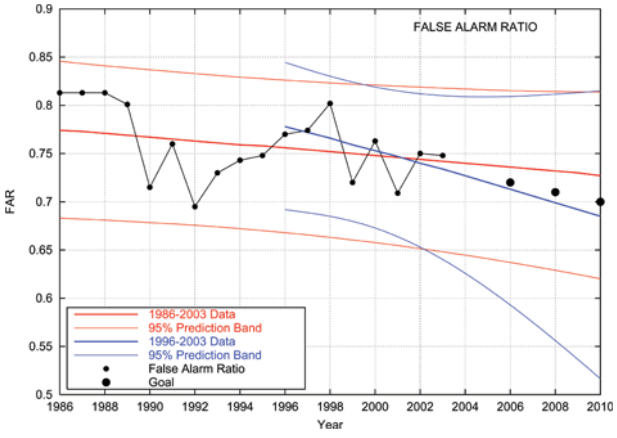

Fig. 2. FAR for tornado warnings, 1986–2003. Weighted logit curves and NOAA targets similar to those in Fig. 1 are shown, except the weights were the number of warnings, the denominator in FAR.

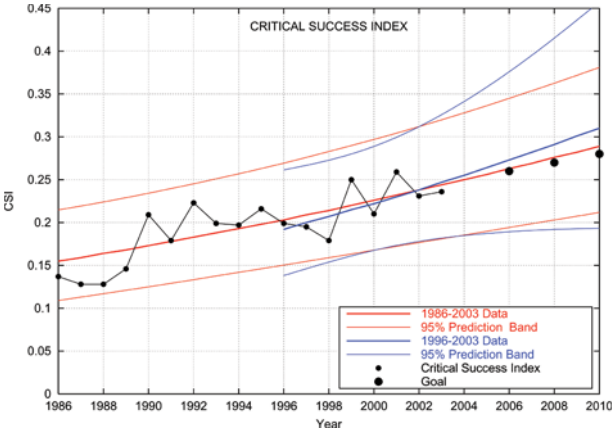

Fig. 3. CSI for tornado warnings, 1986–2003. Weighted logit curves and computed NOAA targets similar to those in Fig. 1 are shown, the weights being the number of warnings.

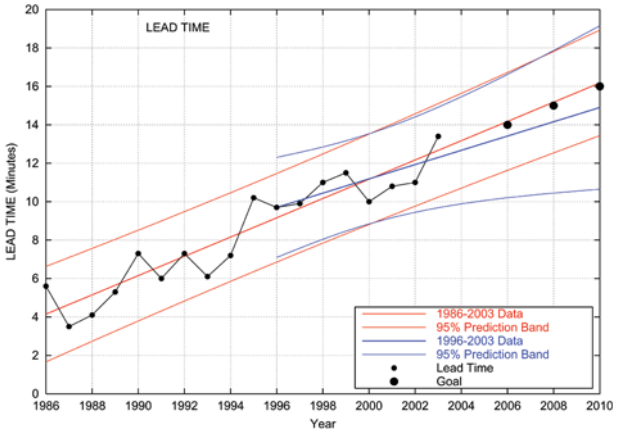

Fig. 4. Average lead times in minutes of tornado warnings, 1986–2003. Weighted linear least squares regression lines and NOAA targets similar to those in Fig. 1 are shown, the weights being the number of events.

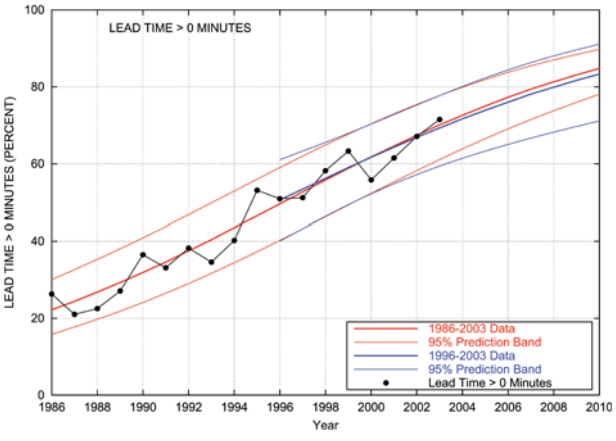

Fig. 5. Percentage of warnings with lead times greater than 0, 1986–2003. Weighted logit curves and prediction bands for the total sample and for the post-NEXRAD period are shown, the weights being number of events.

on either side of the trend line such that if new measurements could be made for individual years, there is an x% chance those measurements would fall inside the band between the lines. The lines in the figures are plotted at the x = 95% level. This is important in extending the trend lines into the future as an aid in establishing expectations.

