Johnson, A. W. and K. E. 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.

Electronic Journal of
SEVERE STORMS METEOROLOGY

Evaluation of Sounding-Derived Thermodynamic and Wind-Related Parameters Associated with Large Hail Events

AARON W. JOHNSON AND KELLY E. SUGDEN
NOAA/NWS, Weather Forecast Office, Dodge City, Kansas

(Submitted 12 March 2014; in final form 09 December 2014)

ABSTRACT

Severe-convective hailstorms are one of the most frequent weather hazards across the United States. However, studies evaluating the ability of various environmental indices to differentiate lower-end severe hail (≤1.25 in, 32 mm) from significant hail (≥2.0 in, 51 mm) prior to storm formation are limited and typically overlap very little with microphysically based research. To bridge this gap, this study builds a database of 520 hail reports that sort into one of four hail-diameter ranges. For each report, various thermodynamic and wind-related fields are then extracted from Rapid Update Cycle (RUC) model analysis to create a parameter-based hail climatology.

Analysis of these environmental indices indicates most wind-based parameters display weaker magnitude winds and resultant shear for the smallest hail-size bin compared to the three largest. Further, the three largest hail diameter bins reveal nearly identical parameter values in the lowest 6 km AGL. In contrast, non-traditional shear layers that include winds in the upper portions of a storm (>6 km AGL) display some skill to differentiate larger hail sizes, especially for ≥3.5-in (89-mm) hail. Thermodynamic variables produced mixed results, with variables such as CAPE displaying a slight tendency to increase as binned hail size becomes larger but still with significant overlap. On the other hand, non-traditional parameters such as the hail-growth-zone thickness reveal a relationship toward decreased depth as the binned hail size increases, but with little to no increase in hail-growth-zone CAPE. Finally, evaluation of the significant severe parameter (SSP) and a new index called the large hail parameter (LHP) display mixed results. Skill at delineating ≤1.25-in (32-mm) report from 2.0–3.25-in (51–83-mm) cases for LHP (SSP) is slightly better (worse) than 0–6-km AGL bulk vector shear. However, the LHP displays improved skill over any other parameter to differentiate ≥3.5-in (89 mm) reports from those with <2.0-in (51-mm) diameter hail. The LHP formula creates improved skill by including non-traditional environmental parameters typically associated with storm longevity, precipitation efficiency, and hail-growth rates.

1. Introduction

The United States has one of the highest frequencies of severe-convective weather in the world with large hail accounting for a sizable percentage of this climatological normal (Frisby and Sansom 1967; Laing and Fritsch 1997; Doswell and Bosart 2001; Cintineo et al. 2012). Estimates of annual economic loss in the United States due to hail damage exceed $1.0 billion (Changnon 1972, 1999). However, a single event with extremely large hail (diameters ≥3.5 in or 89 mm) has the potential to surpass that. A commonly referenced example of this type of case is the 5 May 1995 Fort Worth, TX hailstorm in that people attending the "Mayfest" event were caught outdoors in a storm producing hail with diameters exceeding 4.0 in (102 mm). Numerous people were treated for injuries, some critical, and storm damage exceeded $2.0 billion (Edwards and Thompson 1998).

Despite the impact of these extremely large hail events, the definition of what constitutes

Corresponding author address:
Aaron W. Johnson
NOAA/National Weather Service,
104 Airport Rd. Dodge City, KS 67801
E-mail: aaron.johnson@noaa.gov

large hail is rather arbitrary and varies historically. Within the United States for example, the National Weather Service (NWS) has changed the definition of "severe" hail from 0.75 in (19 mm), to 1.0 in (25 mm) in recent years. Further, the NWS Storm Prediction Center (SPC) issues probability forecasts for "significant" hail (Hales 1988), defined with a diameter ≥2.0 in (51 mm), in addition to probability forecasts for the "severe" hail threshold. Nonetheless, despite varying criteria, the potential for economic and public safety issues arising from extremely large hail events necessitate discriminating these from smaller hail sizes.

When reviewing literature for specifics on hail-size prediction, most research typically focus on two paths of understanding. One primary area of concentration is on the storm-scale processes and microphysics controlling the growth of hail in deep, moist convection (e.g., Browning 1963, 1977; Browning and Foote 1976; Miller et al. 1988). In contrast, a second area of examination has centered more on operational forecast tools and radar techniques assisting with identification of large hail. However, goals for many of these studies overlap very little with the first area of research (e.g., Craven and Brooks 2004; Donavon and Jungbluth 2007; Blair et al. 2011).

