Ortega, K. L., 2018: Evaluating multi-radar, multi-sensor products for surface hail-fall diagnosis. *Electronic J. Severe Storms Meteor.*, **13** (1), 1–36.

Electronic Journal of
SEVERE STORMS METEOROLOGY

Evaluating Multi-Radar, Multi-Sensor Products for Surface Hail-Fall Diagnosis

KIEL L. ORTEGA
University of Oklahoma, Cooperative Institute for Mesoscale Meteorological Studies, and NOAA/OAR National Severe Storms Laboratory, Norman, OK

(Submitted 29 March 2017; in final form 18 February 2018)

ABSTRACT

The operational deployment of the multi-radar, multi-sensor (MRMS) system has made available new products to use for hail detection. MRMS products are provided on a spatial grid and can give information on hail size and the spatial extent and distribution of the hail fall. This information is important to a wide audience, including warning forecasters needing to focus on areas for warning verification and insurance users needing to verify a claim. Products are typically verified and evaluated using hail reports from *Storm Data*, which are reports collected by local National Weather Service Offices. The National Severe Storms Laboratory conducted a project to collect reports of hail, including reports of "no hail" near storms, at high spatial resolution. This project, the Severe Hazards Analysis and Verification Experiment (SHAVE), collected tens of thousands of hail reports over ten years of operations. Three-hundred eighty-nine SHAVE operations, which yielded 21 184 SHAVE reports and 2814 *Storm Data* reports, are investigated. Nine MRMS products were evaluated with the reflectivity at lowest altitude demonstrating the best discrimination for where hail of any size fell and the maximum expected size of hail product provided the best discrimination for severe-sized hail. SHAVE- and *Storm Data*-based evaluations showed marked differences in product skill scores. Discussions on the differences between the hail report databases and explorations of vertical profiles of reflectivity are included.

1. Introduction

a. Identifying surface hail fall

Observing the maximum hail size at the surface is important for understanding hail growth processes. In addition, the degree of damage is likely dependent on the maximal hail diameter (Smith and Waldvogel 1989). However, the importance of observing the totality of surface hailstreaks, spatially continuous areas of hail with temporal coherence, and hailswaths, multiple hailstreaks within certain spatial and temporal bounds, goes beyond simply ensuring the maximum hail size is captured. Changnon (1970) summarized that hailstreak information is important for confirming hailstorm mechanics, developing storm models, assessing radar capabilities to detect hail, and to develop climatologies. There is a large degree of difficulty in sampling not only the largest stone, but also the hailswath in general, because variations in hail size can occur over a few hundred meters (e.g., Morgan and Towery 1975). Changnon (1968) concluded a network density of 1 observation per 2.59 km^2 was necessary to adequately observe the areal extent of damaging hail.

The use of radar to determine the spatial extent of hailswaths has been previously explored. Basara et al. (2007) used a proprietary radar-based algorithm to accumulate hailswaths across the Southern Plains of the United States and illustrated the advantage of using spatial grids instead of single points within an analysis. Cintineo et al. (2012) used 4.5 y of radar data to compile a climatology of hailswaths across the United States. The radar results were compared to a report-based climatology over the same time

Corresponding author address: Kiel L. Ortega, 120 David L. Boren Blvd, Norman, OK, 73072, Email: kiel.ortega@noaa.gov

period, which showed the largest differences in calculated hail frequency occurring in lesser populated areas of the Plains states and Western North Carolina. More practical applications of radar-based hail detections have occurred using insurance claims (e.g., Schuster et al. 2006, Brown et al. 2015) and have shown a poor relationship between radar derivatives and damage claims.

b. Hail detection and radar

The return power of a target volume is dependent on aspects of not only the radar system collecting the data, but also how much radiation is scattered by the collection of particles in the target volume towards the radar. This is quantified by the backscattering cross section, which depends on the scatterers' size, liquid water content and shape. Radar reflectivity factor is the sum of the backscattering cross-sections in the target volume. Thus, only using radar reflectivity factor to diagnose complex scatterers such as hailstones is obviously limited considering that hailstones can have a range of water coatings; can be mixed with liquid water drops; and can have complex geometries. All of which can cause different backscatter to the radar that can make hailstones have a range of radar reflectivity factor values.

With knowledge of the above limitations, the importance of high radar reflectivity factor (hereafter simply reflectivity) within storms in general, and particularly at high altitudes, to identify hail-producing thunderstorms has been long established (e.g., Donaldson 1961; Geotis 1963). The relationship between reflectivity at the lowest elevation scan and hail has been studied for nearly as long (Dye and Martner 1978). For the WSR-88D,, several different reflectivity-derived products and techniques have been investigated. Vertically integrated liquid (VIL; Greene and Clark 1972) was normalized by storm top height by Amburn and Wolf (1997) to produce VIL Density (VILD) for use in detecting hail. The study found that a VILD of at least 3.5 g m^{-3} was associated with 90% of storms producing hail larger than 19 mm. Edwards and Thompson (1998) investigated VIL, storm environmental parameters, and combinations thereof, as related to reported hail size. The study's conclusion was that commonly used hail parameters lacked skill at predicting severe hail size. These parameters included VIL and VILD. Witt et al. (1998a) produced a sophisticated algorithm that vertically integrated reflectivity with weightings applied to both reflectivity and the reflectivity height, to produce estimates of maximum expected hail size and probabilities of hail and severe hail. The skill of the Witt et al. (1998a) algorithm in hail sizing was not directly evaluated due to limitations of the verification data, but the study concluded the algorithm was sufficient to at least categorically forecast severe hail using the algorithm output. Donavon and Jungbluth (2007) used the 50-dBZ echo-top height above the 0°C height to develop severe thunderstorm warning criteria with respect to hail size. This hail-detection methodology produced impressive skill scores, with a probability of detection of 0.9 and a false alarm ratio of 0.22. Implementations of these algorithms on the WSR-88D primarily only provide one estimate per storm, with no ability to determine surface hail fall spatial coverage.

The launch of the multi-radar, multi-sensor (MRMS; Smith et al. 2016) system within the National Weather Service (NWS) has made available many new products for severe-weather diagnosis and nowcasting. The MRMS method for merging single radar data is described in Lakshmanan et al. (2006) and the algorithm suite is described in Lakshmanan et al. (2007a). In order to generate MRMS reflectivity products, the operational system first takes raw level-II reflectivity data from individual radars within the WSR-88D network and applies a quality-control algorithm (Tang et al. 2014). The data are then remapped using the MapReduce algorithm (Lakshmanan and Humphrey 2014) and blended into larger tiled domains using an exponential distance (from radar) weighting over the contiguous United States (CONUS), and finally, those tiles are stitched together to create the final CONUS grid. The major benefit of using multiple radars is to overcome deficiencies in coverage due to radar beam geometry, cone of silence, volume coverage pattern (VCP) selection, and blockages due to terrain or other obstructions.

The operational MRMS system uses a 0.01° × 0.01° horizontal grid spacing with a vertical grid starting at 250-m MSL with 250-m vertical spacing and stretched to 1-km spacing up to the domain top of 20 km MSL. Model analysis information, such as the height of 0°C, can be easily combined with the three-dimensional MRMS reflectivity grid to produce additional products. The operational MRMS system uses a Rapid Refresh (RAP; Benjamin et al. 2016) analysis. The aforementioned algorithms and

techniques can be applied easily within the MRMS system because each is based upon reflectivity and simple environmental parameters. Applying the techniques to each grid point would provide a spatial extent of these algorithms' outputs, and potentially provide information on the spatial extent of surface hail fall and information on the spatial distribution of hail size.

c. Surface hail databases

Traditionally, *Storm Data* is used as verification of algorithms for the WSR-88D network. Collected by the NWS during the course of severe weather warning verification, *Storm Data* is a database of reports, including hail, thunderstorm wind measurements, estimates or damage, and tornadoes. (NWS 2016). Issues concerning *Storm Data* have been well explored, including:

- Reporting sufficiency (Hales and Kelly 1985, Hales 1993, Amburn and Wolf 1997, Trapp et al. 2006),
- Biases due to population and infrastructure, reporting sources, and report collection procedures (Hales 1993, Wyatt and Witt 1997, Davis and LaDue 2004, Dobur 2005, Hocker and Basara 2008, Allen and Tippett 2015),
- Hail-size accuracy (Schaefer et al. 2004, Jewell and Brimelow 2009, Blair et al. 2017), and
- Other, inexplicable inhomogeneities (Doswell et al. 2005).

Witt et al. (1998b) reported on the lack of null or nonsevere[1] reports within *Storm Data*, and problems using *Storm Data* as an algorithm verification database due to these missing data. Amburn and Wolf (1997) and Lenning et al. (1998) used population density and storm intensity thresholds *a priori* to presume nonseverenonsevere events. However, no evaluation on the accuracy of such thresholds has occurred.

After initial efforts to conduct verification and comparisons of single- and multi-radar hail detection and diagnosis techniques (Ortega et al. 2006), CIMMS and NSSL operated the Severe Hazards Analysis and Verification Experiment (SHAVE; Ortega et al. 2009) during the summer months from 2006–2015. The purpose of SHAVE was to make targeted phone calls to the public, usually within an hour of when the storm passed, in order to gather high-spatial-resolution reports. Reports collected included hail, wind damage, flash flooding, and precipitation type. For hail reports, information including the maximum and common size, the start and end time of the hail fall, and a measure of ground coverage was collected. SHAVE reports include 'no hail' verification for locations near the storm track and hail sizes not typically included within *Storm Data* (usually diameters <25 mm).

During the survey, SHAVE operators were trained not to ask leading questions regarding hail size (e.g., "did you get quarter-sized hail?") and to inquire further about hail sizes which were bluntly stated (e.g., "Golf ball is a little under 2 inches; does that sound about right?"). Should a respondent give a broad range of sizes, the operator would work to narrow the possibilities (e.g., "Is it more like a pea or more like a softball…more like a golf ball or more like a baseball?"). Measurements were accepted if volunteered and typically, operators would request measurements for hail-size estimates >50 mm. If questions regarding a reported hail size emerged, SHAVE operators would try to call as close to the location as possible (literally next door in urban/suburban settings to within a kilometer in rural areas) to try to verify the questionable report. If no other locations were available and third-party sources such as social media could not assist either, the report was flagged as questionable and it was excluded from the SHAVE database. In general, SHAVE operators tried to limit sizing error on the part of the respondents, but there is no way to completely eliminate all errors from observers supplying estimates of a quantity instead of actual measurements.

Reported hail-fall times within the SHAVE database are problematic as a general hail-fall time window was collected, and not specific times of maximum hail fall. Informal evaluations using MRMS data of a few cases revealed that many of the start and end times within the SHAVE data may be start and end times of the storm in general (i.e., when rain or lightning began and ended) and not when hail began and ended. However, since SHAVE generally only conducted operations on isolated storms, the general time of hail fall can be determined and techniques can be used to account for the time inaccuracies (Ortega et al. 2016).

[1] A change in the severe threshold from when Witt et al. (1998) was published has allowed for nonsevere hail to comprise ~25% of *Storm Data* for recent years.

This study evaluates the skill of different MRMS products, all of which are previously single-radar techniques adapted to the MRMS grid. In order to explore the impact of higher resolution reports like those within the SHAVE database, SHAVE and *Storm Data* reports will be compared and the evaluations will use both reporting sources, and the resulting evaluations will also be compared.

2. Data and methods

a. Hail reports

The nomenclature of the hail-size categories is as follows: hail with reported diameters <25.4 mm is defined as nonsevere, diameters from 25.4–50.8 mm as severe, and diameters ≥50.8 mm as significant-severe. A brief discussion on why this nomenclature is appropriate for use, especially for the operational community, is found within the reviews and replies of Blair et al. (2011; p. 24). Storms on which SHAVE operators were able to collect high-spatial-resolution reports were used in this study. SHAVE operations from 2006 through 2012 were considered and during that time SHAVE collected 39 951 hail reports. The SHAVE operations used in this study numbered 389 and yielded 21 184 hail reports, of which 9917 reports were of "no hail", 7133 were of nonsevere hail, 3648 were severe hail, and 486 were significant-severe hail. SHAVE operators collected two hail sizes during the verification phone calls: a maximum hail size and a common hail size. This study uses the reported maximum.

Storm Data reports also were collected for the 389 SHAVE operations. The total number of hail reports from *Storm Data* was 2814, of which 685 were nonsevere, 1860 were severe, and 269 were significant-severe. Some *Storm Data* reports had both a starting and ending location. These lines were converted to multiple points by remapping the line to the MRMS grid. For each grid point intersected by the line, a point was placed. No modifications of the reported hail size were made, thus each point had the same hail size. These intermediary points are included in the above counts. The cases used came from across the contiguous United States, with a majority located in the Central Plains region (Fig. 1).

Each SHAVE report was compared with the nearest neighboring SHAVE report to evaluate the consistency of the SHAVE database and to describe the spacing of the reports. The distance to the nearest report and the hail-diameter difference were recorded. A second matching used the 75th percentile of the distances between reports to define a new search area. The purpose of this second matching is to further evaluate hail-size differences between neighboring reports, especially when comparing SHAVE with neighboring *Storm Data* reports. The SHAVE report nearest in diameter within that search area was compared to the originating report. Because the size of the nearest neighboring report is completely arbitrary, as the locations of residences being called are not evenly spaced, this matching technique will further quantify the diameter differences within the SHAVE hail data. Each comparison also was stratified by hail-size category to further evaluate the quality of SHAVE observations.

The two hail databases were compared, with the *Storm Data* reports serving as the starting point for the comparisons because there are fewer *Storm Data* reports. Three different inter-comparisons were completed. The first was the simplest by comparing the *Storm Data* report to the nearest neighboring SHAVE report and recording the distance to that nearest neighbor and the hail-diameter difference. These comparisons were stratified by systematically removing smaller hail reports ("no hail" and diameters <12.7, 25.4 and 50.8 mm) in order to compare only similarly sized hail reports between the two databases. The final two comparisons used a varying search radius from 1–20 km. Two comparisons were made using these distance radii: the diameter of 1) the maximum SHAVE report, and 2) the SHAVE report nearest in size to that of the originating *Storm Data* report.

b. *MRMS data*

The MRMS configuration used within this study differs from the operational MRMS system described above (since the data were processed before MRMS was operational and the settings for the operational system were not yet determined). However, the differences should not make a major impact since beamwidths, even at modest distances <100 km from the radar, are generally as large or larger than the vertical grid spacing employed. Also, high-reflectivity echoes associated with hail cores are most likely preserved regardless of the quality control scheme employed.

