# Evaluating the Performance of WSR-88D Severe Storm Detection Algorithms

ARTHUR WITT, MICHAEL D. EILTS, GREGORY J. STUMPF,* E. DEWAYNE MITCHELL,*
J. T. JOHNSON, AND KEVIN W. THOMAS*

*NOAA/ERL/National Severe Storms Laboratory, Norman, Oklahoma*

3 March 1997 and 15 January 1998

ABSTRACT

This paper discusses some important issues and problems associated with evaluating the performance of radar-based severe storm detection algorithms. The deficiencies of using *Storm Data* as a source of verification are examined. Options for equalizing the time- and space scales of the algorithm predictions and the corresponding verification data are presented. Finally, recommendations are given concerning the different evaluation procedures that are available.

## 1. Introduction

The National Severe Storms Laboratory (NSSL) and the WSR-88D Operational Support Facility (OSF) have developed a computer software package, called the WSR-88D Algorithm Testing and Display System (WATADS; NSSL 1997), for generating and displaying output from radar-based severe storm detection algorithms using WSR-88D level II data (Crum et al. 1993). One of the primary reasons for developing WATADS was to give a wide range of users the opportunity to conduct studies on the performance of the current set of operational WSR-88D severe storm detection algorithms, as well as enhanced algorithms soon to be added to the WSR-88D system. In order to facilitate comparison of the results from these studies, it is desirable for the studies to use a common methodology for evaluating performance.

Due to its wide geographic coverage, *Storm Data* is often used to verify predictions for severe weather events.[1] However, certain characteristics inherent in the observations contained in *Storm Data* make its use for verifying the performance of radar-based severe storm detection algorithms problematic. The nature of these problems is discussed in section 2. Section 3 presents different options for evaluating algorithm performance, while section 4 gives some recommendations. The primary purpose of this paper is to describe some of our ideas and experiences in attempting to evaluate radar-based severe storm detection algorithms and the methods we believe are best to use. A thorough discussion of all issues and possible solutions to the problems presented here is beyond the scope of this paper.

## 2. Verification issues

The development of an appropriate (i.e., scientifically and statistically valid) performance-evaluation methodology for severe storm detection algorithms depends on the character of the verification database, as well as the time and space scales of the algorithm predictions and how well those scales correlate to the verification database. The verification databases range from special field project datasets, such as the Verification of the Origins of Rotation in Tornadoes Experiment (VORTEX) (Rasmussen et al. 1994) or the National Center for Atmospheric Research (NCAR) Hail Project (Kessinger and Brandes 1995), to *Storm Data*. The field project datasets are limited in scope, but generally of high quality. On the other hand, *Storm Data* provides severe weather information for the entire United States, but the information is less detailed, and often less accurate.

The reports that are listed in *Storm Data* are generated primarily through 1) probing calls made by National Weather Service (NWS) offices while deciding whether or not to issue a severe thunderstorm or tornado warning, or calls made in an attempt to verify any warnings that have been issued; 2) reports volunteered to the NWS from storm spotters, emergency management groups, law enforcement agencies, and the general public; and

* Additional affiliation: Cooperative Institute for Mesoscale Meteorological Studies, Norman, Oklahoma.

*Corresponding author address:* Arthur Witt, National Severe Storms Laboratory, 1313 Halley Circle, Norman, OK 73069.
E-mail: witt@nssl.noaa.gov

[1] *Storm Data* is a list of severe weather observations across the United States. It is published on a monthly basis by the National Climatic Data Center (Available from NCDC, Federal Building, Asheville, NC 28801-2696).

