Hail Warning Decision Guidance

Michelle A. Harrold
National Weather Center Research Experience for Undergraduates, and Valparaiso University
Norman, OK, and Valparaiso, IN

James G. LaDue
NOAA/National Weather/Service Warning Decision Training Branch
Norman, OK

Paul T. Schlatter
Cooperative Institute for Mesoscale Meteorological Studies, and NOAA/National Weather Service/Warning Decision Training Branch
Norman, OK

Gregory J. Stumpf
Cooperative Institute for Mesoscale Meteorological Studies, and NOAA/National Weather Service Meteorological Development Laboratory
Norman, OK, and Silver Spring, MD

Corresponding Author Address:
Michelle A. Harrold
1050 East Flake Drive
Palatine, IL 60074
Michelle.Harrold@valpo.edu
(847) 987-8779

Abstract

National Weather Service (NWS) forecasters will have several new hail diagnostic attributes available to them in Advanced Weather Information Processing System (AWIPS; Wakefield 1998), beginning with the Operational Build version 6 (OB6). In a warning decision environment, it is essential that forecasters have the best guidance available to them. Therefore, the purpose of this study is to compare these new hail diagnostic radar parameters with legacy radar parameters to determine which are the "best predictors" for hail warning guidance. This study is the first step in developing official NWS Hail Warning Guidance training information for warning forecasters.

A total of 11 hail producing storm events were analyzed. The events chosen had geographic diversity across the United States and included both the warm and cool seasons. For each individual hail report, the values of 17 different hail diagnostic parameters were recorded. Each attribute was compared to ground truth reports and then statistically analyzed. Statistical analysis included the calculation of various correlation coefficients, as well as analyses of probability of detection (POD), false alarm rate (FAR), critical success index (CSI), and the Heidke skill score (HSS) for varying forecast decision thresholds, as well as different severe hail criteria (i.e., not just 2 cm diameter). Results indicated that the new high-resolution hail diagnostic radar parameters outperformed the legacy hail diagnostic parameters. Suggestions for future work to complete the development of NWS Hail Warning Guidance are offered.

1. Introduction

National Weather Service (NWS) forecasters will have several new hail diagnostic radar attributes available to them in the Advanced Weather Information Processing System (AWIPS; Wakefield 1998), beginning with the Operational Build 6 (OB6). Although these new attributes, along with some legacy hail diagnostic products and a variety of manual data analysis methods are currently in use, questions still arise as to which methods are the "best predictors" for hail warning decisions. The motivation for this study is to determine which hail diagnostic tools provide the best warning decision guidance to NWS forecasters.

Previous studies have looked at the different radar parameters to diagnose hail. Various studies done by NWS Weather Forecast Offices that focus primarily on a particular county warning area (Turner 1996; Amburn and Wolf 1997; Roessler and Wood 1997; Troutman and Rose 1998; Blaes, Cerniglia, and Caropolo 1998; Wallmann 2002). Other studies have also been conducted on a more national scale. (Edwards and Thompson 1998; Witt et al. 1998). Amburn and Wolf (1997) examined 221 different storms. Results from their study showed that a VIL density threshold of 3.5 g/m^3 had a probability of detection of 0.90. Certain limitations existed in this study. Events were confined to the Tulsa, Oklahoma County Warning Area (CWA) and were mainly warm season events. In another study, Witt et al (1998) focused solely on the Hail Detection Algorithm (HDA). Drawing from results extracted from previous papers, this study attempts to cover new ground on the hail warning decision guidance forefront by examining some legacy radar hail diagnostic tools as well as new, high resolution radar

hail diagnostic tools which will be made available to NWS warning forecasters in AWIPS OB6, and operational builds beyond.

In section 2, the hail diagnostic radar products will be explained. Next, the data and methodologies will be discussed in section 3. The resulting correlations and skill scores of the various products will be discussed in section 4. Finally, future work and concluding remarks are provided in section 5.