Prediction bands relate to the prediction of a new datum, and are in distinction to confidence bands, which relate to the estimate of the mean response given by the regression line. A particularly good discussion is given by Neter and Wasserman (1974). The t test they propose (p. 72) was used to establish the prediction bands. The mean square error was about the same whether calculated on weighted data or unweighted data; the latter was used. For the logit, the bands were established in the transformed space, then transformed back to the original space (see Glahn 2002).

In addition to fitting a trend to the total 1986–2003 sample, another trend line was fitted to each set of post-NEXRAD data; these lines together with their 95% prediction bands are also plotted in Figs. 1–5. As can be seen, there is not a great deal of difference between the two sets of lines, lead time and FAR being the most different. The prediction bands are wider in the out years (beyond 2003) for the post-NEXRAD fit, mostly because of the decrease in sample size (from 18 to 8) and the resulting decrease in degrees of freedom.

It seems clear that while the WSR-88Ds were a key component of this phenomenal improvement in NWS tornado warning capability (Polger et al. 1994; Bieringer and Ray 1996), other factors were, and are, playing a very significant role. Even before NEXRAD, the NWS was doing a lot to improve its service. Spotter networks and partnerships with amateur radio operators, emergency managers, and TV stations were being built, and better use was being made of verification data. Throughout this 18-yr period, much was progressively learned about the structure of storms and their involvement with the large scale environment. The term mesoscale convective complex (MCC; Maddox 1980) came into being about 1980 and the MCC relationship to mesoscale convective systems in general was being studied extensively (e.g., Corfidi et al. 1996). Also, the large-scale environment was being predicted much more accurately by the operational models being run by the National Centers for Environmental Prediction (NCEP). Specialized applied research (e.g., Corfidi 1998) has improved interpretation of data and guidance products available from the Storm Prediction Center. Training was a very important element of the NEXRAD program and undoubtedly played a major role in the improvement; training on new science constantly being learned is continuing today. For instance, the Weather Event Simulator (Magsig and Page 2002; Magsig et al. 2005) has played, and is playing, a very important role. Waldstreicher (2005) provides a more complete discussion of the factors affecting tornado-warning performance and the role that collaborative research has played.

In addition to NEXRAD, other legs of the NWS modernization were put into place during this period (Friday 1994). The NWS office restructuring and staffing modifications were completed. The Automated Surface Observing System (ASOS) was implemented, thereby saving labor that was then used more directly in the forecast process. In some tornado-prone areas, a wind profiler demonstration network was put into place (van de Kamp 1993), and Beckman (1993) concludes, ". . . profiler winds have become an integral part of NSSFC [National Severe Storms Forecast Center] operation." Satellite data and its delivery and interpretation improved (e.g., see Menzel and Purdom 1994; Weaver and Purdom 1995). The Advanced Weather Interactive Processing System (AWIPS) was deployed and provides the forecaster a much better means of viewing the various forms of data together so that a better four-dimensional picture of the atmosphere can be visualized (Seguin 2002). AWIPS also, with its sophisticated and constantly evolving software (e.g., see Jones et al. 2004), provides more interactive diagnostic capability (e.g., SCAN; Smith et al. 1999) and a more responsive medium for disseminating warnings once the forecaster decides to issue a warning (U.S. Department of Commerce 1999).[4] In addition, lightning-detection networks started in the late 1970s (Orville et al. 1983; Maier et al. 1984) had by 1998 grown to cover Canada and the contiguous United States (Orville et al. 2002), which may play a role in situational awareness. The decision process has been greatly aided by this technology (e.g., see Andra et al. 2002).