Hail embryos have numerous source regions and subsequent trajectories within deep, moist convection (Browning 1963, 1977; Browning and Foote 1976; Miller et al. 1988). Nonetheless, Knight and English (1980), and Miller et al. (1988) found evidence in supercells that point toward larger hailstones containing embryos that originate close to the rotating updraft. Rotating updrafts introduce an unfair competition for supercooled water due to size sorting (Browning and Foote 1976; Knight and Knight 2001). This process allows larger hail growth and chiefly explains why most hail with diameters ≥2.0 in (51 mm) emanates from supercells (Rasmussen and Blanchard 1998; Thompson et al. 2003; Craven and Brooks 2004; Duda and Gallus 2010). However, if unfair competition exists in each supercell, why some struggle to produce 2.0-in (51-mm) hail while others produce hail ≥3.5 in (89 mm), is not explained solely through unfair competition.

Studies such as Marwitz (1972), Foote and Fankhauser (1973), and Browning (1977) have shown an inverse relationship between storm-precipitation efficiency and storm depth shear and/or storm-relative (SR) wind. Further, Rasmussen and Straka (1998) and Beatty et al. (2009) use this connection to provide an explanation regarding why some supercells favor a low-precipitation (LP) phase (Bluestein and Parks 1983; Bluestein and Woodall 1990) while others display a high-precipitation (HP) phase (Doswell and Burgess 1993). Knight and Knight (2001) indicate this same inverse relationship applies to the level of competition that influences maximum potential hail size within convection.

Predicting ranges of potential maximum diameter hail size in the pre-storm environment remains difficult. Parameter-based climatology studies such as Rasmussen and Blanchard (1998), Thompson et al. (2002a,b, 2003), and Craven and Brooks (2004) give little attention toward discrimination of hail size. Further, of the operational hail-related studies, most are largely radar-based (e.g. Donavon and Jungbluth 2007; Blair et al. 2011). Others have attempted hail-size prediction using various measures of CAPE and temperature level data, but with limited success (e.g., Fawbush and Miller 1953; Foster and Bates 1956; Miller 1972; Renick and Maxwell 1977; Moore and Pino 1990).

Hail-growth models such as HAILCAST (Jewell and Brimelow 2009) have shown a very promising path forward in hail-size predictive capabilities. However, some of the parameters used in hail-growth models are included based on assumptions that are not always true, such as hail embryo quantity. A lack of forecast-oriented studies focusing on variables that may provide insight to items such as hail embryo quantity or trajectories, only further compounds predictability problems.

To bridge this gap in understanding, this study builds a database of over 500 hail reports binned into one of four hail sizes. Creation of a parameter-based climatology of thermodynamic and wind-related parameters follows as detailed in section 2. Examining the physical relevance of a parameter or the relevant parts of a particular parameter space trails in sections 3–5. Finally, a summary follows in section 6.

2. Data and methodology

Building a representative yet high quality hail database in the U.S. is challenging, as few high-density hail-observer networks exist. Most hail-

based literature such as Blair et al. (2011), rely heavily on NCDC storm report database (*Storm Data)*. However, reliability questions exist with both size estimation and representativeness of maximum hail size falling within a storm as discussed by Schaefer et al. (2004), and Doswell et al. (2005).

An informal study by Baumgardt (2014) reveals spotters are more likely to report 1.25-in (32-mm) and 1.5-in (38-mm) hail to the commonly-sized objects of a quarter (1.0 in, 25 mm) and golfball (1.75 in, 45 mm). Jewell and Brimelow (2009) display this same bias with a relative minimum in 1.25-in (32-mm) and 1.5-in (38-mm) reports compared to the two common-size object reports. Some of these issues are inherently due to NWS verification practices related to the issuance thresholds for warnings (1.0 in, 25 mm). Specifically, as noted by Amburn and Wolf (1997) there is an associated bias toward these values rather than potentially larger hail.

Recent projects such as the NSSL Severe Hail Verification Experiment (SHAVE) have attempted to mitigate some of these issues (Ortega et al. 2009). However, with limited resources SHAVE functions over a restricted timeframe and geographic region compared to the temporal and spatial distribution of hail reports (Blair and Leighton 2012). Cintineo et al. (2012) also try to remedy some of these biases by using a radar algorithmic-based climatology to estimate hail sizes. However, this process primarily uses the maximum expected hail-size algorithm that Wilson et al. (2009) display as having poor skill with one-to-one hail-size prediction. Given these limitations, *Storm Data* remains the only database with an extensive list of hail reports nominally suitable for research applications.