2. Hail Diagnosis Products

Overall, seven hail diagnostic radar parameters were used in this study. Table 1 gives a summary description of the characteristics of each of these parameters. Figure 1 shows examples of the 5 gridded products on an AWIPS D2D display. Information about each is detailed in the following sections.

a. Hail Detection Algorithm output

The Hail Detection Algorithm (HDA; Witt et al 1998) is one of the algorithms in Storm Cell Identification and Tracking Algorithm (SCIT; Johnson et al 1998). In the HDA, Probability of Severe Hail (POSH) and Maximum Estimated Hail Size (MEHS) are derived from the Severe Hail Index (SHI). The SHI is computed by an empirical formula that includes weighting functions for environmental temperature data and reflectivity data and is a quantity integrated in the vertical. Both the POSH and MEHS are cell-based products. Cell-based products rely on the accuracy of the SCIT-based algorithm, which tends to have problems during certain types of severe weather events. MEHS was designed to overestimate the hail size because the largest hail within a hail

swath is likely to go unreported (Witt et al 1998). POSH is also empirically derived based on the SHI, and in AWIPS, it is output from 0-100% in increments of 10%.

b. Vertically Integrated Liquid (VIL)

VIL is a radar-derived product that calculates the estimated amount of liquid water in a vertical column (Greene and Clark 1972). The VIL is computed using the following equation (Greene and Clark 1972):

$$VIL = \sum 3.44 \times 10^{-6} \left(\frac{Z_j + Z_{j+1}}{2} \right)^{4/7} \Delta h \qquad (1)$$

It should be noted that VIL, although often thought of as a hail predictor, does not account for ice when it is calculated. If $(Z_j + z_{j+1})/2>$, 56 dBZ, then $(Z_j + z_{j+1})/2$ is set to 56 dBZ in an attempt to remove the contribution of ice to the product output (Stumpf et al 2004). There are two VIL products available in AWIPS. One is a 4 km x 4 km Cartesian grid (VIL), and the other is a slightly newer 1° x 1 km polar grid product, known as "Digital VIL" (DVL).

As with all warning products, there are strengths and limitations for both VIL and DVL. Some storms have tilted reflectivity cores, and since each VIL product vertically-integrates dBZ in columns of different horizontal resolutions, their outputs will be different (Stumpf et al 2004). A forecaster might also expect to see higher 4 km VIL values than DVL values because the 4 km VIL product has a much greater probability of being in the same grid-space as the maximum reflectivities (Stumpf et al 2004). Additionally, unlike the low-resolution VIL product, the high-resolution product *does not*

cap reflectivities at 56 dBZ, resulting in possible DVL values well over 200 g/m^2. However, the final DVL values output are truncated at 80 g/m^2 in AWIPS; therefore, if there is a part of the storm that has actual values above 80 g/m^2 (which because of no cap at 56 dBZ, could happen quite often), the outputted value would only be 80 g/m^2. This will cause misrepresentative DVL values for stronger storms.

c. VIL Density

VIL density is calculated by dividing the value of VIL by the height of the echo top:

$$\textbf{VIL density (g/m3) = 1000 * VIL (kg/m}^2\textbf{) / echo-top (m)} \qquad (2)$$

VIL density was derived from VIL in order to account for varying storm heights and the environment (Amburn and Wolf 1997). Amburn and Wolf (1997) normalized VIL so that VIL density values would increase as hail cores increase both intensity and depth. There are three new VIL Density products available in the System for Convective Analysis and Nowcasting (SCAN; Smith et al. 1999) for AWIPS OB6. They are: 4 km x 4 km VIL density (VILD), 1° x 1 km "Digital" VIL Density (DVILD), and an "enhanced" VIL density (EVILD). All three are grid-based.

The 4 x 4 km VIL density (VILD) uses the 4 x 4 km VIL and the ET. The ET is a Weather Surveillance Radar – 1988 Doppler (WSR-88D) an Open Radar Product Generator (ORPG) product and is calculated as the height of the highest elevation scan

where the reflectivity is greater than or equal to 18 dBZ, given the vertical reflectivity profile over a 4 km x 4 km grid (Stumpf et al. 2004).

The 1° x 1 km Digital VIL density (DVILD) uses the ORPG 1° x 1 km DVL product and the enhanced echo top (EET) product. The EET product computes the echo top on a 1° x 1 km polar grid (Stumpf et al 2004). It too is the height of the highest elevation scan where the reflectivity is greater than or equal to 18 dBZ, however, better vertical interpolation is done to remove the echo top "rings" that were prevalent in the legacy ET product.

A third VIL density product, known as the Enhanced VIL density, combines the 1° x 1 km DVL with a "dilated" EET product. Amburn and Wolf (1997) found that, on occasion, the largest VIL and ET values would not be vertically aligned on the same grid location, due to the tilt of the storm core. This tilt can be caused either by actual tilted storm cores in high shear environments, or by spurious tilt caused by fast storm motion (as upper level elevations are scanned later in time as the storm moved forward). To rectify the tilt, the EET values were dilated to better ensure that the maximum EET and DVL values were vertically juxtaposed. Morphological dilation is the process used and in essence means that the maximum values of a gridded field are "spread out" isotropically a few pixels from their original location. An advantage of dilation is that it implicitly accounts for storm tilt, but a disadvantage is that it combines smaller cores in storms that are in a close proximity (Stumpf et al 2004).

