P2.9 EVALUATING MULTI-RADAR, MULTI-SENSOR HAIL DIAGNOSIS WITH HIGH RESOLUTION HAIL REPORTS

Christopher J. Wilson
National Weather Center Research for Undergraduates, University of Oklahoma,
Norman, Oklahoma and Valparaiso University, Valparaiso, Indiana

Kiel Ortega, Valliappa Lakshmanan
Cooperative Institute of Mesoscale Meteorological Studies, University of Oklahoma, and
NOAA/National Severe Storms Laboratory, Norman, Oklahoma

ABSTRACT

Low resolution verification data, as available from the *Storm Data* database, has hindered the development and evaluation of high resolution hail algorithms as well as the assessment of hail forecasting techniques. Previous studies have highlighted the inadequacies and inaccuracies associated with this verification data. This study uses high resolution ground-truth hail verification data from the Severe Hazards Analysis and Verification Experiment (SHAVE) to evaluate gridded synthetic hail verification and different radar derived parameters used in predicting severe hail.

MESH is found to have limited skill as a synthetic verification tool due to a high probability of false detection and a wide distribution of MESH values for each reported hail size range. In addition, radar-derived parameters are found to provide little skill in the prediction of severe hail as the probability of false detection associated with these parameters leads to low skill scores. The predictive skill of these parameters is also found to decrease with time, limiting the lead time in which skillful prediction of severe surface hail fall is possible using radar derived parameters.

1. INTRODUCTION

Large hail places both human safety and economic interests at risk. A single hailstorm can cause substantial damage to property, and can put individuals in its path at risk of injury. The ability of forecasters to accurately warn of potentially destructive hail and to provide warnings with meaningful lead time is of great importance. Forecasting techniques and hail diagnosis algorithms have been developed to facilitate forecasters' needs for severe hail guidance.

Witt et al. (1998) noted that efforts to evaluate hail forecasting techniques and algorithm performance have been hindered in the past by low resolution ground-truth verification data. Trapp et al. (2006) also documented *Storm Data* deficiencies. In addition, the National Weather Service (NWS) has moved from county-based warnings to polygon warnings, and currently, adequate verification data does not exist to score the accuracy of the geographic bounds of these polygon warnings. High resolution hail data from the Severe Hazards Analysis and Verification Experiment (SHAVE) (Smith et al. 2006) is utilized here to evaluate the feasibility of using the Maximum Expected Hail Size (MESH) algorithm as a synthetic verification tool as well as to evaluate the predictive skill of radar-derived parameters.

* *Corresponding author address:* Christopher J. Wilson, P.O. Box 148, Stronghurst, IL 61480;
e-mail: chris.wilson@valpo.edu.