3) information provided to NWS offices by newspaper clipping services. Since many of the reports in *Storm Data* come from the NWS verification process (Hales and Kelly 1985), the reports will often be on the same time and space scales as the warnings, which are typically issued for one or more counties for up to 60 min in length. Because a warning is actually a short-term forecast for a specific area and time period, a single observation of severe weather is sufficient to verify the forecast (warning). More detailed observations are not necessary for this purpose. Due to limited personnel resources available for verification efforts, once a report is received to verify a warning at an NWS office, usually no further effort (via probing calls) is taken to further quantify the extent of the severe event (Amburn and Wolf 1997). In fact, multiple nontornadic reports occurring within 16.1 km (10 mi) and 15 min of each other and in the same county are considered to be the same severe event for warning verification purposes (Crowther and Halmstad 1994; Polger et al. 1994). Thus, even in areas of the United States where NWS offices devote a large effort toward warning verification, one cannot necessarily expect *Storm Data* to contain severe reports at a much higher frequency or density than is needed to verify warnings.

An example that illustrates the temporal resolution problems that can occur when using *Storm Data* for WSR-88D algorithm verification is this: suppose a long-lived, nontornadic supercell affects six contiguous counties over a 6-h time period, with the storm taking 1 h to traverse each county. If one warning is issued for each county for 60 min, then only six severe reports are needed to completely verify this event (for NWS warning verification purposes). If this event is sampled by a WSR-88D operating in Volume Coverage Pattern 11 (5-min volume scan time), there will be 78 algorithm predictions for this storm. If the severe reports are like most in *Storm Data*, they will have only a single time listed (except for tornado reports, which often have both a beginning and ending time), which leads to 72 volume scans of WSR-88D algorithm output with no direct verification data. The problem then becomes one of reconciling these different timescales. One could take the extreme view that since no severe weather was listed at the times of these 72 volume scans, none actually occurred. However, a more reasonable approach would be to try to equalize the timescales, either by 1) reducing the number of algorithm predictions, or 2) extending the time period over which to apply the severe reports to a number of different algorithm predictions. The best approach to use is debatable, although the latter method allows for a more detailed analysis of the algorithm predictions.

Another example illustrates the spatial resolution problems that can occur when using *Storm Data* for WSR-88D algorithm verification: if two or more individual storms move through a fairly large county at the same time, a single warning could be issued to cover all of these storms. For the NWS to verify this warning, they would only need a single report of severe weather from any one of the storms, since the warning was for a specific area and not for a specific storm. However, because algorithm predictions are produced for each individual storm, it is important, for algorithm evaluation purposes, to know which individual storm cells produced severe weather and which did not, and that is difficult to determine without numerous surface observations.

Another limitation associated with algorithm evaluation using *Storm Data* involves the relative importance of reporting one type of severe event versus another for a given storm. For example, hail that accompanies a tornadic storm is often not reported (Morgan and Summers 1982), due to the tendency for observers to focus on the more spectacular, or urgent, severe event and to ignore lesser ones. Similarly, large hail, and damaging winds are rarely reported from the same storm (Kelly et al. 1985). However, considerable evidence from research studies of supercell storms suggests that these events often occur together. Since algorithms exist for the detection and/or prediction of each type of severe weather phenomenon (damaging winds, hail, and tornadoes), it is necessary to obtain the best possible verification information about each specific type of severe weather.

Another important issue for algorithm evaluation is the general lack of null observations (i.e., observations that severe weather was not occurring). This is especially true for *Storm Data*, since it rarely contains nonsevere reports; but even field project observations can be deficient in this area, since it is very difficult to observe the full extent of a given storm. Without null observations, one usually assumes that if no severe weather was reported with a storm, none actually occurred. Although this assumption may be reasonable for storms located over high-population areas, it may lead to errors in the evaluation (too many algorithm predictions counted as false alarms) for storms located over rural areas.

An even more difficult set of evaluation issues arise when one desires to move beyond the question of whether or not a severe event has occurred and instead attempt to determine the actual magnitude of specific severe weather phenomena (e.g., maximum hail size and hail-swath dimensions, maximum surface wind speeds and area affected, tornado intensity). Many problems exist with assessing the intensity of tornadoes (Doswell and Burgess 1988; Rasmussen and Crosbie 1996). However, the tornado reports in *Storm Data* usually contain some information on the areal extent of damage (pathlength and -width). Unfortunately, similar information is virtually nonexistent for hail or damaging wind reports. Instead, the areal extent and magnitude of a large, continuous severe hail or wind event must be inferred from a number of isolated reports (with single times and locations listed). But since these reports are often widely

separated in space and/or time, it is rarely possible to know the full extent and intensity of damaging hail and/or wind produced by a storm cell on the timescales of a single radar volume scan (5–6 min). Changnon (1968) shows that observation density greatly affects the frequency of damaging hail reports, as well as whether or not damaging hail is observed at all on a given storm day. He states that a network comprising one or more observation sites per square mile is necessary to adequately measure the areal extent of damaging hail. Because of all this, evaluating algorithm predictions of the areal extent and intensity of severe weather events will likely prove to be very difficult.