In addition, an inconsistency in the actual size of softball-size hail is readily apparent in hail reports over the years. Jewell and Brimelow (2009) discuss how the size of a softball has varied from 4.5 in. (114 mm) at the first-ever softball tournament played in 1933, to the modern size of 3.8 in. (97 mm) for a men's softball and 3.5 in. (89 mm) for women's. To account for this inconsistency, any binned hail thresholds including reports of softball-size hail in this study, will also include sizes as small as 3.5 in (89 mm) to minimize the impact of this uncertainty.

Prior to report acquisition and sorting, selection of hail-size bins was necessary for quality control purposes. A simple division in hail diameters starts with the 2.0-in (51-mm) threshold where reports of this size or larger emanate almost exclusively from supercells (e.g. Rasmussen and Blanchard 1998). However, as detailed in Blair et al. (2011), hail diameters ≥4.0 in (102 mm) have the potential for economic and public safety issues well beyond smaller hail sizes. In addition, hail sizes ≥6.0 in (152 mm) are not only rare but result in levels of property damage not found in other significant hail reports (Guyer and Ewald 2004; Blair and Leighton 2012). Based on this literature, reports ≥2.0-in (51 mm) logically sort into a 2.0–3.25-in (51–83-mm) group, 3.5–5.75-in (89–146-mm) range, and a ≥6.0-in (152-mm) hail bin.

Unlike ≥2.0-in (51-mm) cases, additional breakpoints for reports below this threshold are less obvious. A common practice in severe-convective literature is to group <2.0-in (25-mm) diameter hail into one range but typically with little reasoning provided (e.g. Jewell and Brimelow 2009). However, Donavon (2010) notes a difference in radar-based thresholds and potential environments that support 1.75-in (45-mm) hail compared to 1.0-in (25-mm) sizes. Further, due to estimation errors noted previously by Buamgardt (2014), 1.0-in (25-mm) and 1.25-in (32-mm) reports are inseparable in hail bin assignment with the same connected relationship existing with 1.5-in (38-mm) and 1.75-in (45-mm) reports. Given these realities, another logical break exists between a 0.75–1.25-in (19–32-mm) group and a 1.5–1.75-in (38–45-mm) range of values.

Numerous studies have researched the relationship between parameters derived from vertical profiles of observed soundings and convective storm behavior (e.g., Newton 1963; Weisman and Klemp 1982; Rasmussen and Blanchard 1998; Craven and Brooks 2004; Jewell and Brimelow 2009). However, the coarse spacing of the radiosonde network and infrequent launches relative to convective time scales introduce many questions of representativeness when choosing proximity soundings (Orlanski 1975; Brooks et al. 1994b). For this reason, a growing number of studies are relying on objective analysis data from operational numerical weather prediction (NWP) and reanalysis data sets to better delineate environmental features and parameters on a mesobeta or smaller scale (e.g.,

Brooks et al. 2003; Bunkers et al. 2006; Thompson et al. 2007).

One such system used in lieu of observed proximity soundings is the Rapid Update Cycle (RUC) model (Benjamin et al. 2004). The RUC was a numerical weather prediction system specializing in hourly objective analysis run operationally by the NOAA from 1994 until decommissioning in May 2012. Studies focusing on RUC objective analysis error such as Thompson et al. (2003) and Coniglio (2012) found that although small errors exist in the RUC, values were close enough to observed soundings to use in lieu of observed proximity soundings. Based on these studies along with the limitations of observed proximity soundings, this study explicitly uses RUC analysis for environmental data.

Throughout the operational lifespan of the RUC, the model underwent a wealth of changes to improve performance and capability. These changes include the original 60-km horizontal grid spacing switching to 40 km in 1998, 20 km in 2002, and finally 13 km in 2005. Numerous other assimilation techniques and physics packages occurred during these years with specifics available online through the RUC operational change logs (available at http://ruc.noaa.gov). To mitigate potentially unrepresentative analysis fields due to coarse horizontal grid spacing, use of RUC analysis was limited to only full years with 20 km and 13-km grid resolutions (2003–2011). While we acknowledge this resolution difference and other changes might influence the data, no noticeable impacts to data quality and diagnosis results occurred among any of these years.