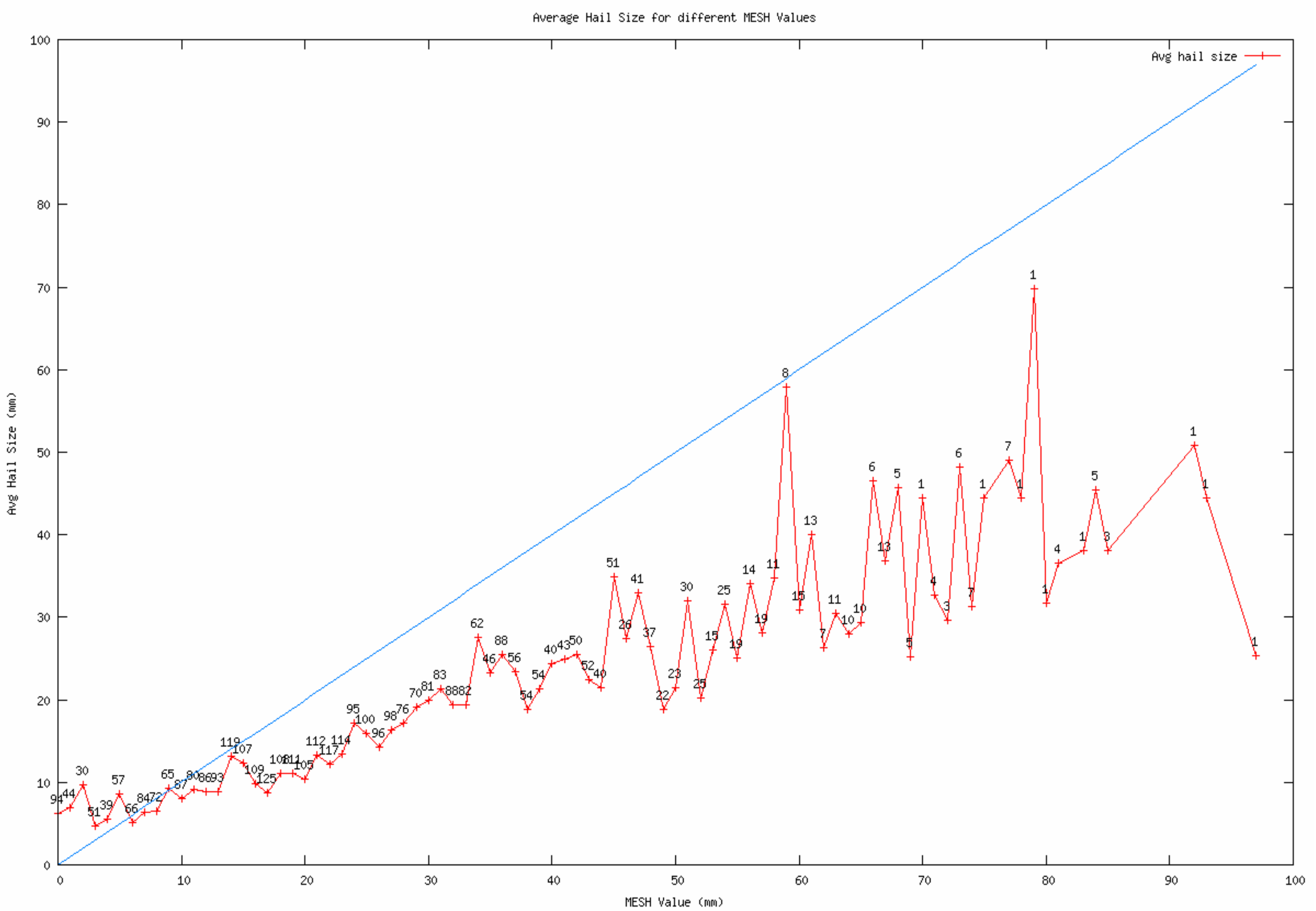

FIG. 1. Average hail size for each MESH threshold (at 1 mm increments) across all reports. The number of hail reports at each threshold is plotted. 3921 total reports are included (1539 severe/2382 non-severe).

2. HAIL DIAGNOSIS TECHNIQUES

Various methods have been employed over the years to forecast severe hail (>19 mm) in a dichotomous yes/no manner as well as to estimate the maximum hail size that can be expected at the surface. Donaldson (1959) investigated the relationship between the maximum height of storm echo tops and whether surface hail fall was reported. Geotis (1963) noted that sustained levels of enhanced radar reflectivity were related to maximum hail size. Greene and Clark (1972) stated that more intense radar returns caused by hailstones aloft could lead to vertically integrated liquid (VIL) values in excess of those calculated in the absence of hailstones. They noted this VIL discrepancy could be used as a means of assessing the severity of storms.

Edwards and Thompson (1998) investigated the usefulness of VIL in predicting hail severity. Though they found that large hail is not likely to be associated with low VIL values, they also showed high VIL values do not guarantee large surface hail fall. Their statistical results, coupled with the difficulties in calculating VIL within a radar's cone of silence, for fast moving storms, and for strongly tilted storms, suggest that VIL-based products are not skillful in predicting severe hail.

The WSR-88D hail detection algorithm (HDA) provides information regarding the probability of hail, probability of severe hail (POSH), and the maximum expected size of hail (MESH) (Witt et al. 1998). The HDA uses two weighting functions, one which weights reflectivity values and one which weights the heights of reflectivity values with respect to freezing level heights. Vertical integration of these weighting functions yields the severe hail index (SHI). Both MESH and POSH are derived products from the SHI. The HDA calculates the probability of hail based on the height of the 45-dBZ echo above the melting level. The POSH involves a thermal weighting in which echoes of 50-dBZ and greater above the -20C isotherm are weighted the greatest, and echoes less than 40-dBZ or below the melting level receive zero weight.

Donovan and Jungbluth (2007) showed that for severe hail, an increasing 50-dBZ echo height is required as the height of the melting level increases. They suggested that the height of the 50-dBZ echo indicates updraft strength, while the height of the melting level provides information about the environmental thermal profile, and they note that use of the 50-dBZ criterion may increase lead time by removing the delay associated with the calculation of volume-derived products such as VIL.