The above comments have largely focused on the quantity of severe reports available from *Storm Data*, and the potential impacts on algorithm evaluation. However, there is also the issue of report quality. Witt et al. (1998) found that roughly 30% of the hail reports in *Storm Data* did not correlate well with storm cell locations from WSR-88D data. Herzog and Morrison (1994) present evidence that there may be a substantial high bias in reported hail sizes in *Storm Data*, relative to hail-size distributions observed during hail suppression experiments. Comparison of tornado observations made during VORTEX with the initial reports in *Storm Data* showed large differences. For four target storms observed by VORTEX personnel during 1995, there were nine tornadoes reported in *Storm Data* that were not observed by VORTEX (even though numerous intercept teams were in the vicinity of each storm). While scoring the performance of different mesocyclone and tornado detection algorithms (Stumpf et al. 1998; Mitchell et al. 1998) on 26 tornadic storm days, we noted many instances where the tornado reports in *Storm Data* did not correlate well with circulation signatures in the WSR-88D data. We estimated that the times of tornado occurrence were in error by $\geq$5 min for 38% of the reports, with locations estimated to be in error by $\geq$5 km for 12% of the reports. Large time errors ($>$40 min) were noted with 12% of the reports. There was a distinct ''late'' bias in the reported tornado times, which is likely due to differences between the actual time of the tornado and the time the report was relayed to the NWS. Some of the large time errors (of $\sim$1 h) may be due to mistakes in correctly converting UTC to LST (which is used in *Storm Data*). Some of the location errors may be due to recording the location of the storm spotter instead of the actual location of the tornado.

Despite all the problems noted above, researchers still need to evaluate the accuracy of severe storm detection algorithms. The evaluation and verification process is, in essence, a multidimensional matching problem (between severe weather forecasts and events). The better the match in dimensions, the better the verification. Since surface severe weather observations are taken at discrete points in time and space, a means of extrapolation is needed. Among the issues that have to be addressed by an evaluation procedure are 1) how to decrease the dimensionality (or extrapolate); 2) how to handle the representativeness of the reports (e.g., does the lack of any observation indicate no severe weather or just no observation); and 3) the relationship between severe weather reports (e.g., few hail reports near tornadoes). In the rest of this paper we describe some of the methods recently used for evaluating algorithms, and then give recommendations for what we believe are the best procedures for handling the first two of the above issues. We have not yet determined an appropriate way to handle the third issue, except to note that it is a concern.

## 3. Evaluation options

### *a. Use of field project data*

In those limited cases where special field project data are available, accurate evaluation can usually be done on an individual volume scan basis, using those time periods when definitive verification data were collected (including null observations). An example of this approach is the NCAR Hail Project, which was conducted to ensure a dataset adequate for verifying the performance of several hail detection algorithms. The main advantage of this type of evaluation is that the least amount of extrapolation is necessary between the forecasts and observations. Thus, the ensuing results, provided that the sample size is adequate, should be highly accurate. The main disadvantage of this option is that the results of the evaluation may only be applicable for a small geographical area.

### *b. Use of Storm Data*

#### 1) EQUALIZING TIMESCALES

Two methods are commonly used to attempt to equalize the timescales when comparing algorithm predictions to severe weather observations. The first, referred to as ''simulated warning evaluation'' [used by Winston and Ruthi (1986) and Witt (1990)], reduces the number of algorithm predictions that are scored. The second method, referred to as ''time window scoring,'' has been used by many investigators (Smart and Alberty 1985; Winston and Ruthi 1986; Winston 1988; Vasiloff et al. 1993; Mitchell et al. 1998; Stumpf et al. 1998; Witt et al. 1998). This procedure extends the time period over which a single severe report is used so that it can be associated with multiple algorithm predictions.