The constraints placed on RUC analysis, limits construction of the database to all *Storm Data* hail reports from 2003–2011 across the contiguous United States (NCDC 2003–2011). Further, only reports emanating from locations east of the Rocky Mountains are included, in order to mitigate stronger orographic influences not easily resolved in proximity data. Starting with the ≥6.0-in (89-mm) hail-size bin, a cursory evaluation of reports for 2003–2011 directed the study toward ≥6.0-in (89-mm) reports in 2003–2004, 2007, and 2010–2011. However, this led to only eight reports in the ≥6.0-in (89-mm) hail range with use of this bin subsequently abandoned due to this limited sample size. This action caused a resorting of reports from this unused range into a single ≥3.5-in (89-mm) hail bin. All remaining raw hail reports then process through our quality-control steps.

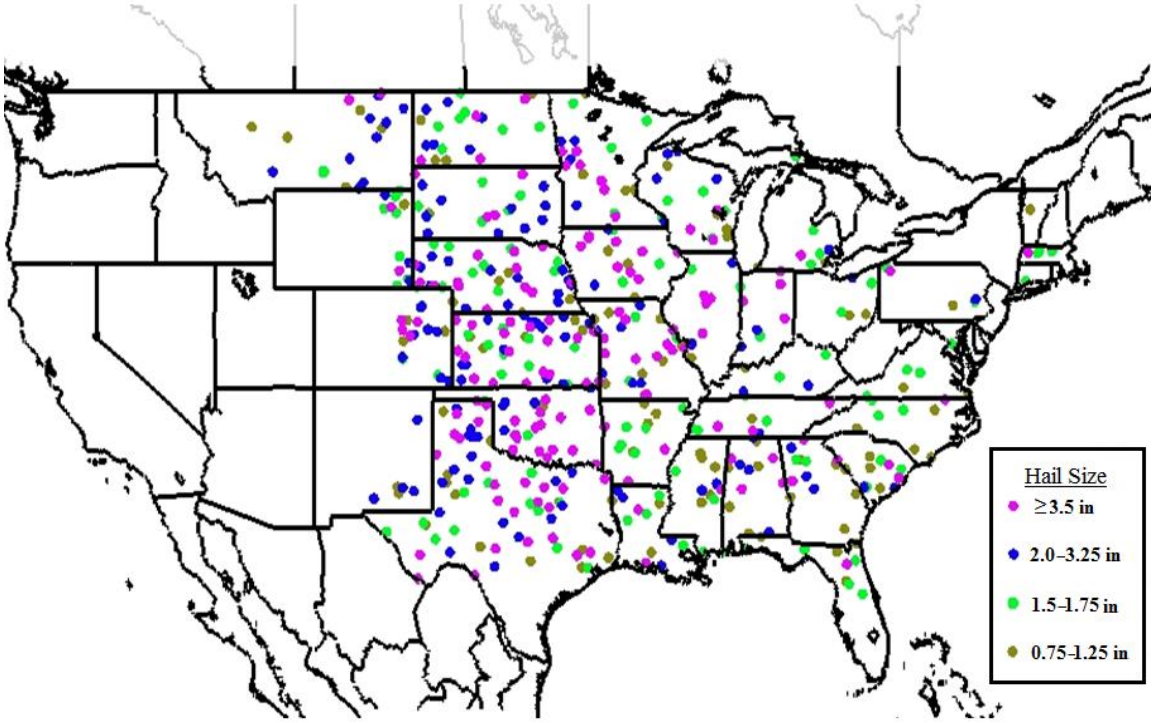

Figure 1: Hail reports from 2003, 2004, 2010, and 2011 used in this study plotted by location and size bin.

To avoid duplication and subsequent biasing of the database toward cases with numerous reports, only the largest hail report was initially included. Further, any subsequent hail reports of similar size were only included if they were beyond 6 h or 250 km from the other report. Identical conditions apply to inclusion of smaller hail except they must also pass the criteria against any of the larger hailstones removed in the previous step. The 6-h and 250-km exclusion thresholds follow similar criteria used in literature such as Thompson et al. (2007) but with values increased slightly to increase confidence in removal of duplicate reports.

Despite the initial quality control criteria removing numerous duplicate reports, concerns involving report inclusion at smaller hail sizes still existed. Specifically, as noted previously by Baumgardt (2014), spotters are likely to report hail size to the nearest common object rather than the actual hailstone diameter. Unlike the two largest hail diameter groups, the two smallest have a narrow range of values with common objects such as golf-ball hail (1.75 in, 45 mm) sitting close to the limits of both adjacent bins. Essentially, at smaller sizes the odds increase that we unknowingly could sort some hailstones into the wrong bin.