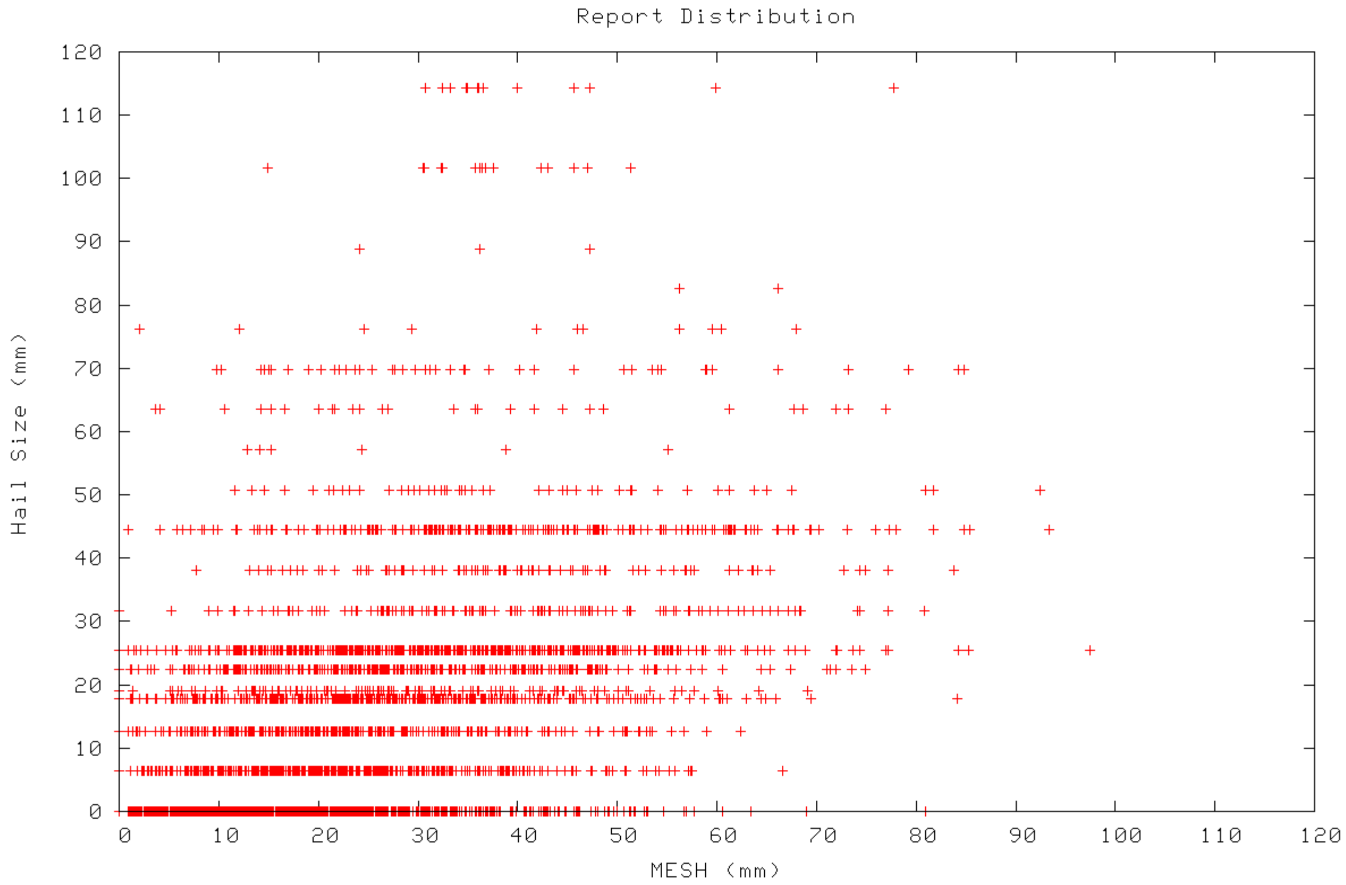

Fig. 2. MESH values associated with each reported hail size from SHAVE show the distribution of reported hail sizes for each MESH threshold (at 1 mm increments).

3. METHOD

One goal of this study was to evaluate gridded hail size prediction (Stumpf et al. 2004) as a synthetic verification tool. Initially, MESH swaths were plotted with SHAVE hail reports overlaid. The MESH swaths were comprised of the temporal maximum MESH value at each location. The MESH values were compared to the SHAVE reports on a point-by-point basis. Figure 1 shows the average hail size reported for each MESH threshold, and Figure 2 shows the MESH distribution for each reported hail size.

Next, MESH values were incremented at 1 millimeter size steps and compared to SHAVE reports at each location. Incrementing began at a MESH threshold of 0 mm and continued through an upper bound of 51 mm. MESH values at or above each threshold at each point were considered "yes" forecasts for severe hail. Figure 3 shows how this scoring was done. The ability of each MESH threshold to verify severe surface hail fall was evaluated. The results of this evaluation were used to populate a 2 x 2 contingency table with hits, misses, false alarms, and correct nulls.