##### *(i) Simulated warning evaluation*

In this method, once an algorithm parameter that is used for predicting a severe weather event exceeds a threshold value, a ''simulated warning,'' with time duration and areal coverage of typical NWS warnings, is ''issued'' for the storm. All further algorithm predictions within the simulated warning area and time period are

then ignored. The main advantage of this method is that it mimics real severe weather warnings. Thus, factors affecting the relationship between severe reports in *Storm Data* and actual warnings should affect a performance evaluation using this method in a similar manner. The main disadvantage with this method is that it does not allow for evaluation of all the primary[2] algorithm predictions for a storm. In fact, a high percentage of the algorithm predictions that are discarded occur when storms are likely above severe levels. This may produce biases in the performance results. Also, for long-lived severe events, such as a long-track tornado, it is desirable for an algorithm to be predicting a high likelihood of the event for the entire time period that it occurs. However, with simulated warning evaluation, an algorithm could produce one initial correct forecast of the event, and then many incorrect forecasts while the event is still occurring, and be scored as totally correct (since the incorrect forecasts have been discarded while the simulated warning is valid). A different algorithm could also produce an initial correct forecast of the event, but then follow that with a continuous set of correct forecasts while the event is occurring and achieve the same score as the first algorithm. It is obvious that the second algorithm would be providing better guidance information to a warning forecaster versus the first algorithm, but this would not be revealed in the performance results. The alternate method, time-window scoring, is used to avoid this problem.

*(ii) Time-window scoring*

In contrast to the simulated warning evaluation method, the time-window scoring method allows one to evaluate all the algorithm predictions. This is accomplished by associating multiple algorithm predictions (over a specific period of time or number of volume scans) with a single severe report. To illustrate the process, take the example case of a storm that produces a tornado lasting for 10 min.: assume an algorithm is providing forecasts of the likelihood of tornadoes for this storm, and that the time window used for evaluation runs from 15 min before the start of the tornado until 5 min. after the end of the tornado (Table 1). In this example, the algorithm forecasts are provided as probabilities that are converted to categorical (''yes/no'') forecasts for scoring purposes by selecting a threshold value (50% in this case). Then, during the time-window period, each algorithm prediction giving a ''yes'' forecast of the event is counted as a ''hit,'' with each ''no'' prediction counted as a ''miss.'' Any ''yes'' predictions outside of the time window (for any observed tornado) are counted as ''false alarms.'' From these results, it is possible to calculate commonly used performance measures, such as the probability of detection, false-alarm rate, and critical success index (Wilks 1995).

TABLE 1. Example of time-window scoring. The tornado occurs from 2130 to 2140 UTC. POT is probability of tornado (in percent), CTF is categorical tornado forecast, H is hit, M is miss, and FA is false alarm. The temporal limits of the time window are 2115 and 2145 UTC.

| Time | POT | CTF | Result |
|---|---|---|---|
| 2100 | 50 | Yes | FA |
| 2105 | 60 | Yes | FA |
| 2110 | 40 | No | |
| 2115 | 30 | No | M |
| 2120 | 50 | Yes | H |
| 2125 | 70 | Yes | H |
| 2130 | 90 | Yes | H |
| 2135 | 40 | No | M |
| 2140 | 80 | Yes | H |
| 2145 | 50 | Yes | H |
| 2150 | 20 | No | |

In the above example, one might be puzzled as to why the time window is selected to extend beyond the time periods of the observed severe event. Several factors are involved in determining appropriate temporal limits for the time window. One factor to consider is the average lead time between precursor radar signatures and the start of severe weather at the ground. For tornadoes, Burgess et al. (1979) show an average lead time of 21 min. for mesocyclone signatures. For damaging downbursts, Eilts et al. (1996) show an average lead time of 10 min for several mid- and upper-altitude radar signatures. Hail located at midaltitudes in a storm (observable as an elevated region of high reflectivity) can take up to 10 min to fall to the ground (Changnon 1970). Because of this, a time window usually extends 10–20 min before the starting time of a severe report. The extra few minutes past the ending time of a report is to allow for synchronization errors between radar and observation times and the general uncertainties with report times.