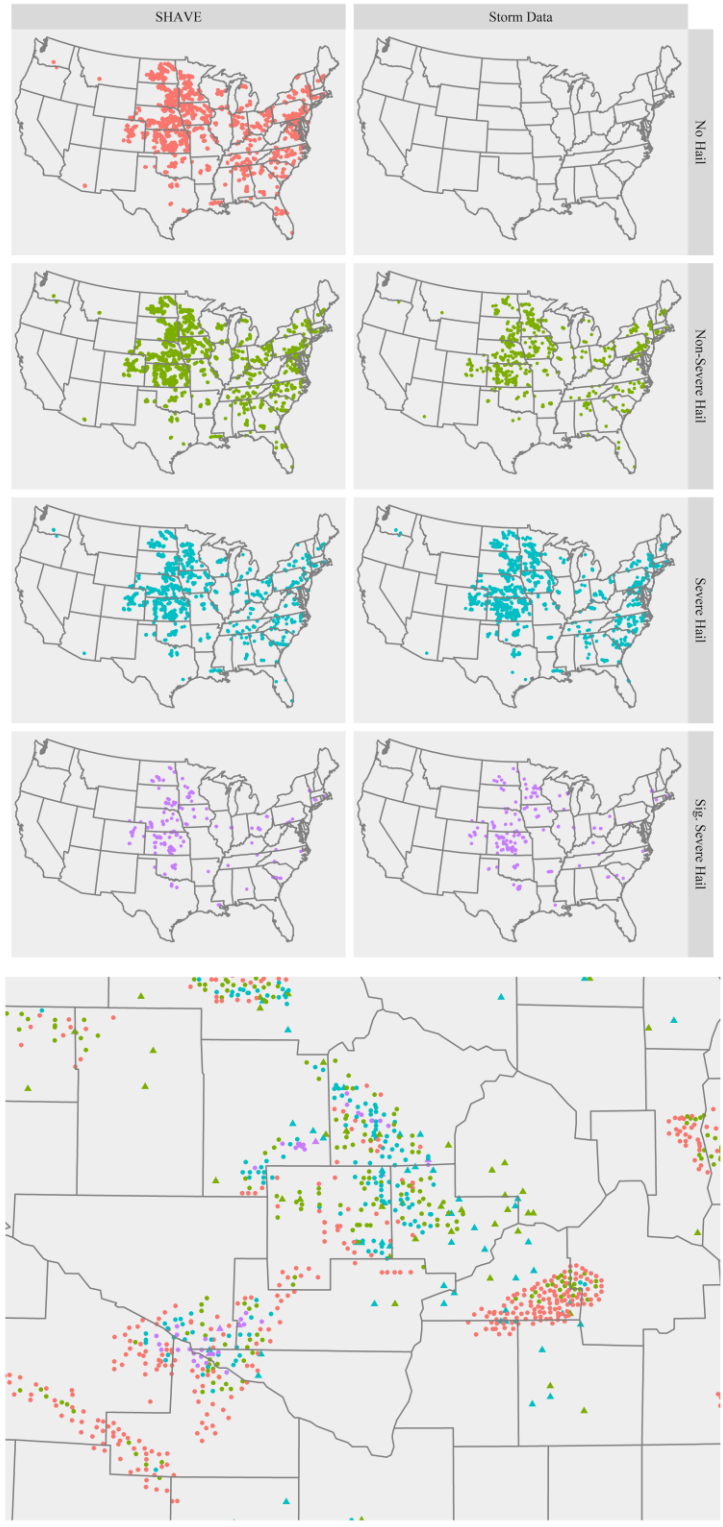


Figure 1: SHAVE (left) and *Storm Data* (right) hail reports used within the study. No hail (top row; red), nonsevere hail (2nd row; green), severe hail (3rd row; blue), and significant-severe hail (bottom row; purple) have been separated for clarity. A zoom of the area near Minneapolis, MN, is included below to show the differences between SHAVE (circles) and *Storm Data* (triangles) in more detail. *Click image to enlarge.*

WSR-88D level-II data were processed using the Warning Decision Support System—Integrated Information (WDSSII; Lakshmanan et al. 2007a) software suite. Reflectivity data were quality controlled using the algorithm described in Lakshmanan et al. (2007b) and then merged using an inverse square distance (from radar) weighting (Lakshmanan et al. 2006) onto a 0.01° × 0.01° latitude-longitude grid with uniform 1 km vertical spacing from 1 km to 20 km MSL. Data from individual radars were only included if the MRMS grid point was within 250 km of the radar. The merged reflectivity was produced every minute. The 20-km Rapid Update Cycle (Benjamin et al. 2004) and 20-km RAP hourly analyses were used to calculate several thermodynamic and kinematic parameters for use within the merger software and in the post-analyses.

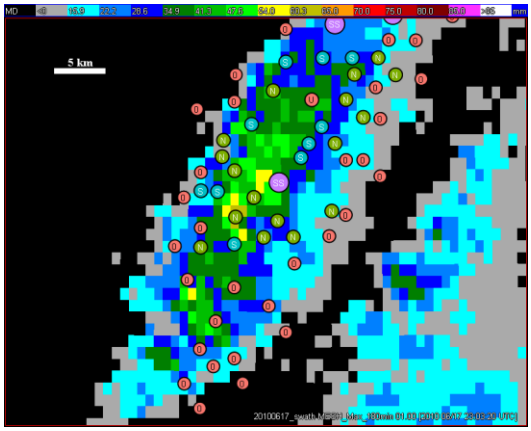


Figure 2: SHAVE reports and maximal MESH for a storm on 17 June 2010 in south-central Minnesota. No hail (red '0' icons), nonsevere hail (green 'N' icons), severe hail (blue 'S' icons), and significant-severe hail (purple 'SS' icons) reports are all shown. The accumulation time for the maximal MESH was 3 h. *Click image to enlarge.*

The domains selected for each event were exclusive to the SHAVE report swath being investigated (Fig. 2). The time period of the merger was determined by reviewing SHAVE reports and radar data. At minimum, at least 30 min was added before the start of the SHAVE report swath and after the last SHAVE report, as determined by both the reported times within the SHAVE data and a manual analysis of when the reflectivity at lowest altitude was passing over the reports. *Storm Data* reports were selected from these spatial and temporal bounds.

c. *Algorithm outputs and reflectivity profiles*

Nine different MRMS outputs were chosen to evaluate the potential of those products to diagnose surface hail fall: maximum expected size of hail (MESH) and probability of severe hail (POSH; Witt et al. 1998a), the 50- and 60-dB*Z* reflectivity heights above the environmental 0°C height (hereafter H50_0C and H60_0C, respectively; Donavon and Jungbluth 2007), VIL and VILD (see below), vertically integrated ice (VII; Mosier et al. 2011; a vertical integration of a reflectivity-ice mass relationship between the –10°C and –40°C heights), composite reflectivity (COMP), and reflectivity at the lowest altitude (RALA). VILD is calculated using VIL and the 18-dB*Z* echo top for the same grid point and COMP is the maximum reflectivity in the column at a given grid point.

The hail reports were paired with the maximal value of each MRMS product during the temporal bounds for each case (Fig. 2). To evaluate the skill of the products in diagnosing surface hail fall, various thresholds for each product were applied and confusion matrices were generated for each product and threshold, and for three hail-size categories: any hail, severe hail, and significant-severe hail. The confusion matrix was filled out such that if the MRMS product exceeded the specified threshold then a prediction for the hail-size category under consideration was made; if the product value was below the threshold, a negative prediction was made. If the hail report was at least of the size category under consideration, a positive observation would be made; if not, a negative observation was made. The hail-size categories follow the classification convention discussed earlier, except without a maximum size cap (e.g., severe hail would be diameters of at least 25.4 mm and not capped). *Storm Data* reports were only used to calculate the severe and significant-severe, since 'no hail' reports are not available. While several statistics can be calculated from the confusion matrices, critical success index (CSI; Schaefer 1990) and Heidke skill score (HSS; Doswell et al. 1990) will be used to evaluate the products. Finally, the Spearman rank correlations (Wilks 2006) were found for each pairing of the nine MRMS products.

For further exploration of the performance of the MRMS products, vertical profiles of reflectivity were created for each SHAVE report. In total three vertical profiles were created for

each report: one was at the time of the maximum MESH, another at the time of the maximum VIL, and the final profile at the time of the maximum RALA. The reflectivity was interpolated within the merger software to isothermal heights, for temperatures from –50°C to 20°C by 5°C increments. Vertical profiles were paired with environmental kinematic and thermodynamic parameters to explore the environment's impact on the profiles.

3. Results

Some of the following boxplots are presented stratified by color. This color corresponds to the originating report's hail-size classification.

a. Report comparisons

The median spacing between SHAVE reports is 1.72 km (Fig. 3), with a median diameter difference of 0 mm (Fig. 4). Separating out the distributions by hail-size category, the distance distributions look similar across all hail-size categories (Fig. 5), while larger hail-size categories typically have larger hail-diameter differences (Fig. 6). Using the 75th percentile (2.39 km) of the distance between SHAVE reports and searching for the neighboring report within that radius with the hail diameter nearest the originating report, the spreads of the hail-diameter difference distributions are reduced and the distributions shift towards no difference (Fig. 7).

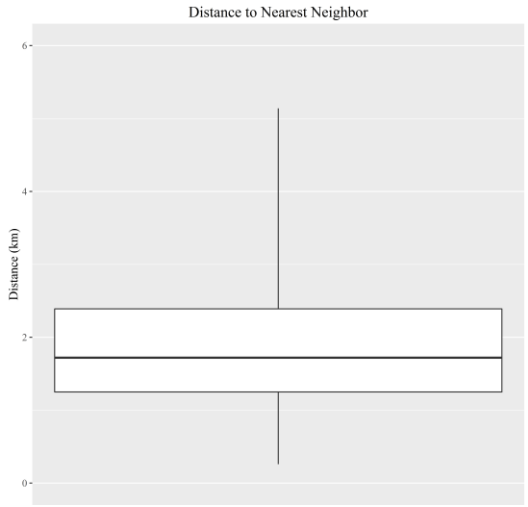


Figure 3: Boxplot of the distance between nearest neighboring SHAVE reports for all SHAVE reports. The whiskers are the 95th percentile, the box the interquartile, and the heavy black line is the median. *Click image to enlarge*.

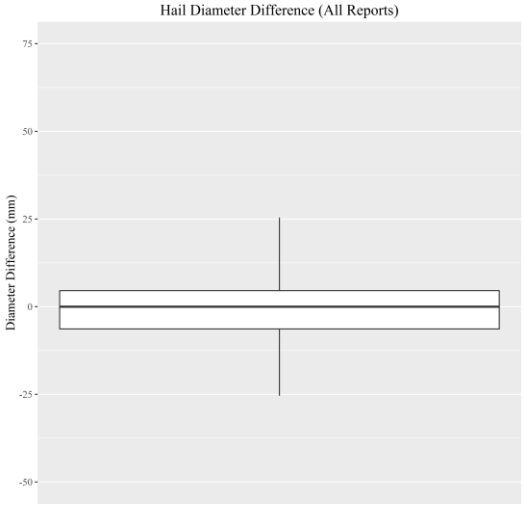


Figure 4: As in Fig. 3, but for the hailstone-diameter difference between nearest neighboring SHAVE reports, for all SHAVE reports. *Click image to enlarge.*

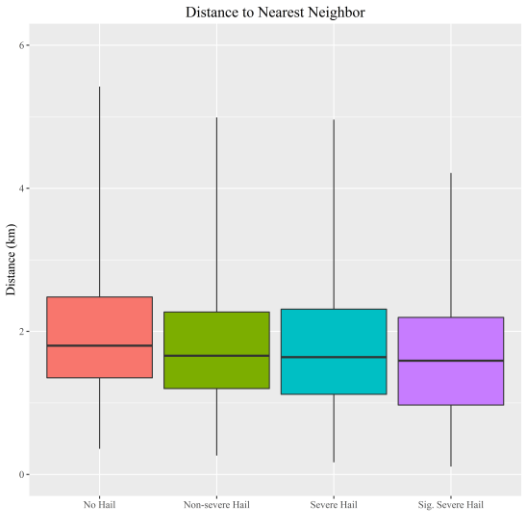


Figure 5: Boxplots of the distance to the nearest neighboring SHAVE report for SHAVE reports of differing hail-size category. The boxplots are as in Fig. 3. The colors of each boxplot are for the originating report. *Click image to enlarge.*

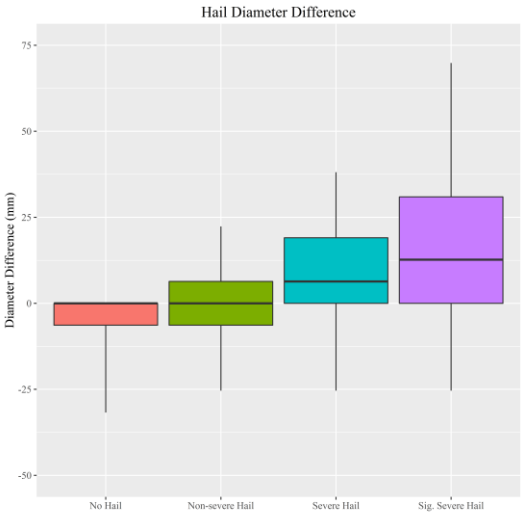


Figure 6: Boxplots of the hailstone diameter difference for the nearest neighboring SHAVE reports for SHAVE reports of different hail-size categories. The boxplots are as in Fig. 3. The colors of each boxplot are for the originating report. *Click image to enlarge.*

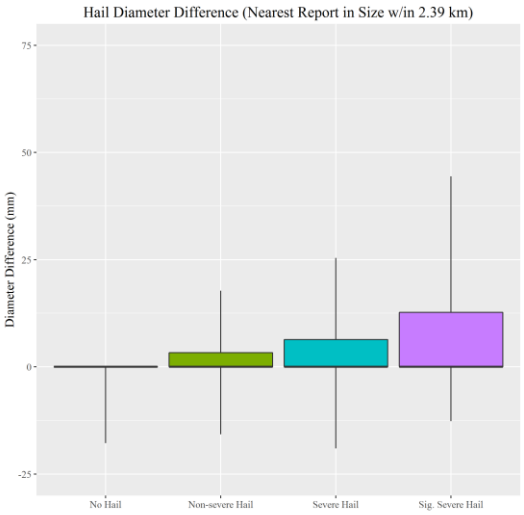


Figure 7: Boxplots of hailstone diameter difference for the nearest SHAVE report in size within 2.39 km for SHAVE reports of differing hail-size categories. The boxplots are as in Fig. 3. The colors of each boxplot are for the originating report. *Click image to enlarge.*

The *Storm Data*-SHAVE comparisons revealed a median spacing of 1.69 km between *Storm Data* and SHAVE reports, with a median hail-diameter difference of 6.35 mm. As smaller SHAVE hail reports are removed from consideration, distances from *Storm Data* to the nearest SHAVE report increase (Fig. 8) and the spread in the hail-diameter difference distributions shrink and generally shift towards values closer to 0 mm (Fig. 9).

Comparisons of *Storm Data* reports to SHAVE reports within an increasing radius demonstrate a stabilization of the diameter difference distributions with a radius of 10 km (Fig. 10). For the SHAVE report nearest in diameter, all diameter difference distributions for all report categories quickly collapse towards 0 mm as the search radius increased, meaning no difference in diameter between SHAVE and *Storm Data* database. For comparisons to the maximum reported SHAVE diameter within the radius, the distributions for smaller *Storm Data* hail diameters quickly drop below 0 mm, meaning the matching SHAVE report is larger. Empirical cumulative distribution functions (ECDF) also show this behavior in the comparison for the maximum SHAVE diameter within the searched area (Fig. 11). As the search radius is increased, the ECDF of the matched SHAVE reports moves towards being similar to the *Storm Data* ECDF and then shows larger proportions of larger hail diameters for search radii ≥3 km.

In summary, the SHAVE-to-SHAVE comparisons reveal a median spacing of 1.72 km, 75% of the reports were within 2.39 km of the nearest neighbor, and small hail-diameter differences between neighboring reports, regardless of hail diameter. The *Storm Data*-to-SHAVE comparisons reveal shifting distributions of diameter differences depending on the minimum size of SHAVE reports used and the search radius used in the matching. Removing smaller diameters and matching only larger sized reports, the matching revealed that as smaller sizes were removed, the distances between reports increased as the diameter differences decreased.

Increasing the search radius around each *Storm Data* report resulted in hail-diameter differences decreasing for the nearest SHAVE report in size. *Storm Data* reports went from generally having a larger diameter to generally having a smaller diameter when matched to the maximum SHAVE hail diameter within the search area, as the search radius increased.