3. Data and Methodology

To produce the needed hail diagnostic attributes, WSR-88D "Level II" radar data was fed into the ORPG to produce the legacy hail diagnostic parameters. Subsequently, the ORPG products were used within the SCAN processor to create some of the newer, high resolution hail diagnostic parameters. Finally, using the AWIPS Display Two Dimensions (D2D) software, the hailstorms were visually analyzed. For each of the hail events, the ground truth information and the hail diagnostic parameters were recorded and entered into a spreadsheet. These data were then statistically analyzed using a number of methods, to determine "best predictors".

A total of 11 hail producing storm events were analyzed (Table 2). Hail ground truth reports came from the National Climatic Data Center's (NCDC) *Storm Data*. Overall, there were 114 hail reports, but only 101 of the 114 could be used due to errors in the reports. The events chosen had geographic diversity across the United States, and included events from both the warm and cool seasons. For each storm cell, several legacy and new hail diagnostic parameters were recorded. These parameters are products from either the WSR-88D ORPG or the AWIPS SCAN.

There are some problems that arise when using *Storm Data* in a verification process (Amburn and Wolf 1997). Often with *Storm Data*, a time or location attached to a storm report is reported inaccurately. Also, depending on trained storm spotter density, there was a lack of reports for some events. Due to the lack of reports for some events, instead of throwing a report away, the time or location was corrected if it appeared to be slightly off from the location of the storm (Witt et al. 1998). This is to ensure that no report is wasted. Another issue surrounding *Storm Data* is once a report of hail that fits

severe criteria and verifies a warning, other severe hail within the area or county sometimes is not reported (Witt et al. 1998). This causes some reports to be omitted.

In an attempt to account for inaccuracies in storms reports, three values were recorded for each product (Fig. 2). First, the value of the product was recorded in the same volume scan and location of the hail report. The second was the maximum value of the product within a 5 km radius of the hail report at the time of the hail report. For the third, the maximum value of the product within a 5 km radius of the report in a 20 minute time window was recorded. These three values will be denoted as actual, maximum radius, and time window from this point onward. For POSH and MEHS, only the value at the same time as the report was recorded.

Several reasons exist for recording 3 pieces of information for each product. The 5 km radius was instated to account for poor storm report placement as well as non-vertical hail trajectories. The time window chosen was 20 minutes; 15 minutes before the report and 5 minutes after the report, hereafter denoted (-15, +5). This, in terms of radar volume scans, is roughly three volume scans before and one volume scan after a hail report. Previous studies on hail diagnostic tools also have incorporated a time window. Amburn and Wolf (1997) chose a time window that included two volume scans previous to the hail report. Witt et al (1998) chose to use two windows: a 60 minute (-45, +15) and a 20 minute (-15, +5). The time window was implemented to account for faulty storm reports and allow time for hail to develop aloft and descend (Witt et al. 1998). Hailstones can take as long as 5 minutes or as short as 90 seconds to fall 10,000 feet (Knight and Knight 2001). Thus, it is imperative that some sort of a time window be incorporated, although the actual length of the time is subjective due to the inaccuracies

of hail reports. The present study incorporated only the (-15, +5) rule because extreme care was taken to match the hail report to each grid based product.

Two different thresholds were varied in the study. The criteria for defining severe hail were varied as to increase the amount of null events in the data set. Severe hail criteria were varied from 2.5 – 5.0 cm in increments of 0.5 cm. In addition to the definition of severe hail, the warning decision threshold was also varied for the 17 different radar product values. A 2 x 2 contingency table for forecasts and observations was implemented to follow Brooks' (2004) detection methods (Table 3). Observations were classified to fit as either "yes" or "no" depending on the choice of the established severe threshold (Brooks 2004). Forecasts were classified as either "yes" or "no" depending on which radar product warning decision threshold values were chosen.

From the 2 x 2 table, the following measures were computed: probability of detection (POD), false alarm rate (FAR), critical success index (CSI), and Heidke skill statistic (HSS). The equations can be seen in Table 3. HSS was also looked at with the most regard due to the fact that it incorporated all four entries in the 2 x 2 table, including the correct forecasts of null events (which CSI does not use). HSS is also used to determine a good forecast decision threshold. An optimal HSS curve would be bell-shaped with the maximum point being the "best" maximum performance of a parameter.