Another reason for extending the time window by some amount of time prior to the severe event is that this allows for an indirect evaluation of an algorithm's lead time capability, which is particularly important to the NWS (Polger et al. 1994). Given the importance of lead time, one might question the validity of the false alarms shown in Table 1. It would appear that these predictions should instead be scored as correct, longer lead time forecasts. One way to do this would be to extend the time window in this example from 15 min to 30 min prior to the beginning time of the tornado. However, if the algorithm had instead given a no forecast for these two times in the example, then using a longer time window would lead to more misses being counted. Because of this dilemma, one may choose to generate multiple sets of performance results using time

[2] In some cases, more than one algorithm prediction is generated for the same storm on a given volume scan, resulting in primary and secondary predictions (see section 3b 2 for additional discussion).

windows of different lengths, in order to obtain a more comprehensive evaluation (e.g., Witt et al. 1998).

In situations where severe reports do occur at relatively high frequency or density, time windows from separate reports will occasionally overlap with the same algorithm prediction, leading to duplicate hits or misses. Since one of the effects of time-window scoring is to ''fill in'' the gap between isolated reports (that may be sampling the same severe event), these duplicate hits and misses should be discarded, so as to avoid producing artificial biases in the performance statistics.

### 2) SPATIAL CONSIDERATIONS

There are two primary spatial issues to be addressed when evaluating algorithms: the choice of analysis domain, and the defined size (spatial extent) of the algorithm predictions (compared to the typical areal coverage of a warning). Concerning the analysis domain choice, one could limit the analysis to only high-population areas (based on census data), where the occurrence of severe weather is least likely to go unreported. Although this option will greatly reduce the number of storm events analyzed, it will lead to higher confidence in the performance statistics generated. Wyatt and Witt (1997) present preliminary results showing a positive correlation between population density and algorithm performance results. However, it is possible that the analysis results obtained for high-population areas may not be applicable for other nearby areas where terrain characteristics and heights are substantially different. The other option is to analyze across the full radar domain (usually out to a range of 230 km). Although this will produce the largest sample size possible, the high likelihood of extensive rural areas means that the accuracy of the ensuing performance statistics will be questionable.

The other issue concerns how best to deal with situations when, for a given volume scan, more than one algorithm prediction is generated for a small area (e.g., for the same storm or county). If observations are not recorded at a similar spatial resolution, then some means of equalizing the space scales is needed. This is usually done by selecting some criteria for declaring a specific algorithm prediction as being primary, and discarding all other secondary predictions within some prespecified distance (i.e., not scoring the other predictions). For example, the WSR-88D vertically integrated liquid (VIL) algorithm produces output at a grid resolution of 4 km (Crum and Alberty 1993). Because of this high resolution, each storm will usually contain several individual VIL values. However, most studies evaluating VIL as a severe weather predictor have simply used the maximum value for the storm, and ignored the remaining values.

## 4. Conclusions

Evaluating the performance of radar-based severe storm detection algorithms is not a simple process, since one is rarely able to obtain a direct one-to-one correspondence between predictions and observations. Algorithm predictions are made on timescales of 5–6 min (every radar volume scan) for weather phenomena that may affect less than 1% of the area of a typical county. Severe weather observations, on the other hand, are sporadic in nature, and, in the case of *Storm Data*, often inaccurate. But, despite the observational problems and deficiencies in *Storm Data*, one can still conduct a meaningful evaluation if appropriate methods are used.