All of these factors and others have undoubtedly contributed to improvement of tornado warning service and reinforces the "can do" attitude prevalent among NWS forecasters, thereby promoting even

[4] Page 18 of the Service Assessment of the Oklahoma/Southern Kansas tornado outbreak of 3 May 1999 states, "AWIPS was critical to the success of this event. It would have been impossible to duplicate the number of successful warnings and lead times and to keep track of the large number of severe storms with a mixture of [old systems]."

greater success. Even the improved working conditions in the modernized Weather Forecast Office (WFO) facilities may have been a factor. For instance, Smith et al. (2003) have presented findings relating social and cultural factors and tornado-warning performance in the NWS. This is undoubtedly a symbiotic relationship—better forecaster morale, better service. The modernization as envisioned by such visionaries as Dick Hallgren, Doug Sargeant, and Joe Friday has paid off.

The effect on these statistics of more intense data gathering (verification) that was evolving during this period because of increased emphasis being placed on saving lives cannot be ignored. Such scores as POD and FAR are only as good as the data that went into them, and the staffs at the Weather Forecast Offices are persistent in trying to determine whether or not a tornado warning is verified; this can be problematic, especially in sparsely populated areas. If a tornado was expected in an area (i.e., the conditions were ripe), the search for the occurrence of one is logically more paramount than for much larger areas where one was not expected. Not observing a tornado when a warning was not issued improves POD and CSI. However, these more persistent verification efforts, while they likely played some role, were not the primary reason for this phenomenal improvement by NWS forecasters. Figure 6 shows the increase and trend lines of the number of warnings and events in the 1986–2003 period. The number of warnings issued notably increased from 1995 onward, while the slow and continued increase in reported events is likely a function of population growth and other observational enhancements.

What do the extensions of the trend lines and the prediction bands for years beyond 2003 mean? Improvements in the past have been due to a number of factors. If technology, data sources, training, and scientific understanding can continue at the same rate as over the past 18 years or 8 years, then error bands on the trend lines indicate that scores for these out years will likely (with a 95% probability) fall within those lines. The NWS goals for 2006, 2008, and 2010 have been plotted as asterisks on the POD, FAR, and lead time charts (the ones for which goals have been defined). By using the POD and FAR goals, CSIs can

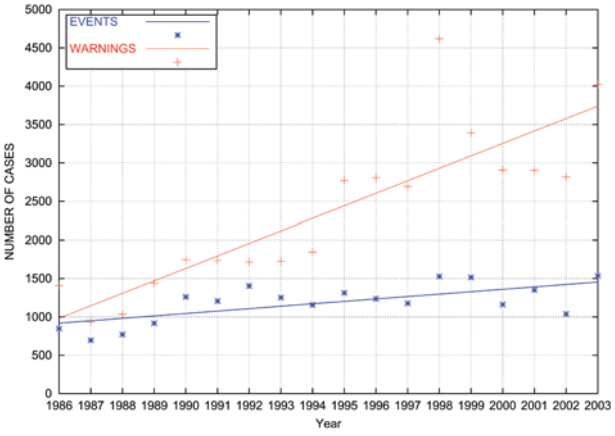

Fig. 6. Number of warnings issued and number of events, 1986–2003. Linear least squares regressions are shown.

be calculated, and these are also plotted. The goals are in general in agreement with the trend lines, the only one seemingly conservative is POD.[5] One should be cautious when setting goals dependent on trends over the last few years. These past years were years of tremendous scientific and technological advancement for the NWS. Improvements are still being made to systems already in place (e.g., NEXRAD, ASOS, AWIPS, satellites). However, one can agree with Brooks (2004) who states, "It is not clear what mix of changes in science, technology, training, and guidance would be necessary to lead to future major improvements in the quality of warnings . . . . It seems likely that continued significant, or even enhanced, investments in all of the areas will be necessary. Large improvements in quality can occur, but they are unlikely to come for free." While phased array radar, better satellite data, and better in situ observing systems are being planned, implementation would be at the latter end of the 2006–10 period, at best. Much is still being learned about tornado formation and forecasting as Johns and Doswell (1992) clearly show, and much still needs to be learned (Hilgendorf and Johnson 1998). However, computer processing power is still increasing at a rapid pace and this will allow full implementation and support of very high resolution numerical weather prediction models that explicitly resolve convection rather than parameterize it and may actually and reliably produce realistic results in the 0–3-h time frame and not just results that look realistic. For better tornado-warning service to occur, substantial investments must be made in science and technology, as they have in the past. NWS forecasters will rise, as always, to the challenge to take advantage of the improved science, technology, and knowledge as they evolve.