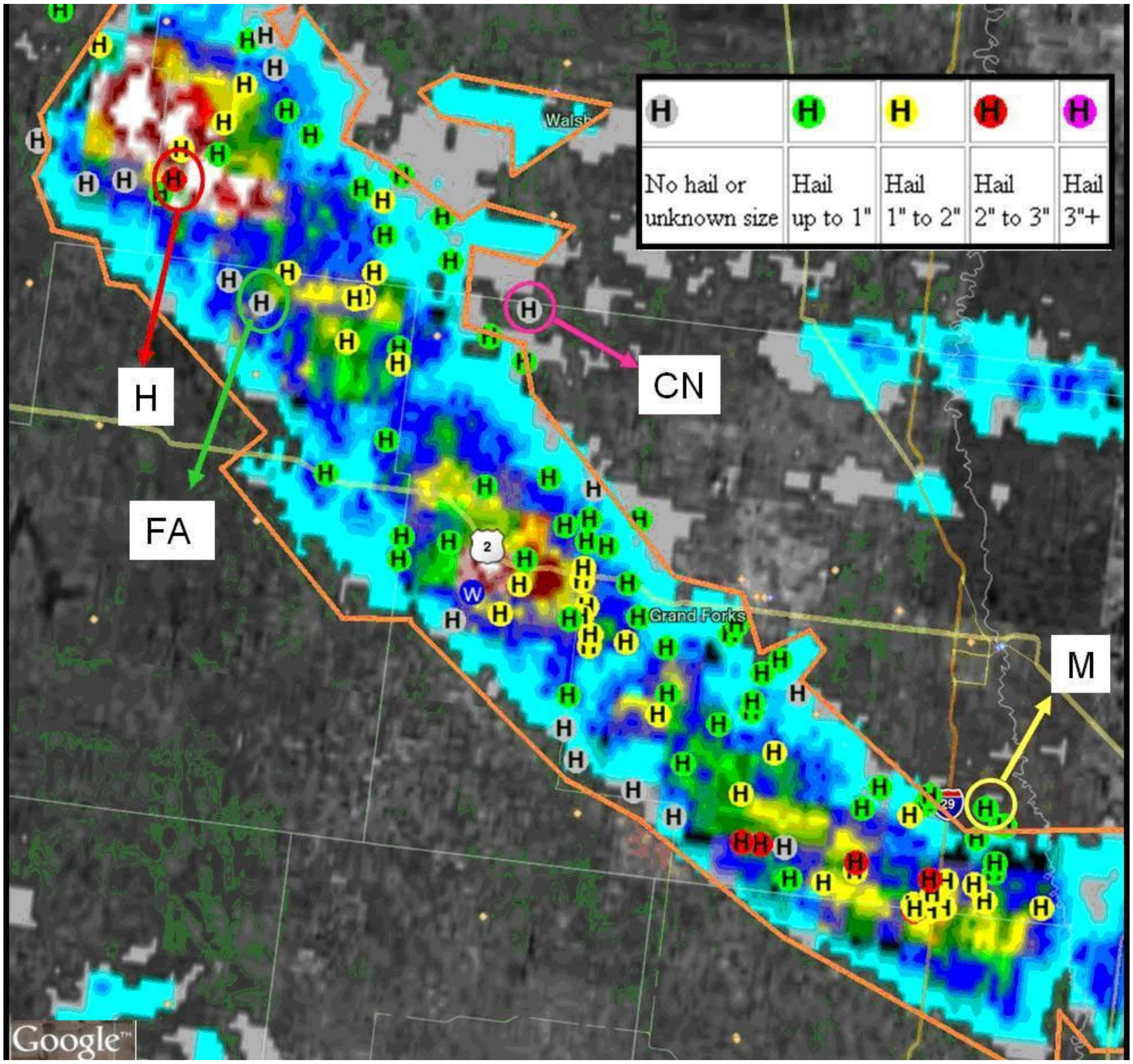

FIG. 3. MESH swath with SHAVE hail reports overlaid. Each small H corresponds to a SHAVE report. The orange outline represents a MESH threshold, and everything inside the contour is a "yes" forecast for severe hail. Severe hail reported in this region results in a hit (H) and severe hail reported outside of this region results in a miss (M). Non-severe hail inside the contoured region results in a false alarm (FA) and outside the region results in a correct null (CN). The MESH threshold is then incremented 1 mm at a time and similar scoring is done for each threshold.

Skill scores were then calculated in order to evaluate the use of MESH as a synthetic verification tool. The Critical Success Index (CSI) (Donaldson et al. 1975) was initially computed. CSI is a biased score which does not take into account correct forecasts of null events, and as a result is dependent on the frequency of severe hail fall events. The True Skill Statistic, or the difference between the probability of detection (POD) (Donaldson et al. 1975) and the probability of false detection (POFD) (Flueck 1987), and the Heidke Skill Score (Heidke 1926) were determined as well. The POFD is the ratio of the number of times an event does not occur when a “yes” forecast is made to the total number of times an event does not occur.

In order to evaluate the predictive skill of radar-derived parameters, SHAVE hail reports were gridded and compared to those parameters. Radar data and products such as MESH, VIL, POSH, reflectivity at isothermal levels, and height of reflectivity values above isothermal layers were also gridded. K-means clustering (MacQueen 1967) was used to identify storm cells on the grid and to gather storm attributes for each cluster. A cluster table was constructed relating the various parameter values at each grid location to the SHAVE hail reports. The maximum observed hail size in each cluster was recorded as well. K-means determined storm motions associated with each cluster, and time trends of various parameters were determined by advecting the clusters back along motion vectors by the desired number of frames and gathering the data values encompassed by the cluster at that time frame.

Figure 4 shows how lag values of each parameter were obtained. To attain lag values, the attributes of each storm of interest were retrieved from one time step earlier to yield lag 1 data, from two time steps earlier to yield lag 2 data, etc. For each parameter, lower and upper threshold bounds were specified. The threshold values between these bounds at each lag were used to evaluate the predictive skill of these parameters by calculating skill scores at each of 9 lags, with each lag step being approximately 5 minutes in length. Parameter values at or above the threshold value at each increment resulted in a “yes” forecast for severe hail. Our evaluation was based on each parameter’s skill in predicting severe hail. The skill of radar-derived parameters to more narrowly predict hail size beyond a severe/non-severe forecast was not assessed.

4. DATA

In this study, SHAVE hail reports were used. The objective of SHAVE is to gather high resolution hail, wind, and flooding data in space and time (Smith et al. 2006). Most importantly, SHAVE data also provides “no hail” reports and non-severe hail reports which is information not usually available using current climatological data.

Only isolated storms in the SHAVE database were selected to ensure that hail reports from multiple storm cells did not contaminate our analysis. A null case was selected which included multiple reports of no hail or non-severe hail as well. For the evaluation of MESH as a synthetic verification tool, 3921 reports were used (1539 severe/2382 non-severe), and 563 total reports (325 severe/238 non-severe) went into the evaluation of the predictive capability of radar-derived parameters. It is important to remember that even though SHAVE data is higher resolution, the possibility still exists that the largest hail size at a given location is either missed or incorrectly estimated.

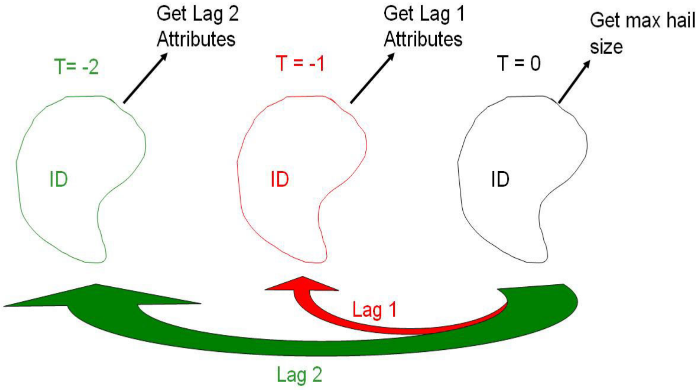

FIG. 4. This figure illustrates the method used to obtain lag values of parameters. “ID” is an identification number assigned to each cluster which allows the cluster to be tracked through time. T=0 represents “current” time, while T=-1 and T=-2 represent one and two lag steps back in time respectively. Each lag is approximately 5 minutes in length.