Based on the comments made above, we offer the following suggestions:

1) When possible, special field project data should be analyzed first and scored to produce a small, but highly accurate, evaluation.
2) If the above option is not possible, or if further analysis is desired, then one should determine if an adequate evaluation, using *Storm Data* for ground-truth verification, can be done when limiting the domain to high-population areas only. This has the advantage over the first option in that a wider geographical and seasonal sample can be obtained. However, the sample size will likely be fairly small, and one must still deal with the quality problems associated with *Storm Data*, although these may be lower in urban areas. This option may provide the best combination of accuracy in the performance statistics and geographical and seasonal coverage.
3) If neither of the above two options are possible, then one is left with performing an evaluation across the full radar domain. Although this will result in the largest sample size possible, the accuracy of the ensuing performance statistics will be questionable. This option is probably best used only for comparing the relative performance of different algorithms, and not for producing absolute measures of performance.
4) In order to allow for the comparison of performance statistics from one study to another, it would be best for all evaluations to contain at least one common scoring procedure. It is recommended that this one method be time-window scoring, using a window running from 15 min before the beginning time of reports to 5 min after the ending time of reports. The choice of 15 min (as one of the temporal limits) is based on an average 10–20-min lead time between the detection of precursor signatures and observed severe weather at the ground. Although an argument can be made for using 20 min instead of 15 min for evaluating predictions of tornadoes, using different time-window lengths for different types of severe weather would make comparison of such results (between different types of severe weather) more difficult. Of course, additional analysis can be done using other methods or temporal limits.

5) Anyone using *Storm Data* for algorithm verification should check the accuracy of each report by comparing its time and location with the corresponding radar data. If an error is suspected with a report (e.g., closest storm is 50 km away), an attempt should be made to correct the likely error (by adjusting the time and/or location of the report). Although one could choose to simply discard all reports that appear to be in error, based on the error frequencies indicated in section 2, doing so would result in the elimination of a large percentage of the reports. Failure to do adequate quality control on the verification data will likely lead to significant errors in the performance results.

*Acknowledgments.* We thank Don Burgess, Robert Maddox, and two anonymous reviewers for providing many useful comments that improved the manuscript. The ideas presented here were the result of an effort by NSSL and the WSR-88D OSF to develop a common methodology for scoring radar-based severe storm detection algorithms. This effort was partially supported by the WSR-88D OSF.

## REFERENCES

Amburn, S. A., and P. L. Wolf, 1997: VIL density as a hail indicator. *Wea. Forecasting,* **12,** 473–478.

Burgess, D. W., R. J. Donaldson Jr., T. Sieland, and J. Hinkelman, 1979: Final report on the Joint Doppler Operational Project (JDOP). Part I: Meteorological applications. NOAA Tech. Memo. ERL NSSL-86, NOAA, Boulder, CO, 84 pp. [NTIS PB80-107/88/AS.]

Changnon, S. A., 1968: Effect of sampling density on areal extent of damaging hail. *J. Appl. Meteor.,* **7,** 518–521.

——, 1970: Hailstreaks. *J. Atmos. Sci.,* **27,** 109–125.

Crowther, H. G., and J. T. Halmstad, 1994: Severe local storm warning verification for 1993. NOAA Tech. Memo. NWS NSSFC-40, 30 pp. [NTIS PB94215811.]

Crum, T. D., and R. L. Alberty, 1993: The WSR-88D and the WSR-88D Operational Support Facility. *Bull. Amer. Meteor. Soc.,* **74,** 1669–1687.

——, ——, and D. W. Burgess, 1993: Recording, archiving and using WSR-88D data. *Bull. Amer. Meteor. Soc.,* **74,** 645–653.

Doswell, C. A., III, and D. W. Burgess, 1988: On some issues of United States tornado climatology. *Mon. Wea. Rev.,* **116,** 495–501.

Eilts, M. D., J. T. Johnson, E. D. Mitchell, R. J. Lynn, P. Spencer, S. Cobb, and T. M. Smith, 1996: Damaging downburst prediction and detection algorithm for the WSR-88D. Preprints, *18th Conf. on Severe Local Storms,* San Francisco, CA, Amer. Meteor. Soc., 541–545.

Hales, J. E., Jr., and D. L. Kelly, 1985: The relationship between the collection of severe thunderstorm reports and warning verification. Preprints, *14th Conf. on Severe Local Storms,* Indianapolis, IN, Amer. Meteor. Soc., 13–16.