Besides skill scores, the Spearman rank correlation coefficient was found for each of the 17 different hail products. Spearman rank correlation coefficient was chosen because the data set did not have continuous readings; rather, it consisted of different reports from different times and locations. Correlation coefficients were used to evaluate

which hail diagnostic tools as well as the different values (actual, maximum, and time window) related to the actual hail size.

4. Results and Discussion

The results showed that no single radar hail diagnostic attribute was far superior to the rest; however, some were better performers. One clear result is that the higher resolution products tended to outperform the lower resolution products. The higher resolution products tended to have the highest HSS maximums as well as some of the higher correlation coefficients between the actual hail size and the product value.

The Spearman rank correlation coefficients showed that the higher resolution products had the stronger correlations to hail size. All 101 reports were used when calculating the correlation coefficients. Figure 3 shows the correlation coefficients for the 17 different products. For the higher resolution products the correlation coefficients increased accordingly for each hail diagnostic tool from the actual to the maximum to the time window. The lower resolution products did not have correlation coefficients higher than 0.4, showing that the correlation between low resolution products and hail size is quite low. While the MEHS performed on a level similar to the high resolution products, the POSH performed on a level similar to the lower resolution products. Note that while certain predictors have the high correlations in this particular data set, overall, the actual values of the correlation coefficients are not overly impressive. With a maximum correlation coefficient of 0.6181, and a perfect correlation being 1, this shows the data

have some decent correlations to products and hail size, but these values should be kept in perspective.

Results from the maximum HSS for each product for each threshold can be seen in Figure 4. The maximum HSS for all thresholds and all products was 0.448 which was for the enhanced VIL density in the maximum category for the 2.5 cm threshold. As the case with the correlation coefficients, the higher resolution products tended to have the higher HSS. Also, as the severe threshold became larger, some 4 km VIL and 4 km VIL density HSS values became negative. This would imply that the forecast decision should be reversed. These results also suggest that the higher resolution products tend to perform better than the lower resolution products. The HDA also performed similarly, with the POSH performing closer to the low resolution products and the MEHS performing closer to the higher resolution. Similar to the correlation coefficients, the maximum HSS may appear to be high on this data set, but it is again important to keep in mind that these maximums are relative. Also, not only is it important to look at the maximum value, but also the behavior of the HSS across the varying decision thresholds as well (i.e., the shape of the HSS curve).

As previously mentioned, the 1 km VIL currently has a maximum value of 80 g/m^2. An interesting result of this study showed some of the problems with capping the 1 km VIL at 80 g/m^2. Figure 5 shows an example of how the POD, FAR, CSI, and HSS appear graphically. As mentioned, a typical HSS curve represents a bell curve; however, in the 1 km VIL and 1 km VIL density, the maximum HSS coincides with the largest value of 1 km VIL available. This leads to many questions as to what the actual maximum would be and what the shape of the HSS curve would look like if the cap was

removed. In other words, might the best 1 km VIL value be greater than 80 g/m^2? Since other products like DVILD and EVILD use the 1 km VIL, their skill score graphs behave in similar ways. It would be a good recommendation to lift the cap on the 1 km VIL in order to determine the optimal warning decision thresholds, and to add value to the use of these products for hail warning guidance.

Figure 5 also illustrates the "inequitibility" of the CSI measure. This is when the CSI does not change much and remains near its highest value for a range warning decision thresholds near the left or right side of the distribution. Using some of these hail diagnostic parameters, forecasters could maximize skill using the CSI measure by issuing warnings for every storm event, since the CSI can be maximized for the lowest value of the decision threshold. Perhaps the inequitibility of the CSI is due to the lack of many null events, but nevertheless warning solely on skill scores is cautioned.

Results were compared to Amburn and Wolf's (1997). They concluded that a 4 km VIL density of 3.5 g/m^3 had a POD of 0.90 using 0.75 in (19mm) as the severe hail size threshold. Using the 2.5 cm severe hail criteria threshold, this study found similar results. The 4 km VIL density time window value of 3.3g/m^3 had a POD of 0.88 (Figure 6), though with a higher severe hail criteria the threshold would be expected to be higher than the 3.5 g/m^3 from the Amburn and Wolf (1997) study. The maximum HSS was around 0.203, again showing that HSS were not overly impressive.

It is difficult to determine which one product is the "best predictor." There is no one "best predictor" because no single predictor consistently shows the best skill scores or correlations upon changing warning or severe hail thresholds. From all the different skills scores and correlation coefficients, the most definitive results that were reached

were that the higher resolution products outperformed the lower resolution products. With the HDA, the MEHS consistently outperformed the POSH.