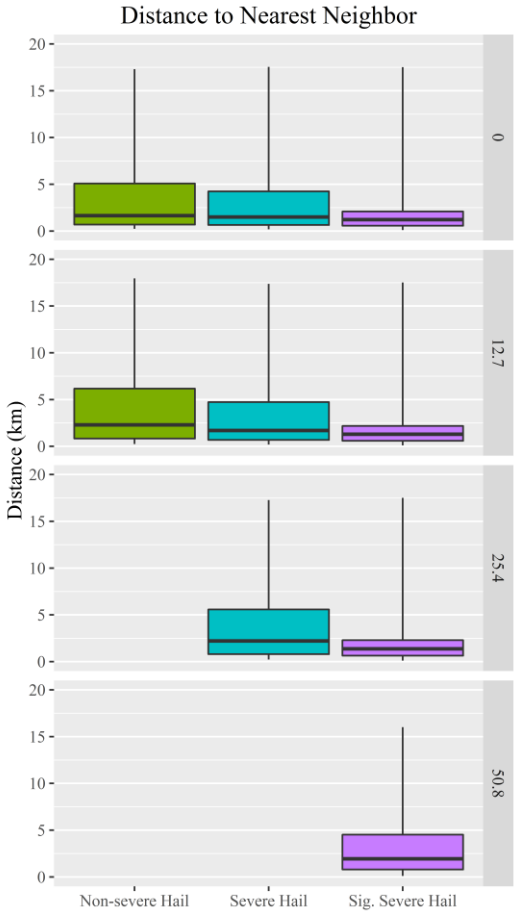


Figure 8: Boxplots of distance to nearest SHAVE report for *Storm Data* reports of differing hail-size categories. Each row excluded SHAVE reports below the threshold (in mm) labeled on the right of each row. The boxplots are as in Fig. 3. *Click image to enlarge.*

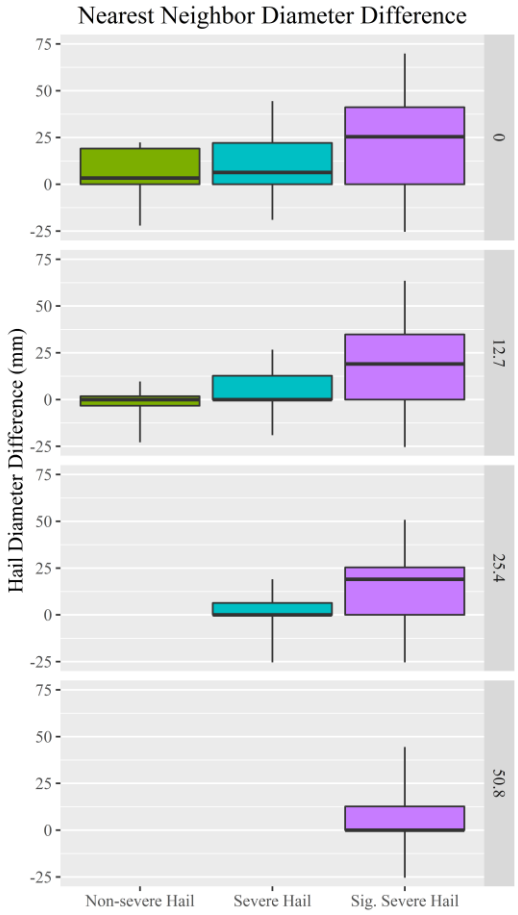


Figure 9: Boxplots of hailstone diameter difference of the nearest SHAVE report for *Storm Data* hail reports of differing hail size. The rows are as in Fig. 8 and the boxplots are as in Fig. 3. *Click image to enlarge.*

b. MRMS product evaluation

Peaks in overall skill were similar for most MRMS grids using SHAVE reports as the validation source (Fig. 12). HSS values are different and generally far lower when using *Storm Data* as the verification source (Fig. 12). Using SHAVE, for any sized hail, RALA = 55 dB*Z* provides the best threshold with an HSS of 0.42. Using SHAVE, for severe sized hail, MESH = 30 mm provides the best threshold with an HSS of 0.38. None of the grids provided a good value for significant-severe hail, with all HSS values <0.20. For further clarification of the HSS values, the receiver operating characteristic (ROC) curves were constructed for the three MESH hail-size categories, using SHAVE as the verification set (Fig. 13). For the best MESH HSS score for discriminating severe hail (MESH = 30 mm), the probability of detection is expected to be 0.6 with a probability of false detection of 0.18.

Using SHAVE reports to calculate CSI (Fig. 14), different thresholds emerge as more skillful than others for the different hail categories, but in general, the relative performance of each product is the same as revealed using HSS. For example, RALA was still top performer for any hail size, with a peak CSI of 0.58 at a threshold of 51 dB*Z*. For *Storm Data*, however, all CSI scores peak at MRMS product thresholds of 0, except for a few products for the significant-severe hail-size category (Fig. 15).

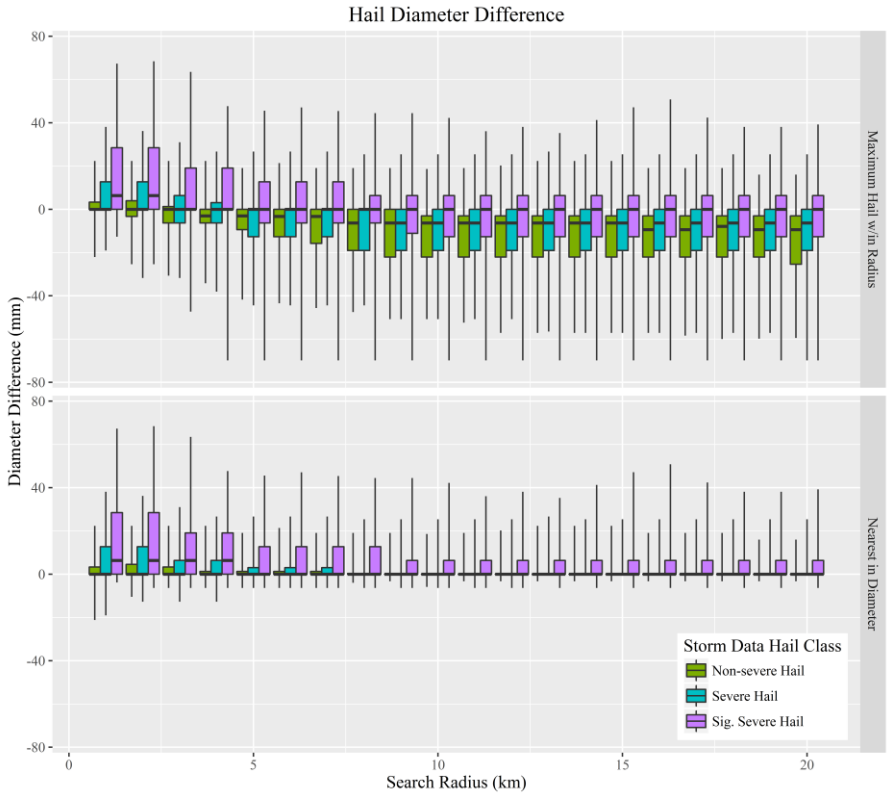


Figure 10: Boxplots for the maximum sized SHAVE hail within the radius (top) and the SHAVE report nearest in diameter within the radius (bottom) hail-diameter differences for different search radii around *Storm Data* reports. *Click image to enlarge.*

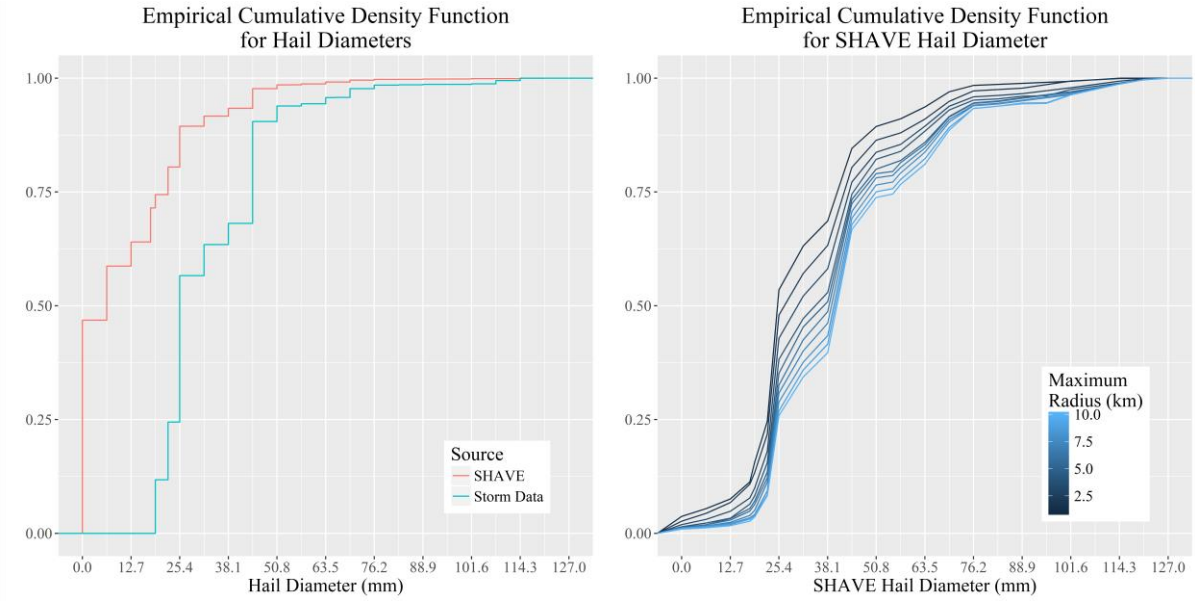


Figure 11: ECDFs for: (left) SHAVE and *Storm Data* reports used within this study; (right) resulting SHAVE ECDFs from different search radii around each *Storm Data* report and matching to the largest SHAVE report within the searched area. *Click image to enlarge.*

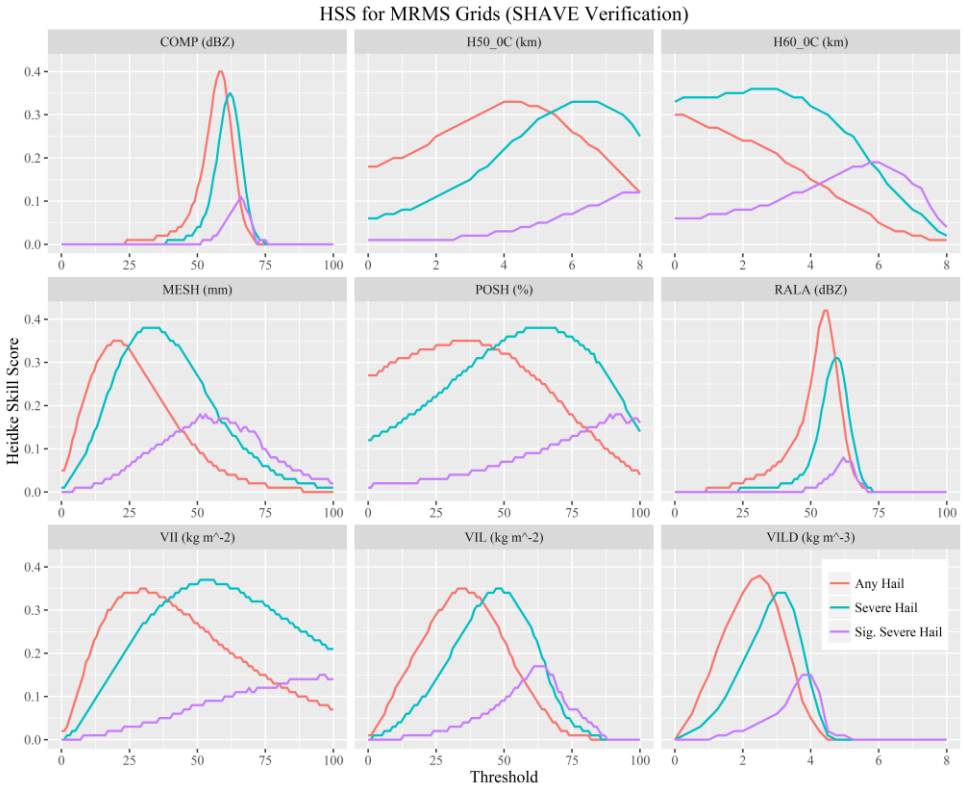


Figure 12: Heidke skill scores for the MRMS grids with different hail-size thresholds using SHAVE as the verification source. *Click image to enlarge.*

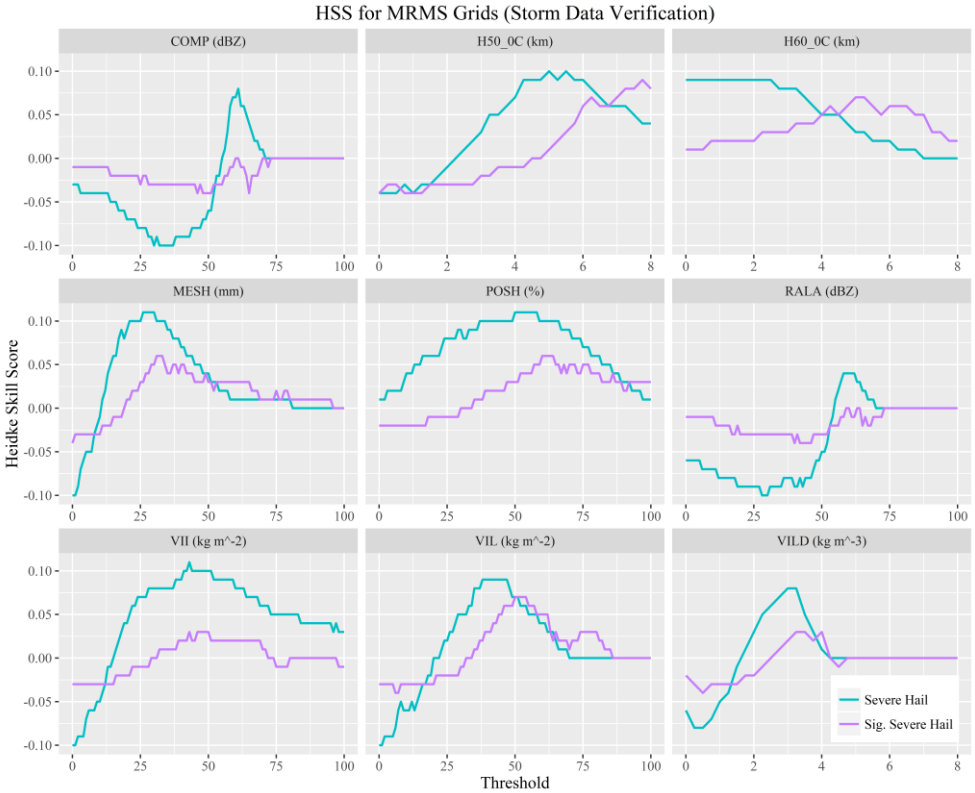


Figure 13: As in Fig. 12, except using *Storm Data* as the verification source. *Click image to enlarge.*

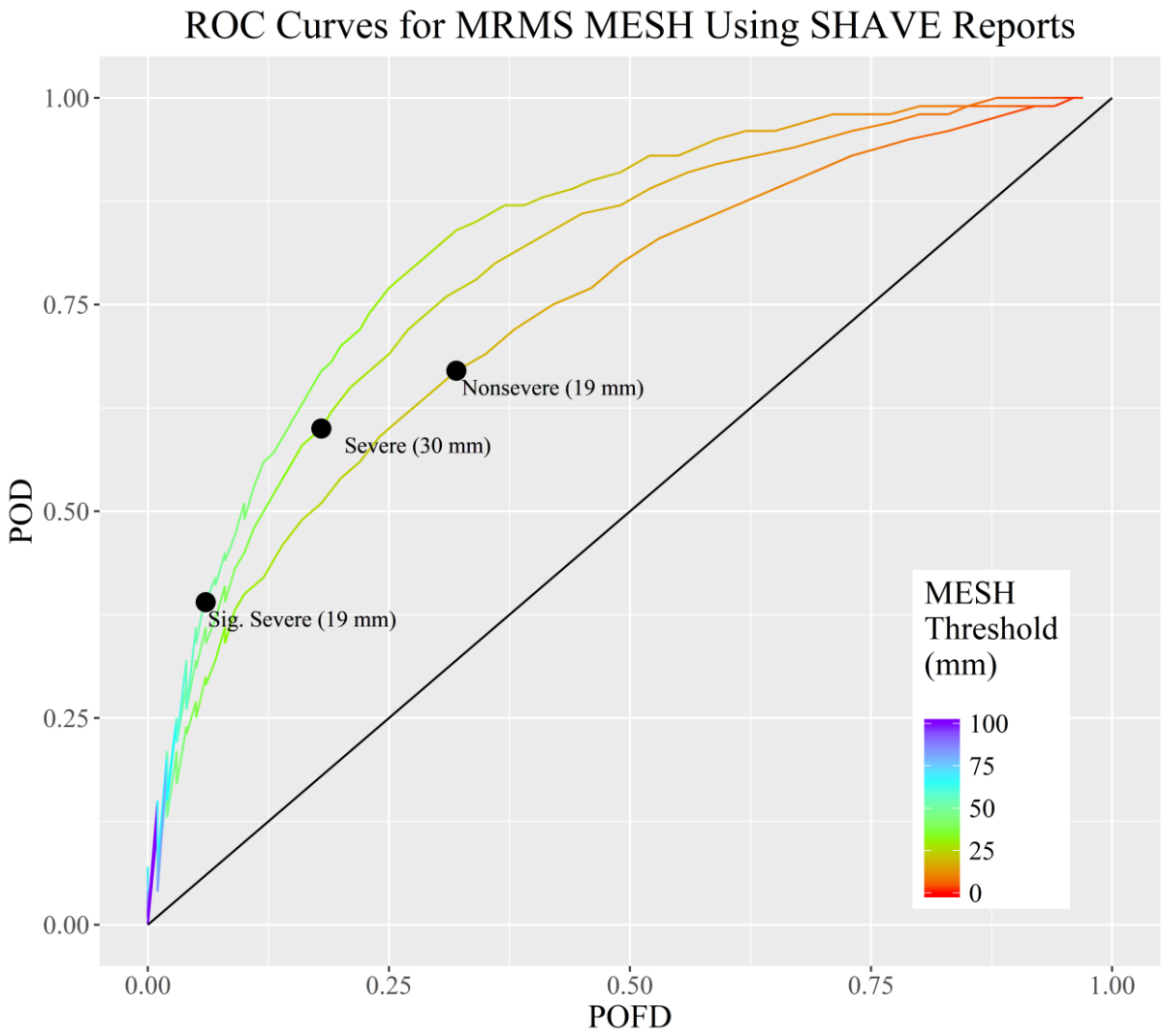


Figure 14: ROC curves for the three hail-size categories for MESH using SHAVE verification. The probability of false detection is on the abscissa and the probability of detection is on the ordinate. The curves are colored by the MESH threshold. The peak HSS values for the three hail-size categories are annotated on the respective curve.