5. RESULTS

5.1 Gridded Synthetic Verification

Initially, a simple evaluation of the performance of MESH as a synthetic verification tool was

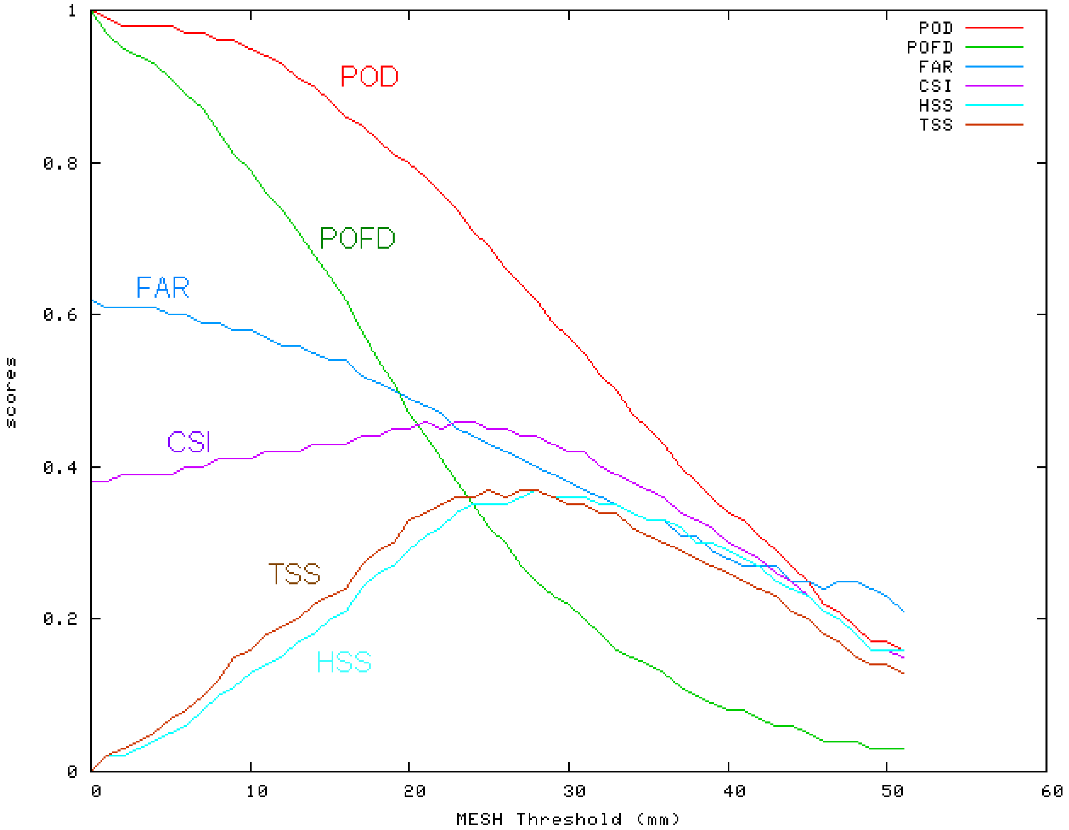

FIG. 5. Plotted are the probability of detection (POD), probability of false detection (POFD), false alarm ratio (FAR), Critical Success Index (CSI), Heidke Skill Score (HSS), and True Skill Statistic (TSS) for each MESH threshold from 0-51 mm. Scores are based on whether the MESH threshold can be used to create a binary hail prediction.

conducted by comparing the average size of SHAVE hail reports to maximum MESH values at each report location. The average hail size report for each MESH value with a y = x line overlaid is shown in Figure 1. The number of SHAVE reports that went into the calculation of each average is plotted for each data point as well. A bias of 9.28 mm and a root mean squared error of 20.34 mm were calculated. This bias suggests MESH usually overestimates the maximum size of hail that will reach the earth's surface. MESH is designed such that approximately 75% of hail will be smaller than the MESH (Witt et al. 1998), but the SHAVE reports being used represent the largest hail size reported at each given location so the relationship should ideally be one-to-one, and the curve should fall along the y = x line which is not what we discovered.

The distribution of hail sizes for each MESH value (at 1 mm increments) is shown in Figure 2. The spread of reported hail sizes that correspond to each MESH value is large. Hail reports ranging from 0 mm to 70 mm correspond to most individual MESH values between 0 mm and 60 mm. For instance, hail sizes of both 0 mm and 70 mm correspond to MESH of 15 mm. Many of the hail sizes between 0 mm and 60 mm correspond to 15 mm MESH as well, including other reports well in excess of 15 mm. This demonstrates that the 70 mm report is not simply an isolated event that stands out as an outlier. This distribution of hail sizes for each MESH value suggests that solely using MESH as a synthetic verification tool is not a feasible option.

The skill scores calculated by scoring how well a MESH threshold verifies severe hail on a point-by-point basis also suggest MESH should not be used alone as a means of synthetic verification. These skill scores are displayed in Figure 5. The Critical Success Index (CSI) peaks at a value of 0.46 at a MESH threshold of 23 mm. The Heidke Skill Score (HSS) and True Skill Statistic (TSS) both peak at 0.37 at a threshold of 28 mm. At the 28 mm threshold, the POD is 0.62 and the POFD is 0.25. At the threshold for severe sized hail (19 mm), the POFD is 0.51, and the HSS is only 0.27.