5. Conclusion

The main motivation of this study was to determine which radar hail diagnostic tool or tools provided NWS forecasters with the best hail warning decision guidance. Overall, a total of 11 hail producing storms yielding 101 useable hail reports were evaluated. After analyzing different severe hail thresholds and hail diagnostic products as well as correlation coefficients and skill scores, it can be seen that there is not one parameter that is distinctly superior. The higher resolution products did tend to perform better than the lower resolution products, but there is not one particular product from the higher resolution products that stands out.

Since three values were found for each grid-based product (actual, maximum radius, and time window), results show that not only do higher resolution hail diagnostic tools perform better, but there are also high correlations between the size of hail report and the maximum radius and time window values. This may be due to the fact that it allows for the trajectories of hailstones as they fall to the ground.

There are many ways in which this study could be expanded and the potential is promising. To have a more-thorough study, more cases would need to be added. In those added cases, an emphasis should be put on the collection of correct of null events. One of the reasons for inconclusive results in this study resulted from the fact that null events not reported in *Storm Data*. By adding more of these types of events the database

would more closely match the natural distribution of hail sizes. One way of overcoming this issue could be to implement the use of NSSL Warning Decision Support System-Integrated Information (WDSS-II). Using WDSSII, a choice of storm cell detection algorithm could be used, in concert with high-resolution population density data, to cull out storm events that may be candidates for null events. The access to a population database would help in the confidence level when asserting that a null event is truly a null event (e.g., a storm cell moving over a city, but no severe hail reports are made).

In addition to recognizing the lack of correct nulls, the current study could add other hail diagnostic tools. Such products could include height of the 50 dBZ ET, reflectivity at 0°C and -20°C constant temperature altitudes, height of 50 dBZ reflectivity above 0°C and -20°C constant temperature altitudes, "sub-freezing" VIL, and a gridded version of HAD (e.g., gridded MEHS). Some of these new products are slated to be included in AWIPS Operational Build 6.1, which will be available for WFO warning operations in the spring of 2006. A goal would be similar to that of this research: examine other hail diagnostic tools and develop additional hail warning guidance prior to field implementation.

Acknowledgements. This research was made possible through the National Weather Center Research Experience for Undergraduates (NWC REU) which was funded by National Science Foundation (NSL) Grant 0097651. The first author would like to thank James LaDue, Paul Schlatter, and Gregory Stumpf for their helpfulness and insight throughout this project. Other thanks go to Daphne Zaras, Dr. Wilson Gonzalez-Espada, the REU staff, and the REU participants that made this experience possible.

References

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

Blaes, J. S., C. S. Cerniglia, and M.A. Caropolo, 1998: VIL density as an indicator of hail across Western New York and Eastern New England. *NWS Eastern Region Technical Attachment 98-8.*

Brooks, H. E., 2004: Tornado-warning performance in the past and future: a perspective from the signal detection theory. Bull. Amer. Meteor. Soc., **6**, pp. 837-843

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.

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

Johnson, J. T., P. L. MacKeen, A. Witt, E. D. Mitchell, G. J. Stumpf, M. D. Eilts, and K. W. Thomas, 1998: The Storm Cell Identification and Tracking (SCIT) algorithm: An enhanced WSR-88D algorithm. *Wea. Forecasting*, **13**, 263-276.

Knight, Charles A., and Nancy M. Knight, 2001: **Hailstorms**. *Severe Convective Storms*, AMS Meteorological Monographs, pp. 223-254.

Roessler, C. A., and L. Wood, 1997: VIL density and associated hail size along the northwest gold coast. *28th Conf. on Radar Meteorology*, Austin, TX, Amer. Meteor. Soc., 370-371.

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

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

Troutman, T. W. and M. Rose, 1997: An examination of VIL and echo top associated with large hail in Middle Tennessee. NWS *Southern Region Technical Attachment SR/SSD 97-15*.

Turner, R. J.: VIL vs. VIL density: A preliminary study for large hail in the NWSO Goodland, KS county warning area. *NWS Central Region Applied Research Paper 17-09.*

Wakefield, J. S., 1998: Operational Risk Reduction: Easing AWIPS into the Field. *Preprints,* 14*th Intl. Conf. on Interactive Information and Processing Systems for Meteor., Oceanography, and Hydrology*, Phoenix, AZ. Amer. Met.. Soc., 389-391.

Wallmann, J. H., 2002: VIL density as a potential hail indicator across Northeast and Central Nevada. *NWS Western Region Technical Attachment 02-10.*

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.