The similar skill scores between products could be explained by the correlations between the products. The correlations between each product are generally above 0.5 (Fig. 17). All of the vertically integrated products have correlations at or above 0.8. Bootstrapped 95% confidence intervals (not shown) of HSS show overlap of most products, suggesting no statistical significance in the differences of performance between each product.

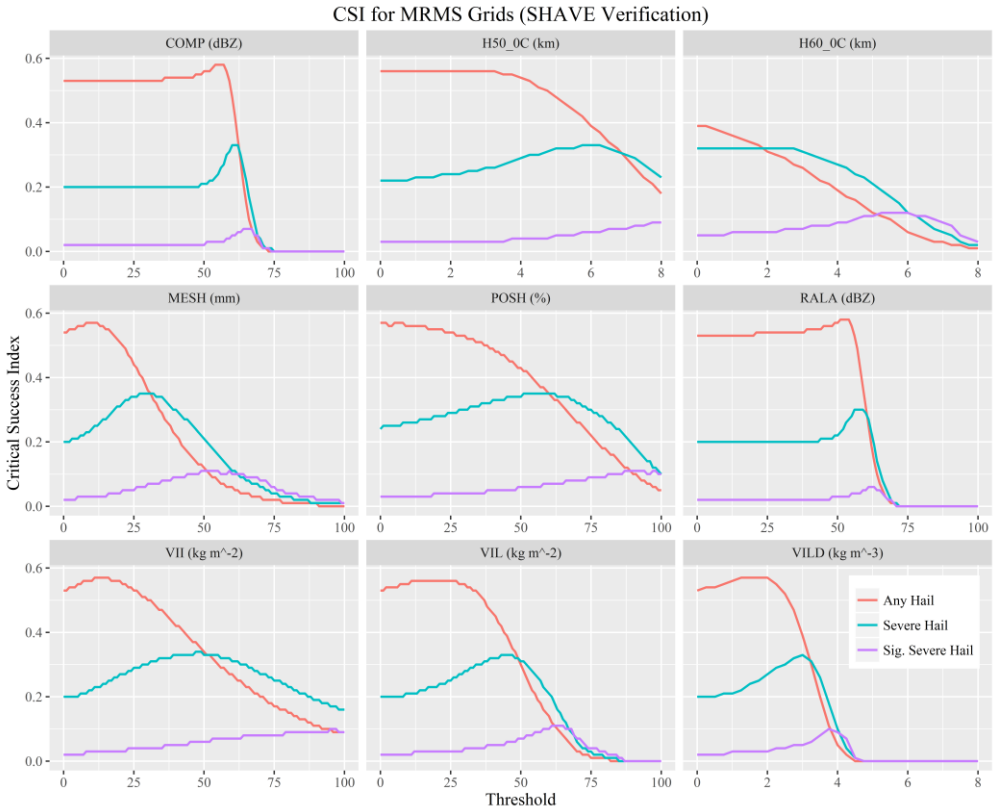


Figure 15: As in Fig. 12, except for critical success index. *Click image to enlarge.*

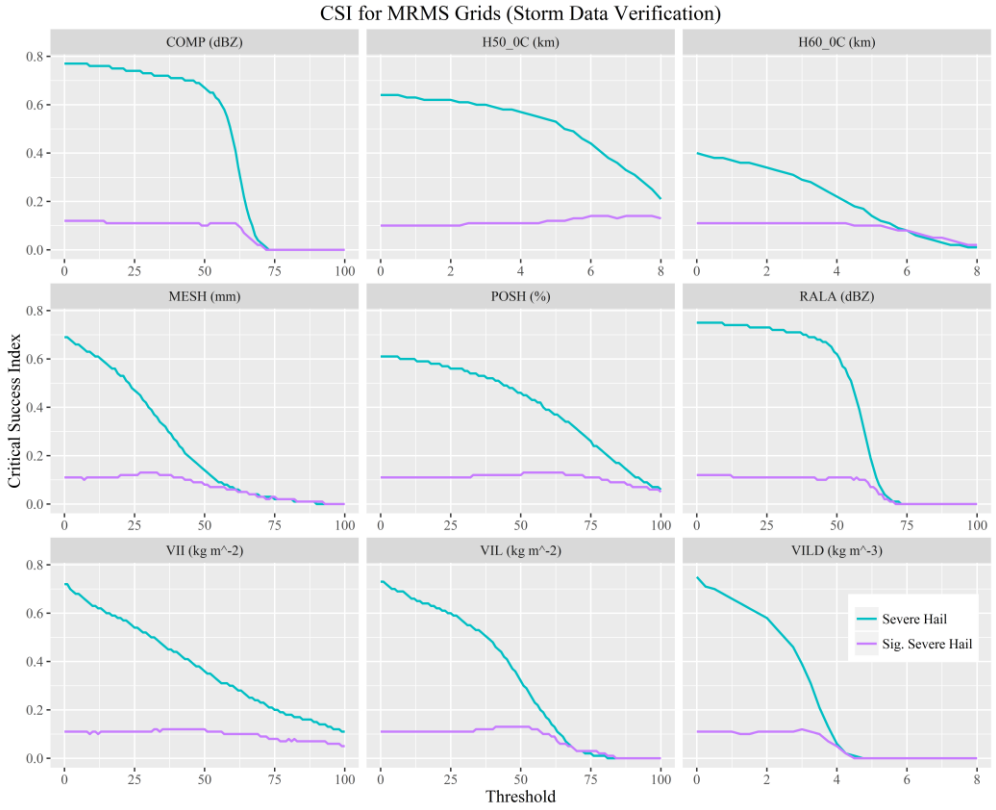


Figure 16: As in Fig. 13, except for critical success index. *Click image to enlarge.*

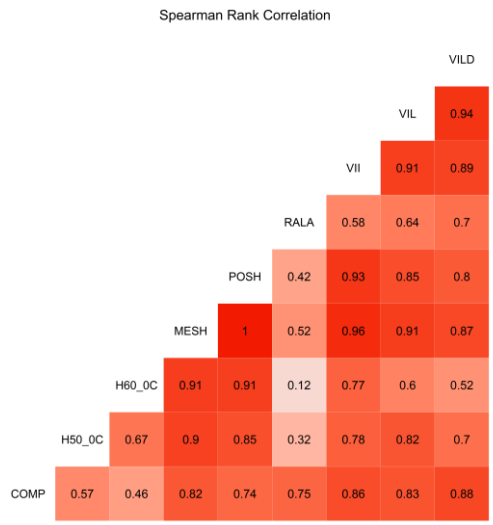


Figure 17: Spearman rank correlation coefficients of the MRMS products. Darker shades of red specify higher correlations coefficient values. *Click image to enlarge.*

c. Vertical profiles of reflectivity

The vertical profiles of reflectivity show some obvious patterns given the results of the MRMS product evaluation (Fig. 18). In general, larger hail-size categories have distributions shifted towards larger reflectivity. Further, at lower temperatures (higher altitudes), the separation of the reflectivity distributions for the hail classes becomes more pronounced than at the higher temperatures (lower altitudes). Generally, smaller hail-size categories typically have broadening of the distribution with increasingly lower temperatures (higher altitudes), while the significant-severe hail distribution narrows. Overall, adjacent hail-size categories show considerable overlap and there is considerable overlap of the distributions for similar hail-size categories regardless of how the profile was selected (e.g., time of maximum MESH compared to the time of maximum VIL).

The vertical profiles of reflectivity do not vary by different environments (Fig. 19). While only instability and shear are presented, the profiles were stratified by several parameters and indices over numerous layers of the column, including surface mixing ratio, storm relative helicity, melting level height, and environmental relative humidity. All showed similar overlap ofthe distributions, and no distinct pattern with respect to the differing environmental parameters.

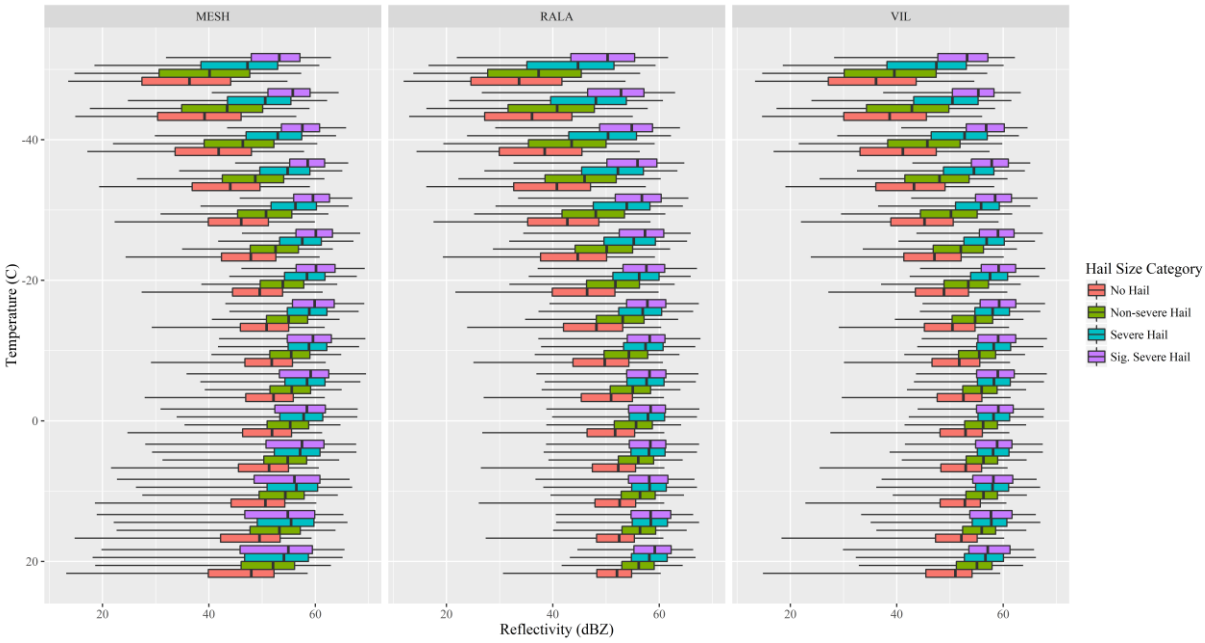


Figure 18: Boxplots of MRMS reflectivity at isothermal levels from 25°C to –50°C by 5°C increments for SHAVE reports of the different hail-size categories. The values were selected at the time of the maximum MRMS product for each report: MESH (left), RALA (middle), and VIL (right). *Click image to enlarge.*

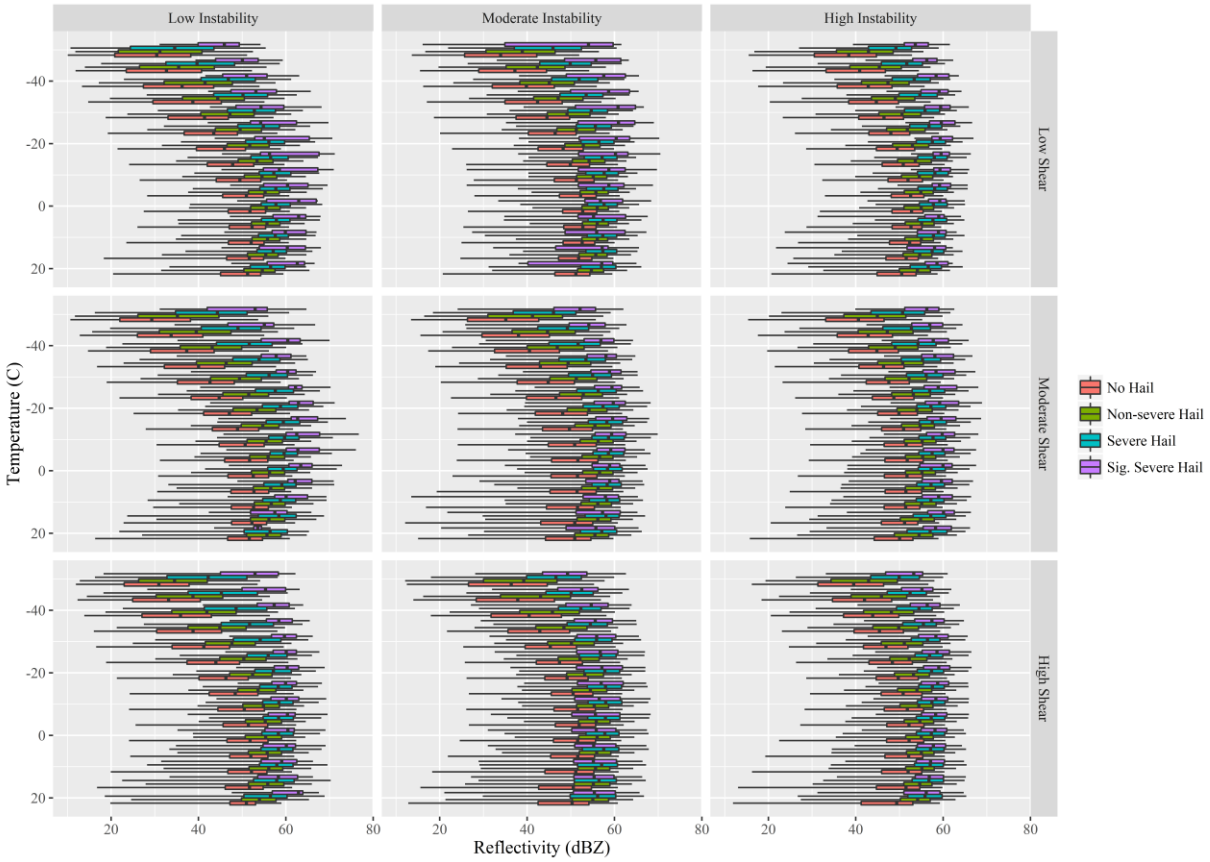


Figure 19: Vertical profiles, presented as in Fig. 16, of reflectivity for SHAVE reports of the different hail-size categories for different near-storm environments. Most-unstable CAPE and the 0–6 km MSL shear-vector magnitude were used for instability and shear, respectively. The thresholds for low, moderate, and high instability (shear) are: 1668 and 2618 J kg^{-1} (17 and 25 m s^{-1}). *Click image to enlarge.*

4. Discussion

a. Hail-report databases

The SHAVE intra-comparisons demonstrate that one of the goals for the SHAVE project, to sample at fairly high resolution, was achieved with a median spacing of 1.72 km, which is just slightly higher than the suggested spacing of 1.45 km of Changnon (1968). That, in turn, is only slightly larger than the 1.6-km typical spacing of the gridded road networks of the U. S. Central Plains. This spacing is also similar to the 2.15-km spacing of SHAVE reports in Ortega et al. (2016), which used a different subset of SHAVE data than the present study. The spacing is also similar across all hail-size categories, suggesting that no areas of the surface hail fall were sampled more than others. The slightly tighter spacing of reports for larger hail sizes was from a tendency to confirm larger hail sizes during SHAVE operations. For instance, if a swath up to a certain point had only yielded a maximum hail size of 44 mm, but a report of 70 mm was found, SHAVE operators would focus around the 70-mm report to try to confirm. NWSChat[2] and social media networks also were leveraged to confirm reports, which also helped to limit tightly spaced reports.

The SHAVE hail-diameter difference distribution (Fig. 4) is similar to the hail-diameter differences found within the SHAVE subset used in Ortega et al. (2016), which had an interquartile range of ±10 mm. The breakdown of the differences by hail-size category also reveal small differences between neighboring SHAVE reports, suggesting that SHAVE captured the general gradient of hail diameters for surface hail fall. The map of SHAVE reports

[2] https://nwschat.weather.gov/

(Fig. 2) helps to explain the differences seen in the distributions between immediate and general neighboring reports (Figs. 6 and 7). While the exact nearest neighbor to a report might be of a completely different hail-size category, a more similarly sized report may still be within the immediate area.

The *Storm Data*-SHAVE comparisons summarized in Figs. 8 and 9 suggest that *Storm Data* reports are not located precisely with the reported diameter in the SHAVE database. As smaller SHAVE reports are removed from consideration, the distance from the *Storm Data* report to the nearest SHAVE report increases. This is combined with results that show as the smaller SHAVE reports are removed, the hail-diameter difference distributions narrow and shift towards zero (Fig. 9). Using a 5-km radius (approximately the 75th percentile in nearly all distributions in Fig. 8), the distributions of hail-diameter difference for *Storm Data* reports (Fig. 10, nearest in diameter) look very similar to the SHAVE intra-comparison distributions (Fig. 7). Within the same 5-km radius, over half of *Storm Data* reports are smaller than the maximum SHAVE report in the area (Fig. 10).