[5] It may seem a little inconsistent that the goals for FAR and calculated CSI agree with the trend lines, but POD seems low. As Polger (1994, p. 204, Fig. 2) shows, CSI is quite insensitive to small changes in POD when POD is in the 0.75–0.85 range and FAR is approximately 0.7.

ACKNOWLEDGMENTS. I appreciate the helpful comments of Stephan Smith and Dennis McCarthy. I also thank Scott Scallion for preparing the figures. The views expressed here are those of the author and do not necessarily reflect those of the National Weather Service.

REFERENCES

Andra, D. L., Jr., E. M. Quoetone, and W. F. Bunting, 2002: Warning decision making: The relative roles of conceptual models, technology, strategy, and forecaster expertise on 3 May 1999. *Wea. Forecasting,* **17,** 559–566.

Beckman, S. K., 1993: Operational application of 404 Mhz wind profilers at NSSFC. Preprints, *26th Int. Conf. on Radar Meteorology,* Norman, OK, Amer. Meteor. Soc., 555–557.

Bieringer, P., and P. S. Ray, 1996: A comparison of tornado warning lead times with and without NEXRAD Doppler radar. *Wea. Forecasting,* **11,** 47–52.

Brooks, H. E., 2004: Tornado-warning performance in the past and future. *Bull. Amer. Meteor. Soc.,* **85,** 837–843.

Corfidi, S. F., 1998: Forecasting MCC mode and motion. Preprints, *19th Conf. Severe Local Storms,* Minneapolis, MN, Amer. Meteor. Soc., 626–629.

——, J. H. Merritt, and J. M. Fritsch, 1996: Predicting the movement of mesoscale convective complexes. *Wea. Forecasting,* **11,** 41–46.

Donaldson, R. J., Jr., M. Dyer, and M. J. Kraus, 1975: An objective evaluator of techniques for predicting severe weather events. Preprints, *Ninth Conf. Severe Local Storms,* Norman, OK, Amer. Meteor. Soc., 321–326.

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

Epstein, E. S., 1962: A Bayesian approach to decision making in applied meteorology. *J. Appl. Meteor.,* **1,** 169–177.

Finley, J. P., 1884: Tornado predictions. *Amer. Meteor. J.,* **1,** 85–88.

Friday, E. W., Jr., 1994: The modernization and associated restructuring of the National Weather Service: An overview. *Bull. Amer. Meteor. Soc.,* **75,** 43–52.

Gilbert, G. F., 1884: Finley's tornado predictions. *Amer. Meteor. J.,* **1,** 166–172.

Glahn, H. R., 2002: A methodology for evaluating and estimating performance metrics. MDL Office Note 02-1, National Weather Service, NOAA, U.S. Department of Commerce, 18 pp. (Available from the Meteorological Development Laboratory, 1325 East-West Highway, Silver Spring, MD 20910.)

Hilgendorf, E. R., and R. H. Johnson, 1998: A study of the evolution of mesoscale convective systems using WSR-88D data. *Wea. Forecasting,* **13,** 437–452.

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

Jones, D. R., and Coauthors, 2004: AWIPS Build 5 in review. Preprints, *20th Int. Conf. Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology,* Seattle, WA, Amer. Meteor. Soc., CD-ROM, 4.1.

Maddox, R. A., 1980: Mesoscale convective complexes. *Bull. Amer. Meteor. Soc.,* **61,** 1374–1387.

Magsig, M. A., and E. M. Page, 2002: Development and implementation of the NWS warning event simulator version 1.0. Preprints, *Interactive Symp. on AWIPS,* Orlando, FL, Amer. Meteor. Soc., CD-ROM J236–J238.