Based on these concerns, this study strengthened the inclusion threshold for the two smallest hail-size bins to maintain some level of uniqueness for these environments. Specifically, reports passing the initial criteria but occurring within 250 km and 12 h of the larger hail must be within one bin size of the largest hailstone or they are excluded from the database. This process restricts the additional inclusion criteria to ≤1.75-in (45-mm) reports falling into this temporal and spatial range. As with the initial criteria the 250-km threshold is increased slightly over previous studies (e.g., Thompson et al. 2007) but expanded even more temporally to 12 h to better isolate reports that only produce the smaller hail.

A final round of quality-control steps examines reports for questionable timing or location issues. These steps exclude reports with duplicate reports caused by an apparent time-logging error. An additional step excludes cases containing *Storm Data* comments that specifically question the reports' timing or location.

After all possible raw hail reports process through our quality-control steps, RUC data were obtained from the NCDC National Operational Model Archive & Distribution System website (http://nomads.ncdc.noaa.gov/). However, during the data-gathering phase of this study archived RUC analysis were missing or only partially available from 2005–2009. Despite these limitations, availability of RUC files for 2003–2004 and 2010–2011 leaves 520 individual hail reports. Sorting reports into the appropriate size bin, results in 152 in the ≥3.5-in (89-mm) range, 137 in the 2.0–3.25-in (51–83-mm) grouping, 116 in the 1.5–1.75-in (38–45-mm) range, and finally 115 in the 0.75–1.25-in (19–32-mm) collection (Fig. 1).

A change in sample size may affect the distribution of reports and statistical interpretation of our results for a selected point granted that only a few cases exist near any given location. However, while we would never argue against a larger sample size, this study was still able to create a sufficiently representative sample size consistent with other parameter-based climatology studies (e.g., Rasmussen and Blanchard 1998; Thompson et al. 2003; Bunkers et al. 2006).

The Java-based utility, Integrated Data Viewer (IDV)[1] was used to format and export RUC data for interrogation by additional software. RUC analysis for each report typically used a point closest to the report and the hour valid immediately before the report time (e.g., 21 UTC RUC analysis used for report occurring at 2120 UTC). However, this process also involved loading RUC surface winds, dewpoints, and surface-based (SB) CAPE in IDV to diagnose any representativeness issues with the analysis data. In <10% of cases, switching to a different grid point within 50 km of the report was necessary to sample the inflow sector.

Extraction of thermodynamic and wind-related variables using the RAwindsonde Observation Program (RAOB)[2] software resulted in over ninety variables. However, many of these were almost duplicates. Through elimination of near-duplicate parameters, 12 thermodynamic and

[1] Version 4.1, Unidata Program Center at the University Corporation for Atmospheric Research; available at http://www.unidata.ucar.edu/software/idv.

[2] Version 6.4, Environmental Research Services, LLC 2014; available at http://www.raob.com.

14 wind-related variables remained for evaluation. Further, the significant severe parameter (SSP; Craven and Brooks 2004) and new index called the large hail parameter (LHP) permit evaluation of parameter combinations.

Seven of the wind-related variables are SR in nature and require a storm motion for parameter calculation. Given the pre-storm focus of this study, observed storm motions were not obtained, but rather the "internal dynamics (ID) method" for predicting supercell motion (Bunkers et al. 2000) was used. As detailed by Thompson et al. (2007), this method may fail for some elevated events and be too deviant relative to the mean cloud-bearing winds for non-supercell cases. Nonetheless, the majority of the database reports still appear to be rooted in the boundary layer as 85% of cases studied have little to no difference in CAPE between an SB parcel and most-unstable (MU) parcel (highest θ_e value in the lowest 500 hPa). In addition, with 289 reports (hail ≥2.0 in, 51 mm) out of 520 total cases likely emanating from supercells (Rasmussen and Blanchard 1998) and numerous supercell cases contained within the 1.5–1.75-in (38–45-mm) range, the ID method provides the best approximation for SR parameters.

3. Results: Wind-related

a. Mean vertical wind structure

Initial examination of the vertical wind structure is broken down into ground-relative (GR) mean hodographs shown in Fig. 2 for each of the four hail-size bins. The three largest size categories are nearly indistinguishable in the lowest 3 km. Further, they only begin to show modest differences in magnitude and/or direction starting at 4 km AGL with the largest separation existing above 6 km AGL. Similarity in hodograph shape and curvature below 6 km for the larger hail sizes implies traditional wind-based supercell parameters focusing on the lowest 6 km may have little utility at distinguishing between larger hail reports.