Thus, the question of which database is more accurate depends on its use, *Storm Data* cannot be used as an exact point report of the maximum size for a given location. This is readily seen in the MRMS product analyses with drastically different skill scores than when using the SHAVE database (Figs. 12–13, 15–16). If *Storm Data* is to be a proxy for hail fall within some defined area, it may be accurate to state hail of that size fell within the defined area. At minimum, this area has a radius of 5 km, but the radius could be increased to 10 km in order to minimize sizing errors (Fig. 10). A statement that *Storm Data* is defining the maximum hail diameter for an area (Fig. 10) likely is inaccurate, consistent with the findings of Blair et al. (2014). For SHAVE reports, it obviously cannot be known whether SHAVE sampled the maximum hail without another independent database. However, the diameter differences between neighboring reports suggest SHAVE is capturing the general gradient of the surface hail fall size. The distances between SHAVE reports are also similar to the MRMS grid spacing, making the reports well-suited for evaluating the products in a grid-point-by-grid-point manner.

b. MRMS Products

The simple decreasing trend of CSI for increasing MRMS product thresholds while using *Storm Data* as the verification source (Fig. 16) serves as a caution in picking not only the verification source, but also the statistics in evaluating products. The NWS definition of severe hail changed in 2010, from 19.05 to 25.4 mm. Thus, the percentage of reports in *Storm Data* used in this study that were <25.4 mm went from being ~30% prior to 2010 to being ~18% during 2010–2012, which can limit the completeness of evaluating the nonsevere/severe thresholds. This follows Amburn and Wolf's (1997) statement that *Storm Data* hail reports are primarily collected to verify severe-weather warnings. This low percentage of nonsevere *Storm Data* reports is compared to the SHAVE reports that are predominantly (80%) nonsevere. Regarding the skill statistics, CSI is limited in what it tells of the skill of the product, because it does not account for correct nulls. When comparing CSI and HSS values (using either SHAVE or *Storm Data* verification), in many areas where HSS implies no or little skill, CSI implies modest skill.

The similarity of skill scores across all MRMS products is not surprising given the large correlations between the products (Fig. 17). Further, each of the MRMS products tested here is based upon merged reflectivity, and sometimes environmental parameters, but no other radar variable. The large correlations and moderate skill scores are explained well by the overlap of the distributions of the vertical profiles of reflectivity (Fig. 18), which show large overlaps of different hail-size categories for the same product, and across different products for the same hail-size category.

The large correlations are operationally problematic. The operational MRMS system generates a large number of products (including all of the products evaluated here), while some algorithms use a combination of these products to generate new analyses (Smith et al. 2016). Further work is needed to explore whether these correlations exist on a time-step-by-time-step basis, and are not just an artifact of using the maximum for a given time period. Should these large correlations be present, a reevaluation of the suite of products produced by the MRMS system would be needed.

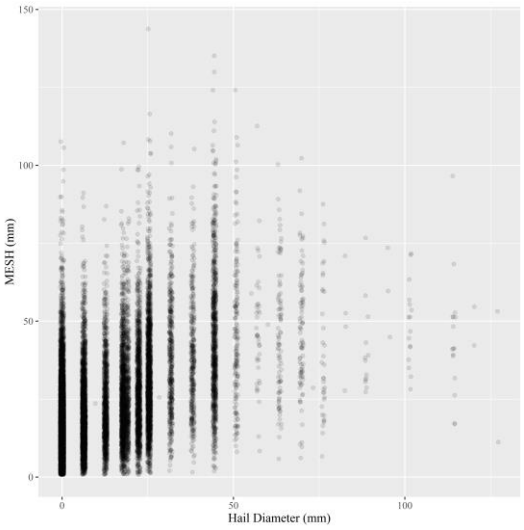


Figure 20: Scatterplot of reported SHAVE hail diameter compared to MRMS MESH values. *Click image to enlarge.*

There is less overlap of the reflectivity distributions for each hail-size category above the melting level than below for all the product-based profiles investigated (Fig. 18). However, there are not large differences between the distributions of the different product-based profiles above the melting level. This suggests that deep columns of high reflectivity must be present for larger hail. Interestingly, for the MESH-based profiles below the melting level the reflectivity distributions are broader and shifted towards lower values than the RALA- or VIL-based profiles. This means that for many reports, there was low to moderate reflectivity at the lower altitudes at the time of the MESH maximum; meanwhile, the high reflectivities would be present throughout the upper altitudes. This suggests that the MESH product is the most removed from direct observation of surface hail fall. Yet, MESH was the best product (per HSS) in discriminating severe hail areas. However, specific MESH values do not align well with specific hail sizes, with essentially most hail sizes having MESH values from 0 to 75 mm (Fig. 20). Further work is needed to evaluate the MESH, and potentially other MRMS products, to determine whether changing the integration's valid reflectivity and temperature ranges impact the estimates and the resulting accuracy and skill of the product.

The broader distributions aloft were not the result of sample-size problems. In fact, even at the –50°C level, nearly 90% of the 'no hail' and 480 of the 486 significant-severe hail observations still have valid reflectivity values. The broadening of the distributions aloft is most likely from a variety of spatial areas of high reflectivity aloft. Broader areas of high reflectivity could spread over small hail-size observations at the surface or the high reflectivity area aloft might be constrained to a small area immediately above the largest hail.

The large overlaps of the distributions at warmer temperatures are most likely the result of two issues. First, for the MESH-based profiles, SHAVE did make a concerted effort to collect reports where surface hail fall began, thus while the MESH was present, precipitation may have only just begun at the surface, leading to the possibility of smaller reflectivity values at lower altitudes. For the VIL- and RALA-based profiles, the overlaps at lower altitudes are most likely the result of saturation of the merger, due to the blending of the individual radar data, in turn causing moderating reflectivity values.

The magnitudes of reflectivity within the reflectivity distributions for the different hail categories are well below those found in single-radar data (Ortega et al. 2016). This is not surprising given that the MRMS system blends the single radar observations, which moderates values. The addition of polarimetric variables (e.g., Ryzhkov et al. 2013; Ortega et al. 2016) within the MRMS system and the addition of storm-rotation variables (e.g., Blair et al. 2017), could assist in discriminating different hail-size categories. Further work is needed on these topics.

The large overlap of the distributions of vertical reflectivity profiles, which limits the discrimination power of MRMS products for hail sizing, for different environments is consistent with Mustered and Ortega (2012) and Edwards and Thompson (1998). One potential reason for the lack of reflectivity profile discrimination using different environmental parameters might be from the coarseness (both spatially and temporally) of the environmental analysis. The accuracy or representativeness of the analysis and the actual impact of an environment on a storm's microphysical processes, with respect to hail production, also could limit the ability to apply environmental parameters to MRMS hail products.

Another potential problem is simply selecting combinations of two parameters in which stratifying the reflectivity profiles was too simplistic. The small sample size of severe and significant-severe hail is another limitation for this analysis. Johnson and Sugden (2014) developed a "Large Hail Parameter" that showed good discrimination between environments capable of producing significant-severe hail and those environments that would only produce marginally severe hail. Further work is needed on combining the environment and MRMS radar data, including investigating the accuracy of MRMS products on a coarser grid and selective application of algorithms on areas correlated with large hail (e.g., deep columns of high reflectivity values); along with using more sophisticated near-storm environmental indices to improve MRMS hail discrimination.

5. Summary

SHAVE and *Storm Data* hail reports were compared. SHAVE reports had a median spacing of 1.74 km and hail-diameter differences between neighboring reports generally <25 mm. *Storm Data* reports had varying differences and distances to the nearest SHAVE report, depending on whether or not smaller SHAVE reports were considered for matching. This suggests *Storm Data* reports are not precisely placed with respect to the reported hail size. In general, the SHAVE-*Storm Data* comparisons suggest the reported *Storm Data* hail size at least describes that hail of the reported size fell within 5 km of the reported location. *Storm Data* location imprecision limits the applicability of its reports to precise grid point-level evaluations, as were conducted here.

An evaluation of MRMS hail-product maximal swaths has also been presented. The products use the three-dimensional MRMS reflectivity grid, combined with model analyses. The MRMS products were evaluated using the SHAVE database. Overall, the RALA product at a threshold of 55 dBZ was the best for discriminating hail where any size fell (HSS = 0.42) and the MESH product at a threshold of 30 mm was the best for discriminating where severe hail fell (HSS = 0.38). No product had HSS >0.2, when considering significant-severe hail.

Vertical profiles of reflectivity were generated for each hail report at the time of the maximum of MESH, RALA or VIL. In general, the profiles confirm that taller columns of higher reflectivities are associated with larger hail sizes. Stratifying the profiles by near-storm environmental parameters did not increase discrimination of different hail-size category. The vertical profiles' lack of stratification and the considerable overlap between different hail-size categories suggests improvements to hail-size identification by a single product (e.g., MESH) may be limited.

The MRMS products are highly correlated with each other, suggesting the slight differences in skill scores are not significant. Bootstrapped confidence intervals of HSS for the different products found overlapping 95th percentile confidence intervals for nearly all product pairings further suggesting a lack of significance between each product's skill. Considering the lack of stratification of different hail sizes by specific MESH values, further work is needed to refine exact hail-size estimates from the MRMS system. Also, exact values of MRMS MESH should not be used as a direct proxy for the actual hail size that fell. Future work should include incorporating polarimetric, velocity data, and more sophisticated environmental parameters into the MRMS severe storm processing to aid hail identification and size discrimination.

ACKNOWLEDGMENTS

The author would like to thank the dozens of University of Oklahoma students who acted as the SHAVE operators during the 10 y of the project and making over 250 000 phone calls and collecting nearly 74 000 total reports. Kimberly Elmore, Alan Gerard, and Tiffany Meyer were helpful in providing valuable feedback improving the manuscript. The author also thanks John Allen, Dennis Cavanaugh, and Matthew Kumjian for their formal reviews that helped clarify and improve the text and figures. Funding was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement #NA11OAR4320072, U.S. Department of Commerce.

REFERENCES

Allen, J. T., and M. K. Tippett, 2015: The characteristics of United States hail reports: 1955–2014. *Electronic J. Severe Storms Meteor.*, **10** (3), 1–31.

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

Basara, J. B., D. R. Cheresnick, D. Mitchell, and B. G. Illston, 2007: An analysis of severe hail swaths in the Southern Plains of the United States. *Trans. GIS*, **11**, 531–554.

Benjamin, S. G., and Coauthors, 2004: An hourly assimilation/forecast cycle: The RUC. *Mon. Wea. Rev.*, **132**, 495–518.

——, and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694.

Blair, S. F., D. R. Deroche, J. M. Boustead, J. W. Leighton, B. L. Barjenbruch, and W. P. Gargan, 2011: A radar-based assessment of the detectability of giant hail. *Electronic J. Severe Storms Meteor.*, **6** (7), 1–30.

——, and Coauthors, 2017: High-resolution hail observations: Implications for NWS warning operations. *Wea. Forecasting*, **32**, 1101–1119.

Brown, T. M., W. H. Pogorzelski, and I. M. Giammanco, 2015: Evaluating hail damage using property insurance claims data. *Wea. Climate Soc.*, **7**, 197–210.

Changnon Jr., 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.

Cintineo, J. L., T. M. Smith, V. Lakshmanan, H. E. Brooks, and K. L. Ortega, 2012: An objective high-resolution hail climatology of the contiguous United States. *Wea. Forecasting*, **27**, 1235–1248.

Davis, S. M., and J. G. LaDue, 2004: Nonmeteorlogical factors in warning verification. *22nd Conf. on Severe Local Storms*, Hyannis, MA, Amer. Meteor. Soc., P2.7.

Dobur, J. C., 2005: A comparison of severe thunderstorm warning verification statistics and population density within the NWS Atlanta county warning area. *4th Southeast Severe Storms Symp.*, Starkville, MS, Mississippi State University. [Available online at https://www.weather.gov/media/ffc/SEconf.pdf.]

Donaldson, R. J. Jr., 1961: Radar reflectivity profiles in thunderstorms. *J. Meteor.*, **18**, 292–305.

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. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting*, **5**, 575–585.

——, H. E. Brooks, and M. P. Kay, 2005: Climatological estimates of daily local nontornadic severe thunderstorm probability for the United States. *Wea. Forecasting*, **20**, 577–595.

Dye, J. E., and B. E. Martner, 1978: The relationship between radar reflectivity factor and hail at the ground for Northeast Colorado thunderstorms. *J. Appl. Meteor.*, **17**, 1335–1341.

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.

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: A new analysis tool. *Mon. Wea. Rev.*, **100**, 548–552.

Hales, J. E. Jr., 1993: Biases in the severe thunderstorm data base: Ramifications and solutions. Preprints, *13th Conf. on Weather Forecasting and Analysis*, Vienna, VA, Amer. Meteor. Soc., 504–507.

——, 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.

Hocker, J. E., and J. B. Basara, 2008: A geographic information systems-based analysis of supercells across Oklahoma from 1994 to 2003. *J. Appl. Meteor. Climatol.*, **47**, 1518–1538.

Jewell, R., and J. Brimelow, 2009: Evaluation of Alberta hail growth model using severe hail proximity soundings from the United States. *Wea. Forecasting*, **24**, 1592–1609.

Johnson, A., and K. Sugden, 2014: Evaluation of sounding-derived thermodynamic and wind-related parameters associated with large hail events. *Electronic J. Severe Storms Meteor.*, **9** (5), 1–42.

Lakshmanan, V., and T. W. Humphrey, 2014: A MapReduce technique to mosaic continental-scale weather radar data in real-time. *IEEE J. Select Topics Appl. Earth Obs. Remote Sens.*, **7**, 721–732.

——, T. Smith, K. Hondl, G. J. Stumpf, and A. Witt, 2006: A real-time, three-dimensional, rapidly updating, heterogeneous radar merger technique for reflectivity, velocity, and derived products. *Wea. Forecasting*, **21**, 802–823.

——, ——, G. J. Stumpf, and K. Hondl, 2007a: The warning decision support system—integrated information. *Wea. Forecasting*, **22**, 596–612.

——, A. Fritz, T. Smith, K. Hondl, and G. Stumpf, 2007b: An automated technique to quality control radar reflectivity data. *J. Appl. Meteor. Climatol.*, **46**, 268–305.

Lenning, E., H. E. Fuelberg, and A. I. Watson, 1998: An evaluation of WSR-88D severe hail algorithms along the northeastern Gulf coast. *Wea. Forecasting*, **13**, 1029–1045.

Morgan, G. M., and N. G. Towery, 1975: Small-scale variability of hail and its significance for hail prevention experiments. *J. Appl. Meteor.*, **14**, 763–770.

Mosier, R. M., C. Schumacher, R. E. Orville, and L. D. Carey, 2011: Radar nowcasting of cloud-to-ground lightning over Houston, Texas. *Wea. Forecasting*, **26**, 199–212.

Mustered, S. K., and K. L. Ortega, 2012: Investigation of radar variables and near surface environments for developing a surface hail fall product. *28th Conf. on Interactive Information Processing Systems*,

New Orleans, LA, Amer. Meteor. Soc., 14 National Weather Service, cited 2016: *Storm Data* preparation. National Weather Service Instruction 10-1605. [Available online at http://www.nws.noaa.gov/directives/].

Ortega, K. L., T. M. Smith, and G. J Stumpf, 2006: Verification of multi-sensor, multi-radar hail diagnosis techniques. *Symp. on the Challenges of Severe Convective Storms*, Atlanta, GA, Amer. Meteor. Soc., P1.1.

——, ——, K. L. Manross, A. G. Kolodziej, K. A. Scharfenberg, A. Witt, and J. J. Gourley, 2009: The Severe Hazards Analysis and Verification Experiment. *Bull. Amer. Meteor. Soc.*, **90**, 1519–1530.

——, J. M. Krause, A. V. Ryzhkov, 2016: Polarimetric characteristics of melting hail. Part III: Validation of the algorithm for hail size discrimination. *J. Appl. Meteor. Climatol.*, **55**, 829–848.

Ryzhkov, A. V., M. R. Kumjian, S. M. Ganson, and P. Zhang, 2013: Polarimetric radar characteristics of melting hail. Part II: Practical implications. *J. Appl. Meteor. Climatol.*, **52**, 2871–2886.

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

——, J. J. Levit, S. J. Weiss, and D. W. McCarthy, 2004: The frequency of large hail over the contiguous United States. *14th Conf. on Applied Meteorology*, Seattle, WA, Amer. Meteor. Soc., 3.3.