——, N. M. Said, N. Levit, and X. Yu, 2005: Build four of NOAA's NWS weather event simulator. Preprints, *21st Int. Conf. Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology,* San Diego, CA, Amer. Meteor. Soc., CD-ROM, P2.38.

Maier, L. M., P. Krider, and M. W. Maier, 1984: Average diurnal variation of summer lightning over the Florida peninsula. *Mon. Wea. Rev.,* **112,** 1134–1140.

Mason, I. B., 1980: Decision-theoretic evaluation of probabilistic forecasts using the relative operating characteristic. *Proc. WMO Symp. on Probabilistic and Statistical Methods in Weather Forecasting,* Nice, France, World Meteorological Organization, 219–227.

Menzel, W. P., and J. F. W. Purdom, 1994: Introducing GOES-I: The first of a new generation of geostationary operational environmental satellites. *Bull. Amer. Meteor. Soc.,* **75,** 757–781.

Montgomery, D. C., and E. A. Peck, 1982: *Introduction to Linear Regression Analysis.* John Wiley and Sons, 504 pp.

Murphy, A. H, 1995: A coherent method of stratification within a general framework for forecast verification. *Mon. Wea. Rev.,* **123,** 1582–1588.

——, 1996: The Finley affair: A signal event in the history of forecast verification. *Wea. Forecasting,* **11,** 3–20.

Neter, J., and W. Wasserman, 1974: *Applied Linear Statistical Models.* Richard D. Irwin, Inc., 872 pp.

Olsen, R. H., 1965: On the use of Bayes' theorem in estimating false alarm rates. *Mon. Wea. Rev.,* **93,** 557–558.

Orville, R. E., R. W. Henderson, and L. F. Bosart, 1983: An East Coast lightning detection network. *Bull. Amer. Meteor. Soc.,* **64,** 1029–1037.

——, G. R. Huffines, W. R. Burrows, R. L. Holle, and K. L. Cummins, 2002: The North American lightning detection network (NALDN)—First results 1998–2000. *Mon. Wea. Rev.,* **130,** 2098–2109.

Polger, P. D., B. S. Goldsmith, R. C. Przywarty, and J. R. Bocchieri, 1994: National Weather Service warning performance based on the WSR-88D. *Bull. Amer. Meteor. Soc.,* **75,** 203–214.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting,* **5,** 570–575.

Seguin, W. R., 2002: AWIPS—An end to end look. Preprints, *Interactive Symp. on Advanced Weather Interactive Processing System (AWIPS),* Orlando, FL, Amer. Meteor. Soc., CD-ROM, J47–J51.

Smith, S. B., G. K. Goel, M. T. Fillaggi, M. E. Churma, and L. Xin, 1999: Overview and status of the AWIPS System for Convection Analysis and Nowcasting (SCAN). Preprints, *15th Int. Conf. Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology,* Dallas, TX, Amer. Meteor. Soc., 326–329.

——, L. Mischkind, and S. D. Duco, 2003: The Impact of Social/Cultural Factors on Tornado Warning Performance. [Available online at www.nws.noaa.gov/mdl/pubs/impactculturalfactors.pdf.]

Swets, J. A., 1973: The relative operating characteristic in psychology. *Science,* **182,** 990–1000.

U.S. Department of Commerce, 1999: Service assessment Oklahoma/southern Kansas tornado outbreak of May 3, 1999. National Oceanic and Atmospheric Administration, National Weather Service, 33 pp.

van de Kamp, D. W., 1993: Current status and recent improvements to the wind profiler demonstration network. Preprints, *26th Int. Conf. Radar Meteorology,* Norman, OK, Amer. Meteor. Soc., 552–557.

Waldstreicher, J. S., 2005: Assessing the impact of collaborative research projects on NWS warning performance. *Bull. Amer. Meteor. Soc.,* **86,** in press.

Weaver, J. F., and J. F. W. Purdom, 1995: An interesting mesoscale storm-environment interaction observed just prior to changes in severe storm behavior. *Wea. Forecasting,* **10,** 449–453.