Above 6 km AGL, only subtle differences exist between the 1.5–1.75 in (38–45 mm) and 2.0–3.25-in (51–83-mm) ranges as they display similar magnitude and unidirectional to slight backing of winds with height. Further, between reports ≥3.5 in (89 mm) and the two middle ranges of hail size, winds in the former category exhibit more of a slight veering profile with height along with greater magnitude above 8 km AGL. In contrast, the 0.75–1.25-in (19–32-mm) range displays noticeably weaker velocity at all levels compared to the three largest hail-size bins. This difference further supports the notion that most hail with a diameter ≥2.0 in (51 mm) requires a rotating updraft, while smaller hail does not (Rasmussen and Blanchard 1998; Thompson et al. 2003; Craven and Brooks 2004; Duda and Gallus 2010).

The inclination toward stronger anvil-level flow and a slight veering profile in ≥3.5-in (89-mm) reports is similar to the Bunkers et al. (2006) composite hodograph for longer-lived supercells. In contrast, the two middle ranges look similar to the short-lived storms with weaker anvil-level flow and a backing profile. In addition, Rasmussen and Straka (1998) evaluated supercell composite hodographs and found LP storms occurring in environments with stronger anvil-level flow similar to hail reports ≥3.5 in (89 mm). However, this same study also contains LP storms with unidirectional to backing winds above 7 km AGL similar to the two middle ranges, while winds were slightly veering for HP environments comparable to the largest hail-size bin. This mixed signal leaves it unclear from hodograph shape alone, whether hail reports ≥3.5 in (89 mm) occur in environments producing lower or higher precipitation efficiency supercells.

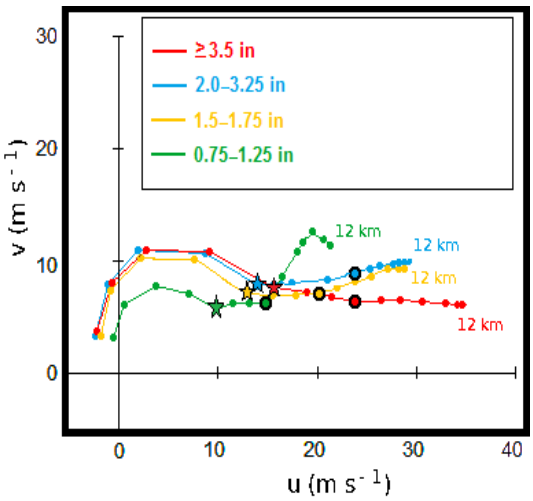

Figure 2: Composite 0–12-km AGL hodographs for each hail-size category. First three points represent surface, 0.5-km, and 1-km AGL winds with subsequent points at 1-km intervals through 12 km. Stars (black rings) on the hodograph represent winds at 3 km (6 km) AGL.

Subsequent sections will further refine some of these tendencies by examining interquartile spacing. Tests of statistical significance use the Student's t-test at a 95% confidence level to determine statistically significant differences in means (Wilks 1995). Unless otherwise noted, the results were statistically significant, often with p <0.0001.

b. Effective and 0–6-km bulk shear

Cloud-model simulations by Weisman and Klemp (1982, 1984, 1986) and Weisman and Rotunno (2000) have shown vertical wind shear over the lower half of a storm's depth as being critical in supporting supercell structures. The magnitude of the effective bulk vector shear (Thompson et al. 2007) and 0–6-km AGL bulk vector shear (hereafter $\text{Shear}_{\text{Eff}}$ and Shear_{6}, respectively) are two of the more popular estimates of shear in this layer. Evaluation of these parameters by binned hail size (Fig. 3) displays mixed results. Hail reports ≥2.0 in (51 mm) occupy a different portion of parameter space compared to the smallest hail-size bin. A value near 17 m s^{-1} is a critical lower limit for the two largest hail-size bins. Further, some level of skill exists with discriminating reports in the 1.5–1.75-in (38–45-mm) range compared to the smallest range at 17 m s^{-1}. However, more interquartile overlap exists with these two ranges compared to larger hail sizes.