Schuster, S. S., R. J. Blong, and K. J. McAneney, 2006: Relationship between radar-derived hail kinetic energy and damage to insured buildings for severe hailstorms in Europe and Australia. *Atmos. Res.*, **81**, 215–235.

Smith, P. L., and A. Waldvogel, 1989: On determinations of maximum hailstone sizes from hailpad observations. *J. Appl. Meteor.*, **28**, 71–76.

Smith, T. M., and Coauthors, 2016: Multi-radar multi-sensor (MRMS) severe weather and aviation products: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 1617–1630.

Tang, L., J. Zhang, C. Langston, J. Krause, K. Howard, and V. Lakshmanan, 2014: A physically based precipitation-nonprecipitation radar echo classifier using polarimetric and environmental data in a real-time national system. *Wea. Forecasting*, **29**, 1106–1119.

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.

Wilks, D. S., 2006: *Statistical methods in the atmospheric sciences*. Elsevier, 627 pp.

Witt, A., M. D. Eilts, G. J. Stumpf, J. T. Johnson, E. D. Mitchell, and K. W. Thomas, 1998a: An enhanced hail detection algorithm for the WSR-88D. *Wea. Forecasting*, **13**, 286–303.

——, ——, ——, ——, ——, and ——, 1998b: Evaluating the performance of WSR-88D severe storm detection algorithms. *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 Meteor.*, Austin, TX, Amer. Meteor. Soc., 368–369.

REVIEWER COMMENTS

[Authors' responses in *blue italics.*]

REVIEWER A (Matthew R. Kumjian):

Initial Review:

Recommendation: Revisions required.

Summary: The author presents a detailed evaluation of the multi-radar, multi-sensor (MRMS) system's radar reflectivity-based products based on Storm Data and SHAVE hail reports. Additionally, a detailed inter-comparison between SHAVE and Storm Data reports is presented. The author finds that the higher-resolution SHAVE reports provide a better means for evaluating MRMS products than Storm Data reports, which may be imprecise and only characterize the size of hail within a 5-km radius of the report. The performance of the different MRMS products were not significantly different, which makes sense given they are all based on a single radar product (reflectivity factor). Of them, reflectivity at lowest altitude is best at detecting any hail, and MESH with a threshold of 30 mm is best at detecting the presence of severe hail. However, no products were useful in discriminating different hail sizes.

The study is interesting and well-suited for EJSSM. The analysis is sound and careful, with robust statistics presented. My main complaint is that a lot of the description of the analysis could benefit from improved clarity or discussion, especially to outsiders who are unfamiliar with some of these techniques. As such, I recommend moderate revisions, with the detailed comments below.

Substantive Comments: What is the difference between a hailstreak and a hailswath?

The definitions are supplied in Changnon (1970). I have added two short definitions following each term.

Section 1b: Not sure if it's my own biased interests or not, but I feel like a few sentences describing what reflectivity factor actually is may be useful, especially given the strong correlations found for each MRMS product and the lack of skill in discriminating hail size. For example, there is an inherent ambiguity in reflectivity factor, as it is increased for increased hail concentration, size (to a certain extent), and liquid water content. Additionally, the "D^6" dependence often cited is only valid for electromagnetically small particles. When dealing with severe or significantly severe hail, this isn't valid anymore.

Agreed. Considering I've pointed out the limitations of single-radar algorithms, it is just as important to highlight the limitations of the physics of radar-based target detection and identification. I've added a short paragraph describing briefly the shortcomings of radar reflectivity factor to diagnose complex scatters like hailstones.

Section 1b: This discussion of products is great, but there is no discussion of any performance evaluation. Clearly, if we're still working on the problem, then these have some limitations? A critique of these techniques would add to the literature review.

I have added quick summaries of the evaluations of the skill of those algorithms, although all but the Donavon and Jungbluth study did not either evaluate the algorithms using the 2×2 table statistics, making the summaries somewhat difficult to phrase that make them comparable to the statistics I am reporting.

Is there any reasoning behind using the 75th percentile? It's OK if it is subjective based on trial and error, etc., but it would be helpful to know why this was chosen. (Update: I see that this comes into play later in the analysis…perhaps some foreshadowing of it here would be useful. For example, some motivation for the future comparisons will help readers understand what this is all about.)

I have added a note. However, there is a slightly not-subjective reasoning for selecting the 75th percentile. Anything below the median would lead to over half of the reports not matching to another report. The subjective part comes in using the 75th percentile as anything larger might be casting too large of a net and end up always matching to very similar sized reports.

Section 2b: I was confused here—are these specifications from the operational version or what you're using in the present study? If they are from the operational ones "described above", then it makes more sense to move this to the earlier section when the operational system is described.

This was more specific information about the real-time system. I have modified the text to say specifically why (essentially, the beamwidth is wide enough the beam is potentially "smeared" across several vertical grid points or just used a single grid point for my coarser resolution) there's no anticipation of a major impact due to the different vertical spacing used in my study compared to the operational MRMS system.

What is the difference between VIL and VII?

I guess out of all of these parameters, VII was never introduced nor is self-explanatory. I have added a small summary of VII.

Provide a reference for or describe what a "confusion matrix" is.

In meteorology, it's popularly been referred to as a "contingency table" or "2×2 matrix". Specifically, contingency tables deal with cross-tabulation of variables rather than prediction and observation, though you could include such things by having variables "observed A", "observed B", "predicted A", and "predicted B." A confusion matrix is a specific form of contingency table, but "confusion matrix" is the more precise term to use since I'm using it to calculate skill scores from the observations and predictions (side note: and those skill scores have different names depending on field, though meteorology has seemingly settled on probability of detection, false alarm ratio, etc.). I hesitate to call it a 2×2 matrix since the 2×2 nature of the matrix can be derived from predictions and observations of more than 2 classes—which I have done here by using different size categories and parameter thresholds. A quick search of AMS literature reveals the table referred to as both terms, without reference.

Section 3: The detailed inter-comparison of SHAVE and *Storm Data* reports is highly important, but I found myself trudging through numerous boxplots with no clear indication as to why or what the point was. Additionally, many of the descriptions in this section were confusing to me. (Perhaps I'm just dense?) For example:

Reviewing the captions and some of the text, clarification is definitely needed. I added a note at the start of the results section to clarify how the plots are set up. For all plots, the color of the box specifies what the size category of the originating report. Except for the analysis summarized in Figs. 8-9, there were never restrictions on matching with respect to hail size, only using a maximum distance.

The boxplots in Figs. 5 (6) are showing the nearest neighboring distance (diameter difference) of the same hail size *category*, right? Or just nearest neighboring report of any size class, and then you parse them into different size categories?

I've modified the captions in Figs. 3-6 to be more precise.

Boxplot in Fig. 7: this is the 75th percentile- was this chosen to be consistent with the suggestion of Chagnon (1968) mentioned in the introduction?

No, see explanation above on the 75th-percentile selection.

Figures 8 (and 9): Is my interpretation correct: the top row shows the distance (diameter difference) to any report (even null/no hail), the 2nd row the distance (diameter difference) to any report >12.7 mm, etc.? That discussion or the figure captions need to be crystal clear.

You are correct. I feel the text is clear about this analysis, I've modified the captions to better match the text and hopefully clarify the figures. I also realized the text did not clearly state what was implied in the figures: both databases had the minimum diameter applied when doing the matching. For example, in the 4th row the reports from each database were first filtered down to all reports with diameters equal to or greater than 50.8 mm and then those reports were put through the matching.

Section 3b: Revising statements like "Using SHAVE, for any sized hail, RALA equal to 55 dBZ provides the best threshold with a HSS equal to 0.42" can be clearer. For example, if I'm interpreting it correctly, something like "Using SHAVE reports, RALA greater than or equal to 55 dBZ provides the best detection of hail of any size (HSS=0.42)" or similar. Is the MESH threshold of 30 mm similar to the findings of Cintineo et al.? If so, it should be referenced here.

It's similar, but for a different hail size. Full story: the Cintineo et al. study used a little less than half of the data set used in this study; at the time I was still in the process of finishing data processing for the data used in this study (which eventually did not finish out until 2014). However, it looks like Cintineo were using a 19-mm threshold (the old penny-sized severe threshold for the NWS) and not 25.4 mm (quarter-sized hail severe threshold) as I am here. Also, it looks like we used a window technique (and not just using the point estimate) to score MESH vs. the SHAVE reports. So, yes, the HSS is similar, but for a different hail threshold, using a different methodology. It was a bit of cart before the horse; I was hoping to get something like this out before we had a Cintineo et al. like study, but here we are (…5 years later). Given the differences, I would prefer to not reference the study so as to not cause confusion by conflating that the 2 studies have similarities (besides the resulting HSS value).

I think you mean "all CSI scores peak at thresholds of 0". What is the physical interpretation of this? That, using *Storm Data* reports for validation, most of the MRMS products are useless? Something to help interpret this result is needed.

I definitely dropped the ball here. I've added a paragraph (which should have been there to begin with) in the discussion. The point of even showing the CSI plots and the results using Storm Data *was to highlight the need for a good verification data set along with a good selection of statistics to calculate in order to do a good evaluation of the products. I put the data in, but not the context.*

Do you mean that RALA is statistically significantly better and excluding RALA+H50_0C is statistically significantly worse?

I went back and forth whether to include that sentence because it did muddy things a bit. I've decided to remove it but yes, your interpretation is partly correct. If you use the bootstrapped CIs, RALA would be significantly better than all other products for any sized hail (though the magnitude of the HSS increase is quite small). For severe hail, you would not consider using RALA nor H50_0C.

"Colder (higher) altitudes" and similar statements are clunky. How about "lower temperatures (higher altitudes)"? Also, is the increased overlap of distributions at higher temperatures related to the impact of melting on reflectivity? This is relevant for the earlier comment. Finally, are the narrowing and broadening of the reflectivity distributions aloft a result of sample size? In other words, do small hail cases have fewer numbers of observations at such low temperatures, increasing the distribution width? This would be consistent with your findings that high reflectivity aloft is associated with larger hail.

The trends in the boxplots are not the result of sample size. In fact, ~90% of "no hail" reports have valid reflectivity values for the –50°C level; for sig. severe hail there still are 480/486 reports included in the distributions. The broader distributions aloft I think are more indicative of the range of spatial sizes of high reflectivity areas above the melting level. Some storms have big, broad areas of higher reflectivity (so these may extend over areas of the surface that receive no or small hail) and others do not. The observation/conclusion I make about high reflectivity aloft is associated with larger hail is made mostly looking at the medians of the boxplots (well, the upper-50 percentile to be precise). The median for significant-severe hail is still above 50 dBZ at -50C for all of the profiles, the median for severe hail is above 40 dBZ, while for the two smaller categories the median has fallen below 40 dBZ.

The increased overlap of the distributions at the warmer temperatures is a probably caused by two different issues. First, for the MESH profiles, it is most likely the result of matching to a parameter that is only the result of reflectivity aloft. So the peak MESH over an area may occur when there's a local minima of reflectivity—say for instance when a storm first formed, or maybe the local maxima—say when the largest hail was actually falling. SHAVE did try to collect observations (especially starting in 2008 and moving forward) at the very beginning of a storm to get an idea of where the first, even pea-sized, hail fell. Second, for the VIL and RALA profiles, I think the increased overlap is predominately from the saturation of the merger. It is incredibly difficult to get high reflectivity values, even if the nearest individual radar has a few pixels of very high reflectivity. This is because of the weightings applied to the individual radar observations (distance from radar, spread of the beam at a given location, etc.) and the coarseness of the MRMS grid (~1 km compared to ~250 m observations from the individual radars). I've added a paragraph highlighting these points in the discussion section.

I think an impactful sentence summarizing these points is needed, like: "In other words, MRMS products are useful for hail detection but not for hail size discrimination." I'd even add that dual-pol information should help!

I did have a note on dual-pol in the discussion, but it should be repeated here. However, I think saying MRMS is not useful for sizing is a bit misleading. For specific numerical sizing, yes it does a bad job—but I've yet to see an algorithm that does a good job at specific numerical sizing. But for categorical sizing, it's not as great as, say, the polarimetric HSDA (which is also categorical sizing), but it is comparable—especially for being single-pol and on a coarse grid compared to super-res WSR-88D observations. I clarified that final sentence to mean "exact, numerical sizing" and added a note on incorporating dual-pol.

[Minor comments omitted...]

Second Review:

Reviewer recommendation: Accept with minor revisions.

General comments: The author has done a great job at improving the clarity and flow of the manuscript. All of my comments from the original round were adequately addressed. I have one lingering re-tooling suggestion and some very minor typos/fixes. These are discussed below. Because of the minor nature of these comments, I recommend the paper be accepted for publication once these are fixed. I do not need to see the manuscript again.

All feedback has been incorporated. Thanks for the feedback.

[Minor comments omitted...]

REVIEWER B (John T. Allen):

Initial Review:

Reviewer recommendation: Accept with minor revisions.

Synopsis: The author presents an analysis of SHAVE observations of hail as compared to the data available from *Storm Data*, and compares the discrimination skill (between non-severe, severe, and significant severe hail categories) of different MRMS radar products with these two datasets. Overall, it constitutes a well-written study with a sufficiently detailed analysis, and some interesting confirmations of hail data characteristics (e.g. the spatial error in *Storm Data* reports), and the lack of independence between different radar products used in MRMS as potential hail predictors. My concerns for the manuscript are mostly minor and seeking clarification, or the addition of relevant references and material to the content to better convey some of the arguments.

Substantive comments: A couple of things are not clear regarding SHAVE. It is not entirely obvious to me whether SHAVE verification phone calls would obtain an accurate size measurement beyond that of *Storm Data* (other than the improved spatial density of reports and nulls). Were participants encouraged to use a particular method to measure hail size, or like *Storm Data*, were reference objects used? This is an important point given it may explain the wider confidence bounds for some of your results. It does not surprise me that the neighborhood variation between SHAVE reports is ±1", given the potential for "measurement" variation. Secondly, how many storms or storm days does this dataset reflect, as based on the figures it is clear many reports are from each individual storm?

The procedures SHAVE used to collect the data have now been summarized. In general, yes, SHAVE used common objects to get size estimates (unless the respondent was estimating actual lengths), but SHAVE tried to get precise with the objects. More specifically SHAVE would only record at ¼" increments for estimates, but for measured reports SHAVE would record the data in decimal inches.

I hesitate to let observer "measurement" error definitely be the primary reason for broad confidence bounds of the size differences, but I am sure it is near the top of the list. Many storms typically do not have a gradual gradient in maximum hail size in certain areas. I was reminded of this during a few of our storm intercepts this spring. The large hail simply stopped, as opposed to a steady drop to smaller sizes before hail fall completely ended. Even temporally at a static location there may be no lead up to larger sizes. I recall a SHAVE case for Burlington, CO, in which the storm suddenly began precipitating hailstones greater than 3" in diameter—no rain or small hail led up to this as reported by the respondents. I personally recall a storm chase where a storm I was under began to precipitate golf ball-sized hail without any smaller hail and barely any rain falling. All that said, the procedures that SHAVE operators followed hopefully limited big observer errors from getting into the database, but obviously cannot prevent all observer error from getting into the database.

Unique days number 229 for the data used here. Best I can say is the number of storms is around 389, as reported in the text as SHAVE operations. I'm certain several of these operations include multiple storms as I would "bundle" together those into a large case if the temporal and spatial closeness was satisfactory (e.g., storms in the same ~5 county region would be most likely in a single "operation" as long as the times the storms were sampled overlapped).

This [report count] contrasts the statement [above]. It is of course sensible given your database stretches prior to the elevation of the hail size criteria in 2010, and thus contains reports between 0.75" and <1.00" (only 6 reports in the full 1955-2014 climatology are sub .75"). There needs to be more context provided here as to the nature of the reports (which are inferior to the SHAVE sub-severe dataset), and to balance out how this contrast occurs between different paragraphs in the manuscript.