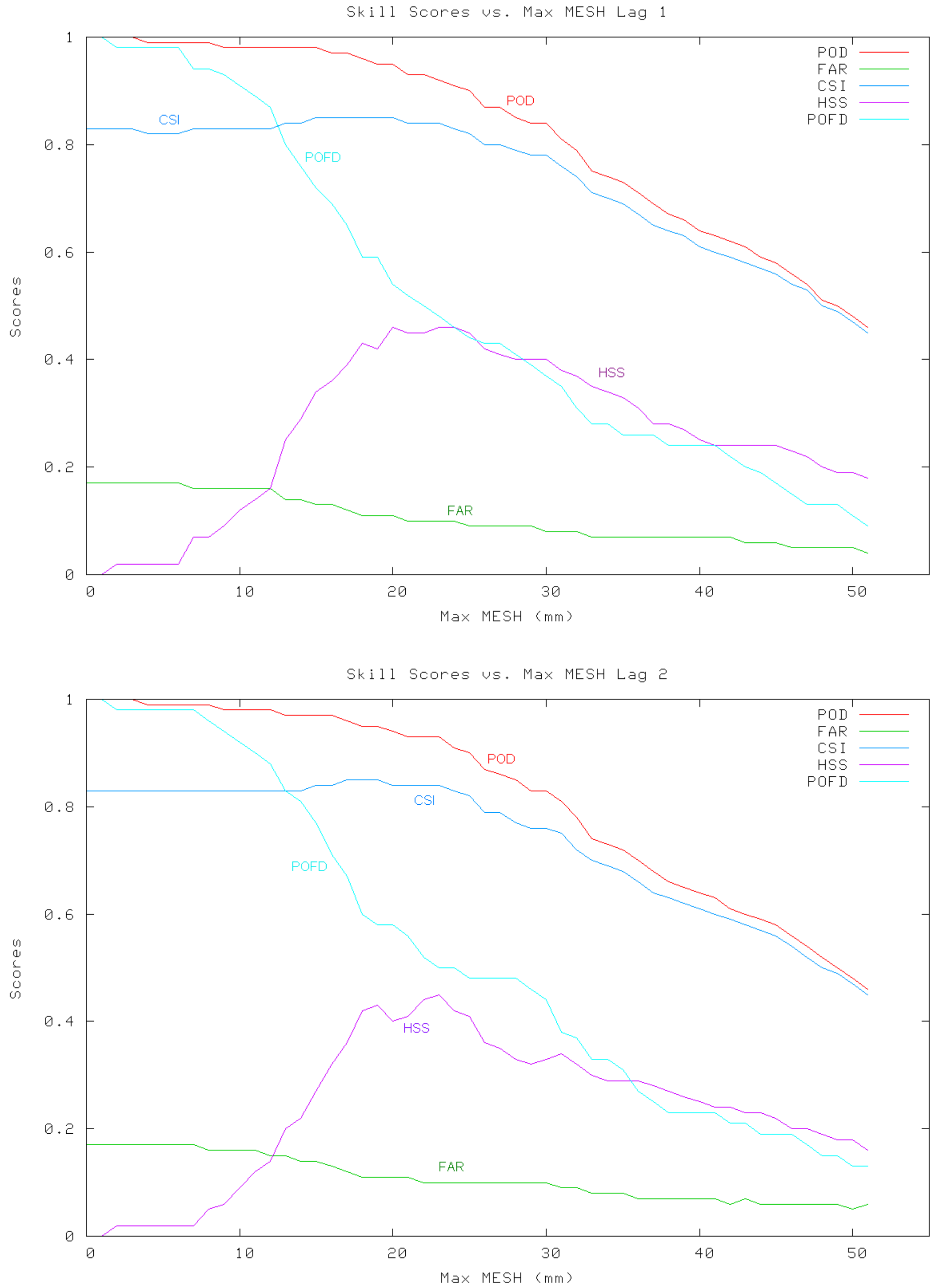

FIG. 6 (a) Skill score plot evaluating the skill of maximum MESH thresholds to predict severe hail at lag 1 (~5 minutes lead time); (b) same as (a) but at lag 2 (~10 minutes lead time).

These results confirm what was hypothesized from Figure 2, that the skill of MESH to determine locations of severe hail fall is greatest at MESH values greater than 19 mm. The low HSS and TSS values at all thresholds are a result of the high probability of false detection. The relatively low skill score values also suggest that MESH alone, while providing some skill, is not an adequate tool for synthetic verification.

5.2 Prediction Using Radar Derived Parameters

Time series analysis of radar-derived parameters was performed by comparing time trends of parameter values at different lags to hail reports. This analysis revealed little to no predictive skill associated with using time trends to predict severe hail.

The ability of radar-derived parameters to predict severe hail was determined by again calculating skill scores. The skill scores computed for the radar-derived products as well as the maximum MESH in each cluster at each of nine different lags suggest that the predictive skill of these parameters is limited. Looping from lower to upper bounds of each parameter and incrementing to incorporate intermediate values provided thresholds of each parameter at each of nine lags. The parameter values were then compared to the SHAVE reports on a point-by-point basis for each lag, and were scored based on their skill in predicting severe/non-severe hail at the surface. Plots were constructed to allow for visualization of the results. In no instance (any threshold at any lag) did the HSS exceed 0.5. Skill scores for maximum VIL, maximum MESH, and maximum height of 50-dBZ echo above the environmental melting level are shown. These were chosen as MESH is a newer tool still needing evaluation, and maximum VIL and the maximum height of 50-dBZ echoes above the melting level are commonly used to diagnose and predict severe hail operationally.

a. Maximum MESH

The maximum MESH in each cluster was increased at 1 mm increments from 0 mm to 51 mm over nine lag steps. Each lag represents approximately five minutes lead time. Figure 6a shows the skill of maximum MESH at lag 1 to predict severe hail was greatest for the 23 mm MESH threshold. At this threshold, the POD was 0.92, but the HSS was only 0.32. The HSS was low as a result of the POFD being high at 0.48. The HSS drops markedly on either side of the 23 mm peak value. Figure 6b shows at lag 2 the HSS again peaks at a MESH threshold of 23 mm and has increased to 0.45 with a POD of 0.93. This threshold/lag pair yields the greatest predictive skill score using maximum MESH to predict severe hail fall. The POFD at the 23 mm/lag 2 pairing is 0.5. This lag would provide about ten minutes lead time. The HSS for lags 3 through 9 remains below the lag 2 value and decreases with increasing lead time. The low skill scores across all lags and all thresholds of maximum MESH suggest maximum MESH provides limited skillful predictive capability.

b. Maximum VIL

Our results support the conclusions of Edwards and Thompson (1998) who stated that VIL offered little predictive skill. Here, maximum VIL was increased at 1 kgm^{-2} increments from 0 kgm^{-2} to 65 kgm^{-2} over the same nine lags as used for maximum MESH. The skill score results for using maximum VIL in severe hail prediction were the lowest of those for the three predictive parameters focused on in this paper. Figure 7 shows the lag 1 HSS peak of 0.31 occurred at 37 kgm^{-2} and corresponded to a POD of 0.84. The POFD at this point was 0.5. At lags 2 and 3, the HSS peak dropped to 0.25 at a threshold of 46 kgm^{-2}. Remaining lags yield HSS below 0.2. The decline in HSS values across all lags between thresholds of 30 kgm^{-2} and 35 kgm^{-2} is likely attributable to insufficient data. Our skill score results show that maximum VIL cannot consistently be used to skillfully predict severe surface hail fall with meaningful lead time. Issues associated with VIL calculation, including inadequate sampling of storms in a radar's cone of silence, fast moving storms, and strongly tilted storms, likely decrease the predictive skill of maximum VIL.