In contrast, considerable overlap exists among the three largest hail groups with little separation between the two larger ranges. Specifically, the differences in the two largest hail-size bins are not statistically significant. Several parameter-based climatology studies (e.g., Rasmussen and Blanchard 1998; Thompson et al. 2003; Craven and Brooks 2004) have shown $\text{Shear}_{6\text{m}}$ and later $\text{Shear}_{\text{Eff}}$ (Thompson et al. 2007) to discriminate well between supercell and nonsupercell cases. The dominant convective mode for each report in this study is unknown, yet if hail with diameters ≥2.0 in (51 mm) emanate almost exclusively from supercells, then results shown in Fig. 3 may be the same signal with limited ability to discriminate larger potential hail sizes.

c. 0–1-km and 0–3-km SRH; 0–1-km bulk shear

Davies-Jones et al. (1990) and Davies (1993) revealed a slight relationship between increasing values of 0–3-km AGL SR helicity (hereafter SRH_3) and increasing tornado damage rating. Additional research involving this relationship

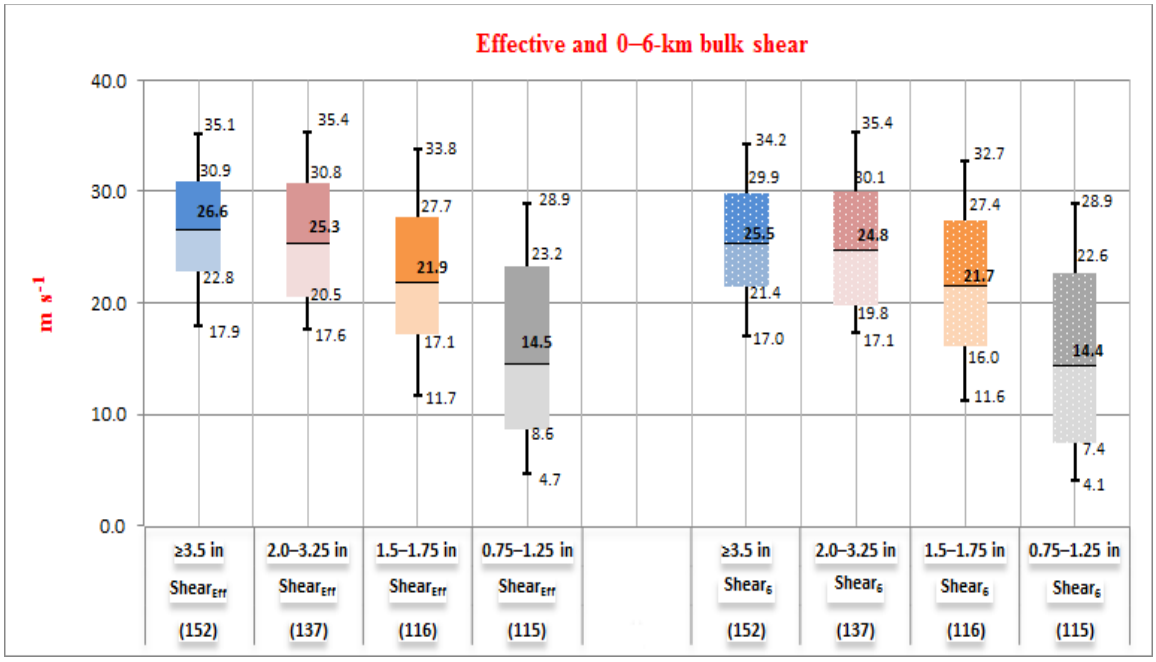

Figure 3: Box and whiskers plots of the magnitude of effective bulk shear (left) and 0–6-km bulk shear (right) for each binned hail size. Boxes denote 25th to 75th percentiles, with the line separating slight shading difference in each box representing the median value. Thin vertical lines (whiskers) extend to the 10th and 90th percentiles. Sample sizes given in parentheses.

from Rasmussen and Blanchard (1998) found SRH_3 values discriminate supercell versus nonsupercell environments rather than tornado damage rating. Nonetheless, studies such as Rasmussen (2003), Thompson et al. (2003), Craven and Brooks (2004), Miller (2006), and Esterheld and Giuliano (2008), evaluate shallower, near-ground shear layers with promising results. The magnitude of the 0–1-km AGL bulk vector shear and SR helicity (hereafter $Shear_1$ and SRH_1, respectively) were shown in these studies to discriminate well between tornadic and nontornadic supercell environments.