I've added a footnote near the Witt reference to clearly state that with the change to 1" severe, Storm Data *(using 2013–2016 as a reference) is now about 25% nonsevere reports. I also added a discussion in Section 4b on the change in sub-1" reports in the data I used.*

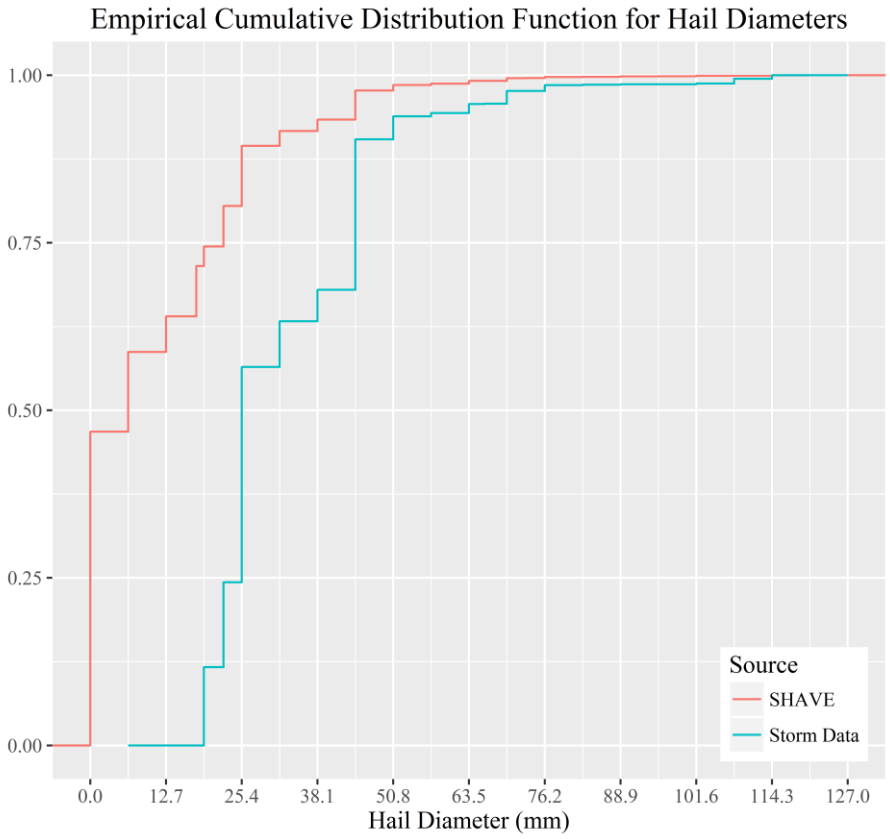
It's interesting you say there are only 6 reports that are sub-0.75" for 1955–2014. I grabbed the latest CSVs from NCEI and found 22 sub-0.75" reports in 2012–2015. Excluding 2015, there's 13 (still higher than the 6 you are reporting). There might be a conversation here (outside of these review replies, of course) on how intact the hail reports, and severe weather events in general, are as they are passed from the NWS offices to the NWS Verification Branch to NCEI to the SPC (not that adding 0.01% to the report totals is going to greatly affect climatologies or other use of the report database).

I'd be very cautious using the hail path data in this way [intermediary points included in hail counts] (especially given its availability is highly conditional on the WFO policy)—it makes a fairly gross assumption about the propensity of a storm to produce hail, and that the storm data maximum report along that path is reflective of the true hail size distribution within the storm's path. I'd be more comfortable if the maximum size were only used at the start point of the path as it is not clear what the WFOs are using to generate this path data.

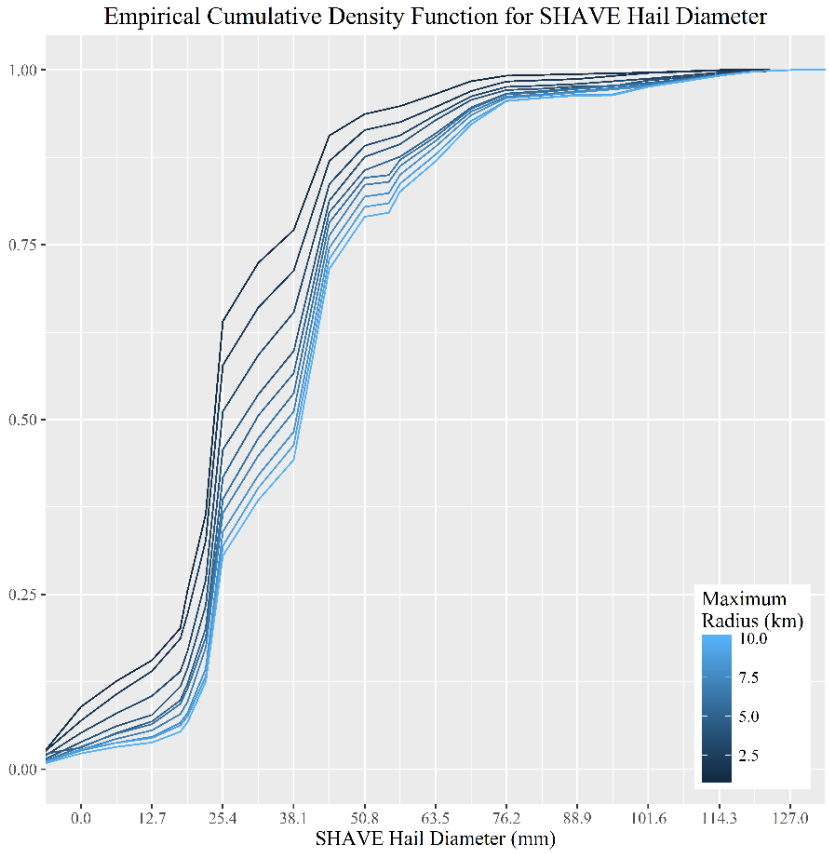
The hail path reports are problematic, but using the start point only is just as arbitrary as using the entire path. Fig. 1 in Ortega et al. (2009) there's a hail path (not determined from SHAVE data) with comment about golf balls extending pretty much along the entire path (which given the SHAVE reports, we can see that's not true). Further, the start point for that path is on the NW side of Madison, so it's not even the correct location even given the attached narrative said the hail was SE of Madison. I recall seeing path data in Alabama for storms in 2004 (or around that year) that just extended from county line to county line (for several counties), presumably along the storm path (so the start points are just county lines). But the generation of point reports probably have just as much uncertainty in their generation also: was it the max or was it the maximum a mobile observer was willing to have fall on them before they moved? Did they get an exact location (and ensure the lat/lon matches that location) or were the observers guesstimating their location as a distance and location from a town or landmark? If a static observer submits several reports, which report is used: the first severe, the max, them all? That's all kind of the point of the paper: there's a lot of uncertainty in Storm Data *and I've quantified a bit of it. But if you use* Storm Data *as a point verification source for radar, it's not going to work (and I think I've shown that). Worse yet, the MRMS grids are at 0.01° spacing, which is the precision most* Storm Data *report locations are recorded (now* Storm Data *reports can have up to 4 decimal place location precision in the lat/lons if they are entered by clicking on a map—this is implying 10s of meters of precision!).*

Figure 10 / Results: It would be useful to see a PDF or ECDF of the maximum hail size observations that appear in this version of SHAVE for the reader's reference and how it compares to Storm Data rather than having to infer this from Fig. 10.

Are we talking in general here or for these matches? Here is for all the data used in the study:



Here is the ECDF of SHAVE diameters for the different search radii and finding the maximum SHAVE report within that area.



Much of the central Plains doesn't have this sort of grid network—for example Fig. 11 of Allen and Tippett (2015) illustrates that gridded roads do not necessarily mean gridded reports, rather there is a clustering toward major roads.

While major roads (paved state highways at minimum size) there is not a gridded network, for unpaved local roads on which many residences are located, the Central Plains has a fairly consistent layout of grids at about 1-mi spacing. The clustering near major roads found in Allen and Tippett is probably because many storm chasers/observers are not going down the local roads; the clustering is also a hint that while many folks may be impacted and observe hail, they are very rarely reporting it. SHAVE being a phone-based, remote collection project did not care about road type just whether a residence existed and if a phone number could be obtained for that residence.

While I agree with this interpretation [re: lack of separation of the reflectivity profiles with respect to different environments], a third potential hypothesis here (and I suspect it is a combination of all three) is that the non-reliability of size observations in storm data (noted by Blair et al. 2017), the quantization of the size report data (Allen and Tippett 2015), and the relatively small sample size in SHAVE may mean that meaningful environmental discrimination is not possible—there are only a small number of observations for large hail sizes, especially in SHAVE. The analysis by Johnson and Sudgen (2014) suggests that there is potential for better discrimination given sufficient observations of the desired threshold. This point should be worked into this argument.

Sample size may well be an issue. I think the other issue here is the fact we run the MRMS algorithms everywhere, for every environment. I think there's some investigation to be had to only do hail analyses where we have background information that suggests large hail (both from the environment and the vertical reflectivity profiles). I've added a sentence on sample size and worked the Johnson and Sugden results into that paragraph.

[Minor comments omitted...]

Second Review:

Reviewer recommendation: Accept with minor revision.

Synopsis: The author has addressed the majority of my concerns and provided the clarity necessary where the text was unclear. I have a couple of other minor revisions; including both of the ECDFs the author produced as an additional figure, and a suggestion to modify Fig. 1 for readability, however given the extremely minor nature of this second round I do not need to see the manuscript again.

I incorporated all further feedback, including adding the new figures. Thanks for the feedback.

[Minor comments omitted...]

REVIEWER C (Dennis Cavanaugh):

Initial Review:

Recommendation: Accept with major revisions.

General comments: Overall the quality of presentation of this paper is of very high quality. The paper is well-organized, the figures look good and are well-explained and the quality of the writing is fantastic. This paper is certainly relevant and does a good job demonstrating the value of the SHAVE dataset compared to the National Weather Service Storm Data dataset of hail. MRMS data are widely used operationally in the National Weather Service to make warning decisions, where the goal is to increase the resilience of communities all across the United States to the impacts of hazardous weather. The author's work is important, and this paper does a good job documenting all of the hard work the author has done to collect, interpret, and evaluate hail data and MRMS products. I do have a few major concerns regarding the scientific content of this research, which are detailed in order of concern below.

Substantive comments: Regarding Figs. 11–14, which shows the skill scores for various MRMS fields in their ability to discriminate hail size categories, there appear to be some fundamental errors in calculating the Heidke Skill Scores (HSS) and Critical Success Indices (CSI) in these figures and cited in parts of sections 3-5 in the manuscript. This is a major concern because these two statistical skill score calculations are presented as the basis of the evaluation of the MRMS data throughout the study. The author does not specifically state how the graphs were generated in figures 11–14 from the confusion matrices, so I apologize if I'm misunderstanding or misinterpreting the results shown on the graphs of these figures. Based on my interpretation of these figures, I have the following comments:

I've added some clarifying text in section 2c on the method. I think your interpretation below is correct, but to be clear: each point on each line of CSI and HSS is derived from a different confusion matrix. For example, for MESH using SHAVE data, with 101 thresholds and 3 hail size categories, 303 different confusion matrices were tabulated and then the CSI and HSS scores were calculated and plotted. The line graph, while technically incorrect since there's no relationship between each point, provides a cleaner representation of the statistics than a point graph of upwards [of] 909 points on some of the graphs.

a. In all of the figures, the author appears to be showing the skill score of each MRMS product in forecasting the presence of hail, severe hail, and significant severe hail for various thresholds of each MRMS product. [For] Fig. 14, the author states that the peak CSI scores are at thresholds of 0 for the *Storm Data* database. If I'm interpreting the graphs and calculation of CSI correctly in these instances, it appears that the author is showing that at a reflectivity of 0, MRMS has greater than 0 positive forecasts for severe hail based on the equation for CSI in the form of X/(X+Y+Z) where X denotes the number of positive forecasts that correspond to an occurrence of the event, as stated in Schaefer (1990). For CSI to be a nonzero number, X must not be equal to zero, yet in Fig. 14 at a reflectivity of zero, CSI is listed at just below 0.8. It's not clear how a reflectivity of zero can be counted as a positive identification for severe hail when a reflectivity of zero should be completely clear air (e.g. even non-

meteorological clutter should be absent). This apparent problem is present in each calculated field in Fig. 14 and exists in Fig. 13 as well. It is possible that the author was counting a proper forecast for the absence of hail at these thresholds as a positive forecast of an occurrence of no hail, but if that is the case, the figures should be redrawn/recalculated to show only positive forecasts for hail when MRMS fields would indicate that hail (or hail categories) should be present. If my interpretation of these graphs is correct, then the author should note that CSI peaks at some MRMS values that are nonzero when using the *Storm Data* dataset.

The thresholds tested were done over a wide spectrum because: 1) this is the first thoroughly quantified evaluation of MRMS products for hail identification and size discrimination, and 2) the selection of the lower and upper bounds of the thresholds tested were to keep the plots nicely spaced between the multiple products. The filling out of the confusion matrix was consistent for each threshold, regardless of the physical meaning (if any) behind the threshold selection. A discussion (which I inadvertently left out) on the meaning of the CSI peak at 0 value for all MRMS products is now included in Section 4b.

b. Based on my interpretation of HSS, the HSS values in Fig. 12 should not start out as negative numbers for zero values of MRMS fields. Based on my understanding of the HSS, the only way that value should be negative is when "expected correct" forecasts exceed the sum of "hits" and proper forecasts for non-events. Again, just taking the zero reflectivity value in the MRMS fields as an example, a negative HSS at 0 reflectivity implies that a correct forecast for hail is expected, but that neither a hit nor a correct forecast for no hail was made. The fact that Fig. 11 does not show this quirk indicates that the author may have used a different technique for calculating the verification of SHAVE data versus *Storm Data* hail data. Where Figs. 11 and 12 seem to converge on a logical solution is at very high reflectivity and VIL values, where it is likely no observations existed; both calculations move to zero as you would expect (e.g. for reflectivity values >75). It's not clear what happened here, but it appears that different verification techniques were applied to SHAVE versus *Storm Data* hail. Is it because SHAVE data contained proper forecasts of no hail (e.g. no precipitation, or a simple rain report) whereas *Storm Data* does not have this information? If that is the case, some effort should be made to estimate correct forecasts for non-events in *Storm Data*, or perhaps simply do not use *Storm Data* to evaluate MRMS fields at all and make a case to only consider SHAVE data in the MRMS field evaluation earlier in the paper.

There was no difference in the treatment of Storm Data *and SHAVE when calculating the confusion matrices, except for not doing scoring of 'any sized' hail with* Storm Data *since there are no 'no hail' reports available through* Storm Data. *Regarding the negative HSS scores, it is entirely reasonable for those to be there and for reasons explained within the text. A negative HSS score comes about if the number misses multiplied by the number of false alarms is greater than the number of hits multiplied by the number of correct nulls. To get a miss (and correct null) with a value of 0, the parameter would have to be missing completely (i.e., no MRMS value, so the report was not within the swath of MRMS product; these were assigned a value of -99900 in the data files, so they can be scored properly). So it is possible to get a full confusion matrix even at 0 value threshold. That said, it's possible to get nearly 0 correct nulls and end up with a negative HSS value. The fact that the HSS for* Storm Data *flatlines at 0 for many of the products above reasonable values (as it also does for SHAVE) for those products gives me confidence the calculations are correct.*

Now included in section 4b is a discussion on the number of available reports <25.4 mm in Storm Data. *Since* Storm Data *is skewed towards sizes > 25 mm, the chances of a report resulting in a miss—especially considering the spatial uncertainty associated with the* Storm Data *reports—is greater than compared to using the SHAVE database, which is predominately of the smaller sizes. The whole point is depending on the goal at hand (in this case evaluating a product at ~1 km grid spacing) the selection of verification and skill score is important as to how to interpret the results. If someone had come along (or does come along) and evaluates MRMS products with* Storm Data *and finds little skill, the explanation is far more complex than MRMS is doing poor job—especially if it's done a grid point-by-grid point level.*

c. Based on my interpretation of Figs. 11–14, the recommendation is to re-run these skill scores so that CSI and HSS approach 0 as MRMS fields approach 0 for the positive identification of hail. Rerunning the skill scores with near-zero values of MRMS data approaching zero HSS and CSI should help identify the true peak in skill associated with each MRMS field. As it stands, the true peak appears as though it is masked by the early/low-value MRMS fields that currently have high HSS and CSI. If I have simply misunderstood the graphs and calculations performed by the author, I apologize in advance.

The stated purpose of the paper is to evaluate the ability of MRMS products to accurately model the hail fall characteristics of convection. The author seems to spend quite a bit of time comparing the SHAVE and *Storm Data* hail databases leaving less focus on evaluating MRMS fields for accuracy. The author makes a strong argument that SHAVE data is more accurate in terms of capturing the true hail fall characteristic of convection when compared to *Storm Data*, but this does not seem to help evaluate MRMS data in this study as both *Storm Data* hail and SHAVE data are used in MRMS product evaluation. If the author wants to show that SHAVE data is superior at capturing the hail fall characteristic of convection when compared to *Storm Data*, there are a number of statistical techniques that could be employed to prove the point. If this is a goal of the paper, then I would encourage the author to employ some of these techniques to prove that SHAVE data is indeed doing a better job capturing the hail fall characteristics of convection when compared to *Storm Data*. Subjectively, there is little doubt that SHAVE data is of superior quality to *Storm Data* hail, but there is primarily circumstantial evidence provided in the paper in its current form.