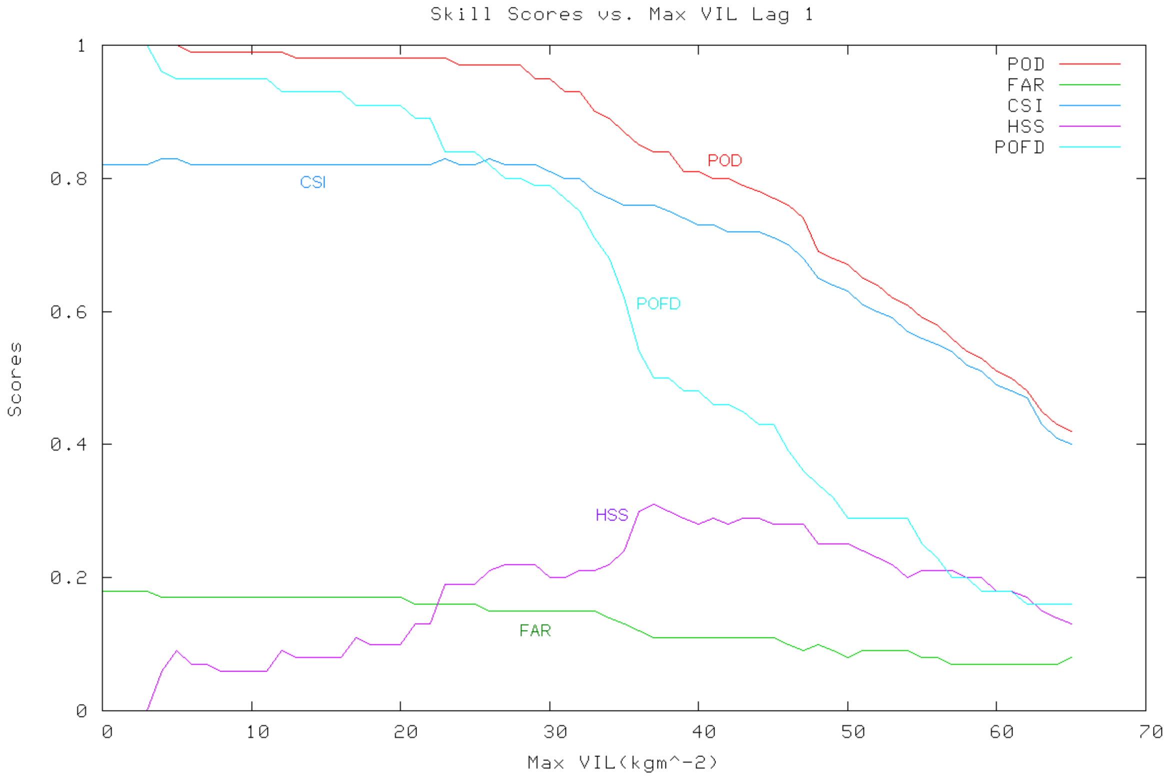

FIG. 7. Skill score plot evaluating the skill of maximum VIL thresholds to predict severe hail at lag 1 (~5 minutes lead time).