Herzog, R. F., and S. J. Morrison, 1994: Hail frequency in the United States. Haag Engineering Co. Rep., 18 pp. [Available from Haag Engineering Company, P.O. Box 814245, Dallas, TX 75381.]

Kelly, D. L., J. T. Schaefer, and C. A. Doswell III, 1985: Climatology of nontornadic severe thunderstorm events in the United States. *Mon. Wea. Rev.,* **113,** 1997–2014.

Kessinger, C. J., and E. A. Brandes, 1995: A comparison of hail detection algorithms. Final Rep. to the FAA, 52 pp. [Available from the authors at NCAR, P.O. Box 3000, Boulder, CO 80307.]

Mitchell, E. D., S. V. Vasiloff, G. J. Stumpf, A. Witt, M. D. Eilts, J. T. Johnson, and K. W. Thomas, 1998: The National Severe Storms Laboratory Tornado Detection Algorithm. *Wea. Forecasting,* **13,** 352–366.

Morgan, G. M., Jr., and P. W. Summers, 1982: Hailfall and hailstorm characteristics. *Thunderstorms: A Social, Scientific and Technological Documentary,* Vol. 2, *Thunderstorm Morphology and Dynamics,* E. Kessler, Ed., U.S. Government Printing Office, 363–408.

NSSL, cited 1997: WSR-88D Algorithm Testing and Display System. [Available online at http://www.nssl.noaa.gov/~watads.]

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.

Rasmussen, E. N., and C. Crosbie, 1996: Tornado damage assessment in *VORTEX-95.* Preprints, *18th Conf. on Severe Local Storms,* San Francisco, CA, Amer. Meteor. Soc., 153–157.

——, J. M. Straka, R. Davies-Jones, C. A. Doswell III, F. H. Carr, M. D. Eilts, and D. R. MacGorman, 1994: Verification of the Origins of Rotation in Tornadoes Experiment: VORTEX. *Bull. Amer. Meteor. Soc.,* **75,** 995–1006.

Smart, J. R., and R. L. Alberty, 1985: The NEXRAD hail algorithm applied to Colorado thunderstorms. Preprints, *14th Conf. on Severe Local Storms,* Indianapolis, IN, Amer. Meteor. Soc., 244–247.

Stumpf, G. J., A. Witt, E. D. Mitchell, P. L. Spencer, J. T. Johnson, M. D. Eilts, K. W. Thomas, and D. W. Burgess, 1998: The National Severe Storms Laboratory Mesocyclone Detection Algorithm for the WSR-88D. *Wea. Forecasting,* **13,** 304–326.

Vasiloff, S. V., G. J. Stumpf, A. Witt, P. L. Spencer, D. L. Keller, and M. D. Eilts, 1993: An evaluation of several mesocyclone and tornado detection algorithms versus ground verification. Preprints, *17th Conf. on Severe Local Storms,* St. Louis, MO, Amer. Meteor. Soc., 395–398.

Winston, H. A., 1988: A comparison of three radar-based severe-storm-detection algorithms on Colorado High Plains thunderstorms. *Wea. Forecasting,* **3,** 131–140.

——, and L. J. Ruthi, 1986: Evaluation of RADAP II severe-storm-detection algorithms. *Bull. Amer. Meteor. Soc.,* **67,** 145–150.

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

Witt, A., 1990: A hail core aloft detection algorithm. Preprints, *16th Conf. on Severe Local Storms,* Kananaskis Park, AB, Canada, Amer. Meteor. Soc., 232–235.

——, M. D. Eilts, G. J. Stumpf, J. T. Johnson, E. D. Mitchell, and K. W. Thomas, 1998: An enhanced hail detection algorithm for the WSR-88D. *Wea. Forecasting,* **13,** 513–518.

Wyatt, A., and A. Witt, 1997: The effect of population density on ground–truth verification of reports used to score a hail detection algorithm. Preprints, *28th Conf. on Radar Meteorology,* Austin, TX, Amer. Meteor. Soc., 368–369.