If showing that SHAVE data is better than *Storm Data* hail is the primary purpose of the paper, it would probably also be useful to mention some of the limitations of the SHAVE database. The author enumerates the multiple faults of the *Storm Data* hail database; however, there were no limitations mentioned for the SHAVE database as an objective comparison. Alternatively, if the author is trying to provide evidence to use SHAVE data instead of *Storm Data* hail in the evaluation of MRMS products in this study, then the author is encouraged to state this in the paper and consider not using *Storm Data* hail as a basis of evaluation for the MRMS products. I think the author makes a strong and valid argument that SHAVE data is more accurate than *Storm Data* hail, and if that is the case, why use Storm Data hail to evaluate MRMS fields at all? The evaluation of MRMS products for their ability to identify hail in this study seems lost in the continued comparison of *Storm Data* hail to SHAVE observations. It seems like there are two competing papers in the manuscript at times; one that seeks to prove that SHAVE data is superior to *Storm Data* hail, and one that seeks to evaluate the ability of MRMS fields to identify hail in convection for operational meteorologists. I think the paper would benefit from a focus on one topic or the other.

*Regarding time spent on the SHAVE-*Storm Data *comparisons, this is mostly a function of the ease in which the analyses can summarized and explained. However, thousands of confusion matrices have been summarized in 4 figures plus the thousands of reflectivity profiles summarized in another few figures, plus the thousands of reflectivity profiles not even shown. The amount of analysis to make the conclusion about MRMS performance, and the reasons for that performance, was as deep as the comparisons of the databases, just more easily summarized.*

The point of the analyses go beyond specific applications for the operational community, though you can certainly apply these findings to using MRMS products in the operational community. The purpose for showing Storm Data *and SHAVE evaluations of MRMS products was to show how different the answers are. The bigger picture here is having completed a dual-pol hail algorithm evaluation that only used SHAVE data, having completed this analysis, and completing other analyses, I'm finding that sometimes you need SHAVE data (especially when doing grid point-level analyses like was done here) and sometimes* Storm Data *is good enough to do an evaluation. So, just my own personal knowledge tells me to report the findings using both databases. And to be honest, I'm not sure where the harm is there. It only informs the community what analyses you can and cannot complete with certain databases.*

Further, to reviewer's comment that comparisons of SHAVE and Storm Data *could occur in an isolated setting, I've tried that before and the feedback was it was not scientific enough to stand on its own. If that's the feedback, that's fine; here is the comparison again, this time combined with analysis using both*

databases to show why it is important to quantify the differences between the databases, in both the actual reports and the analyses using those reports.

If comparing data from 389 cases over 7 years, yielding tens of thousands of reports from one database and thousands from other, using several different methodologies and reporting those differences is only circumstantial, then I guess I am at a loss for what would be considered a non-circumstantial evaluation. I can easily run (in seconds) statistical tests on the data and find that yes, the Kolmogorov–Smirnov test yields a statistically significant result that the distributions of the SHAVE and Storm Data *reports are different. Well, I know that from the fact SHAVE reports are predominantly not of the sizes found in Storm Data, I do not need a statistical test to validate that. I can do a permutation test on the SHAVE nearest neighbors to Storm Data and find it is statistically different. Great, but what's the hypothesis I'm rejecting? That the databases are not different. Okay, but what do I do with that? It's not telling me SHAVE is better. It's not telling me* Storm Data *is better. It's just saying they are different. The point to significance testing is to test a hypothesis and either reject or do not reject that hypothesis. I can't think of a meaningful hypothesis to test using just the neighboring report sizes alone that would result in establishing anything of consequence. Further, I use the reports in an analysis. Both databases are treated the same. The resulting skill score calculations are markedly different, in fact bootstrapped 95% confidence intervals come nowhere close to overlapping, suggesting statistical significance. Now I could have run permutation tests to definitely establish the significance, but considering HSS calculated using SHAVE compared to* Storm Data *can at times be an order magnitude different and possibly of a different sign, I'm not sure what taking the time to execute those tests do other than provide a p-value.*

The reviewer is correct about potential errors in the SHAVE database, and per other reviews also, I added more detail on SHAVE collection strategies and shortcomings of the SHAVE reports.

In Fig. 16, it's unclear how the vertical changes in VIL were calculated and then evaluated or discussed in the supporting manuscript. It's unclear to me how VIL values have significant overlapping IQRs when calculated from the +20°C level and the –50°C level. If VIL is calculated as the vertical integration of reflectivity from the identified isothermal level to the "top" of the storm, you would expect that VIL calculated from the –50°C level would be much smaller than VIL calculated from the +20°C level as the amount of reflectivity below the –50°C level should all be removed from the calculation. With around 90% of the storm existing below the –50°C level, I would expect that VIL would be around one order of magnitude lower than the VIL calculated from the bottom of the storm. The author mentions that the hail producing storms displayed higher values of reflectivity aloft than in the lower portion of the storm, but the –50°C level should be near the echo tops in many cases, and reflectivity should drop off on average with the temperature being too cold to support a mixed phase of hydrometeors. Even if the values are wrong, I'm not sure if recalculating them would have value for evaluating this particular MRMS field, but these values seemed strange based on how I understood VIL to be calculated in this figure.

VIL was calculated for the entire storm depth. Figure 16 shows the profiles of the MRMS reflectivity at the time of each of those MRMS product maxima. The methodology is explained in the last paragraph of Section 2c.

Some opportunities for discussion seem left unanswered. For instance, is it possible that SHAVE data can improve MRMS hail detection? Does the author have any recommendations to change MRMS products based on these results? One of the more interesting findings in the paper is that evidence was provided that environmental data offered little to no utility in discriminating hail detection by MRMS products, yet there is no recommendation that MRMS discontinues the use of environmental data in its hail detection algorithms if it's not helping. Based on some of the stated results that taller columns of higher reflectivity values are there proposed new MRMS products that the author suggests investigating? I think there is a lot of good and operationally relevant content here that the author could discuss if more emphasis is placed on using the presented research to identify what's working and areas for improvement for the current MRMS suite of hail detection algorithms.

I'm a little puzzled here on the comment. The manuscript provides evaluations of 9 products and identifies the skill of those products in identifying where hail of 3 size categories fell. The manuscript also provides

some insight as to why the products perform as they do by plotting the vertical profiles of reflectivity at the time of the maximum of 3 of those products and showing the correlations of those products to each other. The manuscript demonstrates that using Storm Data to do a grid point-level evaluation of MRMS products is a non-starter. This is hugely important considering SHAVE has long since stopped operating and developing future products might rely on using only Storm Data *(though for the foreseeable future SHAVE data should get us to the next level of MRMS outputs). I conclude that exact hail size estimates from the MRMS system at this time are not really possible, not because of any fault of the algorithms and techniques employed, but because the vertical reflectivity profiles for even broad size categories generally have large overlapping distributions of reflectivity with their neighboring categories. I also showed that using simple combinations of environmental parameters did not help stratify the reflectivity profiles (and thus would not stratify MRMS products); thus rules of the day based on an environment (e.g., minimum MESH of X means hail size Y) cannot be used (I've added an explicit statement of this into the text). Throughout the discussion and conclusion sections several different avenues of future work, including exploring polarimetric and storm rotation variables in the MRMS framework. Some of these have been added in the revision process, however, many were already present in the text.*

[The end of section 4] seems speculative. There's no evidence presented that supports the claim that the coarseness of the environmental analysis (either spatially or temporally) can explain the lack separation of reflectivity profiles in varying environmental parameter space. The resolution of the environmental parameters used in MRMS or in this study are not stated, and it is assumed that a large number of cases were collected and used due to the high precision of the thresholds chosen as breaking points for the low, moderate and high categories given in Fig. 17 (e.g. thresholds of CAPE chosen at the nearest 1 J/kg). If thousands of points of environmental data are included in the study, it doesn't seem reasonable to assume that the representativeness of the analysis is a problem.

I used the 20-km RUC and RAP grids and have added it to the data section. I also do not explicitly say that the coarseness is the only factor, but it certainly has to be considered especially because storms do modify the environments in which they form and how that modification occurs and impacts the resulting storm- to micro-scale processes is not well explored. I added a note that the analysis may be too simplistic (but recall that within the MRMS system, most things currently are simple combinations and sophisticated techniques to diagnose severe storms are still being explored) Per another review, I added the results of Johnson and Sugden (2014), which shows possibly the need to add more sophisticated environmental indices in exploring this topic.

[Minor comments omitted...]

Second Review:

Reviewer recommendation: Accept with major revision.

General comment: The author has addressed several of my concerns in this revision, and I certainly appreciate that. I have only one real major concern that remains, then there are several minor comments/concerns that are primarily grammatical in nature.

Major concern: The calculation of skill scores and their representation in figures 11–14 still seem incorrect. The entire graph does not look incorrect, and I think the peak magnitude of the skill scores likely will hold that SHAVE data skill scores peak at a significantly higher magnitude than Storm Data skill scores, but I can't be certain because from the graphs, it appears that there may be fundamental issues with the way skill scores were calculated. This is of significant concern because the calculation of the skill scores is derived from the hundreds of confusion matrices that are used as the primary point of evaluation of various MRMS fields using SHAVE and *Storm Data*.

Looking at Fig. 13 for instance, the calculation of CSI using SHAVE data as the evaluation basis for MRMS grids, the COMP, H50C, H60C, and RALA grids stand out as having a high skill score in what appear to be trivial null circumstances. For instance, how many correct forecasts for "any hail", severe hail, or significant-severe hail could be expected at a composite reflectivity of 0? I would anticipate that 0

forecasts for hail occur at this threshold yet there is a flatlined skill score of somewhere between 0.5 and 0.6 represented on the graph.

```
#$thresh = looped index for the current MRMS threshold
#$hail_thresh = looped index for the current hail size threshold
#$hail = vector of hail sizes
#$mesh = vector of MRMS values
#$o = index for the current hail/MRMS pair being evaluated
#$hail and $mesh are the same size
if ($hail[$o] >= $hail_thresh && $mesh[$o] >= $thresh) { # hit
$h++;
} elsif ($hail[$o] >= $hail_thresh && $mesh[$o] < $thresh) { # miss
$m++;
} elsif ($hail[$o] < $hail_thresh && $mesh[$o] >= $thresh) { # false alarm
$fa++;
} elsif ($hail[$o] < $hail_thresh && $mesh[$o] < $thresh) { # correct null
$cn++;
}
```

Above is the exact code snippet that does the comparisons to fill-in the confusion matrix. It's used for both SHAVE and Storm Data skill score calculations and used for all thresholds. I think the reviewer is confusing the point of the thresholds. There is not an anticipated number of forecasts for a given threshold; I am using the threshold as a forecast. The purpose is to explore what threshold is most successful in discriminating the hail size classification under investigation. As I said in my first reply, since this is a first quantitative analysis of the MRMS products, a thorough investigation of the product values, even potentially unrealistic ones, should take place. Additionally, the bounds on some of the products were used to keep the axes of the resulting figures fairly clean and even.

Without seeing the data one of two things appear to be happening:

1. The author is using correct forecast for "no hail" at these lower thresholds of reflectivity, in which case I would expect the skill score to be at 1.0 OR
2. The graphs are being extrapolated from the first "hit" (e.g. there was some hail reported at a composite reflectivity of 35) back to the left to zero because there were no data to score.

What's truly confusing is that on the far-right hand side of the graphs, where reflectivity values are equally unlikely to be associated with hail (e.g. composite reflectivity of 100), the skill scores drop to "0" as you would expect when there are no hail cases to evaluate. This leads me to think that there is a problem with the way confusion matrices are generated/scored to the left of the peak values in the SHAVE evaluation, and just on the left-hand side in general in the Storm Data evaluation.

At very high thresholds where there are no matching reports as the threshold is too restrictive, there would generally only be misses and correct nulls. This results in a 0 in the numerator portion of the HSS calculation and a 0 overall for the HSS value. For lower thresholds, a flatline along 0 most likely means generally only hits and false alarms being scored (since we essentially have positive predictions for every data point) and thus a 0 HSS value.

Where the graphs and calculations appear valid, the skill scores have a non-zero slope indicating that there are data present to evaluate. The figure that "looks right" is Fig. 11: the HSS for MRMS grids using SHAVE verification. It seems strange that the HSS can flatline at 0 from 0–25 dBZ while CSI remains stable above 0.5 at the same thresholds.

Since at low thresholds we will generally only register hits and false alarms, the flatlined CSI scores are essentially just the proportion of hail categories (since we only have H / (H+FA) for the CSI calculation). As 46% of the SHAVE reports are of 'no hail', the CSI is ≈0.54 for composite reflectivity (it's a little less probably due to some correct nulls being registered for reports with no value at all, e.g., missing values).

Finally, in Fig. 12, the negative HSS associated with *Storm Data* don't appear to be "entirely reasonable" based on discussion in the text or presentation on the graphs. Based on lines 464–466, the author states that

"no hail" reports are unavailable using *Storm Data* to evaluate MRMS data. If "no hail" reports are unavailable, how are they being scored or used to generate confusion matrices when there is no hail present at lower MRMS threshold values? One possible explanation is that SHAVE data are being used as a point of comparison to *Storm Data* where no data exists. If this is the case then do the graphs represent using *Storm Data* to evaluate MRMS products or do they represent the difference in *Storm Data* and SHAVE data? If the difference in skill is being shown in these graphs, then I would simply suggest changing the label of the graph to show that this represents the difference in skill scores between the two datasets. If SHAVE data are being inserted in place of Storm Data for "no hail" reports, then *Storm Data* is not truly being used to evaluate MRMS data.

Even without the negative HSS, it is easy to show that SHAVE data does a better job evaluating MRMS data as the peak HSS is likely to remain .2 or more below the SHAVE evaluation. It seems entirely reasonable for HSS to be negative at some point in the analysis, but immediately at the beginning (zero or near-zero values for MRMS products) of each graph, it doesn't make sense. If the skill scores are wrong for some points, they may be wrong for all points. I realize that sharing the component confusion matrices for the hundreds of points of data evaluated in this study is absurd, but maybe sharing the composition of a few significant thresholds (e.g. near zero, min, max scores) would be useful here. My main concern here is that there is really no way to evaluate that the skill scores are being calculated properly except from these graphs of aggregate skill scores. Again, while there is little doubt that SHAVE data will outperform *Storm Data* skill scores, the graphs themselves don't appear to be conveying realistic skill scores at low MRMS thresholds, and that's really all I have for evaluation.

"No hail" reports are not being used in Fig. 12 since Storm Data *does not have "no hail" reports. I'm unsure why the reviewer would think so as there's nothing in the paper to even suggest that. SHAVE- and* Storm Data*-based product evaluations were conducted separately and reports from one database were not used in the other's evaluation. Figure. 12 is the HSS for each of the MRMS grids using Storm Data as the evaluation, not a difference. As to the negative values, as mentioned above, for low thresholds, since no correct nulls can be logged (since generally only hits and false alarms will be logged due the very loose prediction value), the numerator of the HSS is negative and thus a negative HSS value.*

Here's the confusion matrix for the composite reflectivity, 0 dBZ threshold, for severe hail for SHAVE and then Storm Data:

Pred ↓ / Obs →	***Severe***	***Non-Severe***
Severe	*4134*	*17047*
Non-Severe	*0*	*3*

Pred ↓ / Obs →	***Severe***	***Non-Severe***
Severe	*2555*	*719*
Non-Severe	*61*	*0*

These tables lead to an HSS for SHAVE of 0.0001 and an HSS for Storm Data of -0.035, which matches the graphs.

Now for peak HSS score (SHAVE = 62 dBZ and Storm Data = 60 dBZ) and with the same parameters as above:

Pred ↓ / Obs →	***Severe***	***Non-Severe***
Severe	*2224*	*2701*
Non-Severe	*1910*	*14349*

Pred ↓ / Obs →	***Severe***	***Non-Severe***
Severe	*1336*	*1280*
Non-Severe	*286*	*433*

This leads to HSS values of SHAVE = 0.354 and Storm Data *= 0.075, which match the graphs.*

[Minor comments omitted...]
Third Review:

Reviewer recommendation: Accept.

General Comment: I appreciate Kiel's efforts to share his code with me and a couple of the confusion matrices. The ROC curve that he added looks right, and makes a lot more sense to me than the CSI and HSS plots. If he generated the ROC curve from the confusion matrices, then I'm a lot more confident that the skill scores (CSI/HSS) were created properly. I'm still not sure about the left hand side of the *Storm Data* CSI scores in particular, especially what the meaning of those curves are, but there's no sense holding up his publication because I don't understand part of the graph. It is possible that it's created properly and he's just using a technique or presentation of the data that's unfamiliar to me and therefore difficult to process. The ROC curve looks great, and makes sense, so if that is generated from the same collection of confusion matrices, the other graphs are probably fine as well. Thank you!