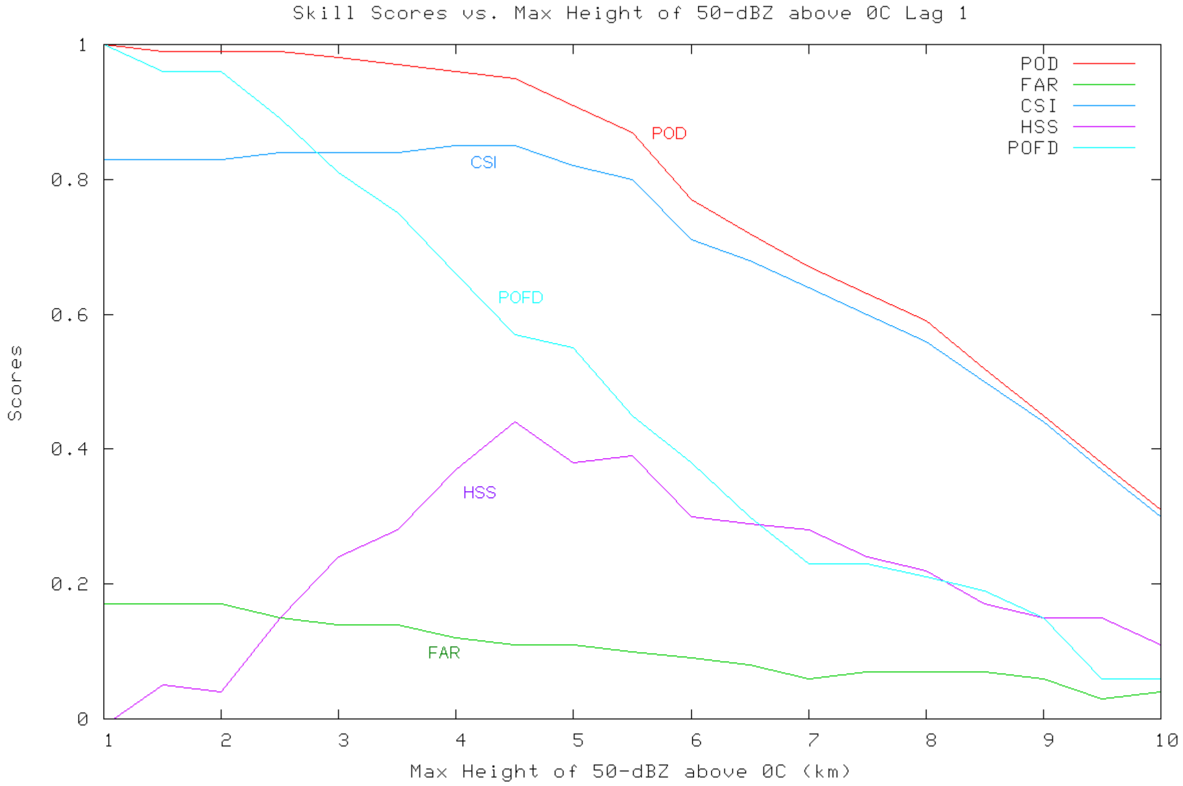

FIG. 8. Skill score plot evaluating the skill of threshold values of maximum height of 50-dBZ echo above the freezing level to predict severe hail at lag 1 (~5 minutes lead time).

c. Maximum Height of 50-dBZ echo above 0C

Evaluation of the predictive skill of the maximum height of the 50-dBZ echo above the melting level in predicting severe hail reveals a predictive capability similar to that of maximum MESH. This parameter was incremented between 0 km and 10 km using a 0.5 km step size. Figure 8 shows a CSI peak of 0.85 and HSS peak of 0.44 occur at lag 1 at a threshold of 4.5 km. The POD is very high at this threshold at 0.95, but the POFD is also high at 0.57. At lags 2 and 3, the HSS drops to 0.40 and 0.38 at a threshold of 4.5 km. At both of these lags, the POFD at the 4.5 km threshold at which HSS is maximized is above 0.60. The HSS continues to drop over lags 4 through 9. Again, low HSS values suggest that the maximum 50-dBZ echo height above 0C provides only limited skill in the prediction of severe surface hail fall.

6. CONCLUSIONS

SHAVE data offers a high resolution alternative to the low resolution hail verification data found in the *Storm Data* database. The enhanced temporal and spatial resolution of SHAVE data makes the evaluation of hail forecasting techniques and high resolution hail algorithms more complete. SHAVE data also includes reports of "no hail" as well as non-severe reports which make more comprehensive evaluation possible.

The relatively low skill scores computed in evaluating the ability of MESH to verify severe hail at the surface, as well as the distribution of MESH values for each reported hail size range, demonstrate that using MESH alone is not a feasible option for use as a synthetic verification tool.

Analysis of lag values of radar-derived parameters reveal that such parameters, across all thresholds at each of nine lags, offer limited skill in predicting severe hail fall at the surface. The probability of false detection associated with the predictions made by these parameters is often high and probably leads to low HSS and TSS values. Time trend analysis also shows that time trends in radar-derived parameters offer little skill in predicting severe hail.

Acknowledgements. The authors would like to thank Kevin Manross, Angelyn Kolodziej, and Arthur Witt for their helpful insight. This material is based upon work supported by the National Science Foundation under Grant No. ATM-0648566. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Donaldson, R.J., 1959: Analysis of severe convective storms observed by radar-II. *J. Atmos. Sci.*, **16**, 281–287.

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

Donavon, R.A., and K.A. Jungbluth, 2007: Evaluation of a technique for radar identification of large hail across the Upper Midwest and Central Plains of the United States. *Wea. Forecasting*, **22**, 244–254.

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

Edwards, R., and R.L. Thompson, 1998: Nationwide comparisons of hail size with WSR-88D vertically integrated liquid water and derived thermodynamic sounding data. *Wea. Forecasting*, **13**, 277–285.

Flueck, J.A., 1987: A study of some measures of forecast verification. Preprints, *10th Conf. Probability and Statistics in Atmospheric Sciences*, Edmonton, Alberta, Amer. Meteor. Soc., 69-73.

Geotis, S.G., 1963: Some radar measurements of hailstorms. *J. Appl. Meteor.*, **2**, 270–275.

Greene, D.R., and R.A. Clark, 1972: Vertically integrated liquid water—A new analysis tool. *Mon. Wea. Rev.*, **100**, 548–552.

Heidke, P., 1926: Berechnung des Erfolges und der Gute der Windstarkevorhersagen im Sturmwarnungsdienst. *Geografiska Annaler*, **8**, 301-349 (In German).

MacQueen, J.B., 1967: Some methods for classification and analysis of multivariate observations. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1:281-297.

Smith, T.M., K.L. Ortega, K.A. Scharfenberg, K. Manross, A. Witt, 2006: The severe hail verification experiment. *23rd Conference on Severe Local Storms*, Amer. Meteor. Soc., St. Louis, MO.

Stumpf, G.J., T. M. Smith and J. Hocker, 2004: New hail diagnostic parameters derived by integrating multiple radars and multiple sensors. Preprints, 22nd Conf. on Severe Local Storms, Hyannis, MA, Amer. Meteor. Soc., CD-ROM, P7.8.

Trapp, R.J., D.M. Wheatley, N.T. Atkins, R.W. Przybylinski, and R. Wolf, 2006: Buyer beware: some words of caution on the use of severe wind reports in postevent assessment and research. *Wea. Forecasting*, **21**, 408–415.

Witt, A., 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**, 286–303.