Examination of SRH_1 and $Shear_1$ (not shown) reveal little interquartile separation among the three largest binned hail sizes with any differences not statistically significant. Further, SRH_3 (not shown) does display some marginal difference in parameter space occupied between the two largest hail-size bins and the 0.75–1.25-in (19–32 mm) cases at a value near 150 $m^2 s^{-2}$. However, results display less interquartile separation than $Shear_6$ and $Shear_{Eff}$ among the three largest hail groups with differences between the two largest not even statistical significant. From a low-level environmental wind shear perspective, tornado and maximum expected hail size have little forecast overlap.

d. Upper-tropospheric shear and SR wind

Although development of supercell structures may be highly dependent on vertical wind shear in the lower half of a storm's depth, the composite hodographs in Fig. 2 indicate recognizably different structure in the upper half. The magnitude of storm-depth shear or SR wind above 6 km AGL has received considerably less literary attention compared to shallower, near-ground layers. Nonetheless, Bunkers et al. (2006) found the 0–8-km AGL bulk vector shear and 8-km SR winds were stronger with longer-lived supercells. Further, Rasmussen and Straka (1998) in evaluating supercell vertical wind structure, found the magnitude of the 0–9-km AGL and 4–10-km AGL bulk vector shear, along with 9–10-km SR flow, stronger in LP versus HP supercells. Beatty et al. (2009) also found stronger upper-level SR winds with forward reflectivity mode supercells.

In evaluating upper-tropospheric winds, equilibrium-level (EL) calculations use a 1.5-km non-pressure weighted mean wind. The top of this layer uses an EL height derived from an MU parcel with a virtual temperature correction applied in the calculation (Doswell and Rasmussen 1994). The mean wind approach used follows results from Rasmussen and Straka (1998). Specifically, they find the strongest signal at discriminating LP versus HP supercells occurs with upper-tropospheric flow typically 1–2-km below the tropopause. While this study also notes most supercells reach heights above the tropopause, the layer approach attempts to sample this zone of winds occurring below the EL. This process also permits sampling a broader range of winds near anvil height than a single level.

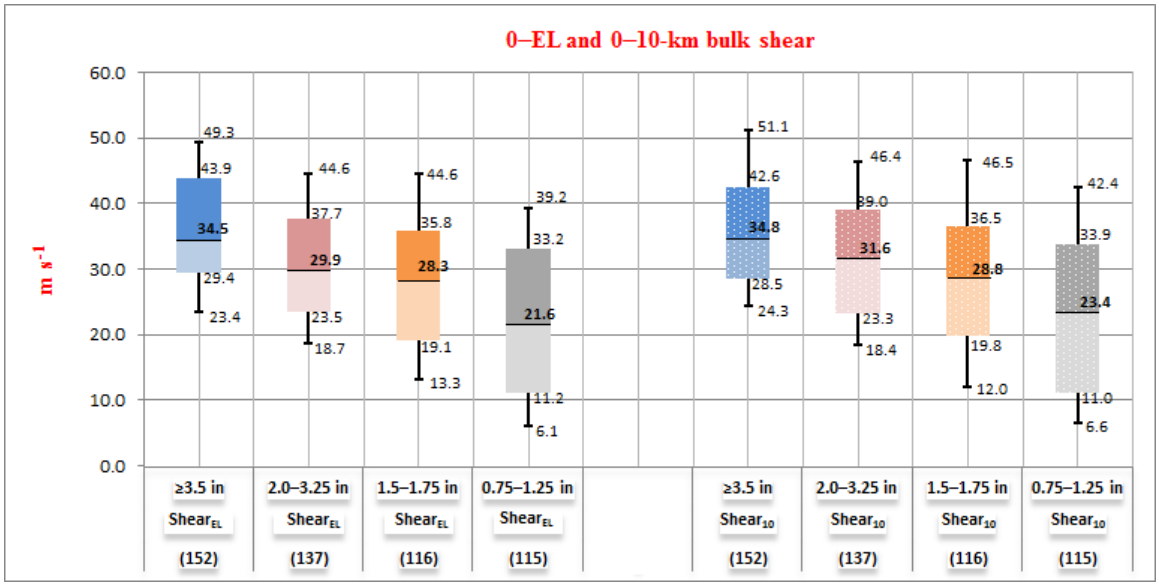

Figure 4: As in Fig. 3 except 0–EL and 0–10-km bulk shear.

Comparison of binned hail size to storm-depth shear for both a fixed layer 0–10-km AGL and variable depth 0–EL bulk shear (hereafter, $Shear_{10}$ and $Shear_{EL}$, respectively) are shown in Fig. 4. In addition, Fig. 5 displays the magnitude of the EL and 10-km AGL SR wind (hereafter, SRW_{EL} and SRW_{10} respectively). Upper tropospheric bulk shear using a 6 km–EL layer (hereafter $Shear_{6–EL}$) is calculated but not shown. Unlike $Shear_6$, $Shear_{10}$ and $Shear_{EL}$ display some interquartile separation at around 25–30 m s^{-1} between the two largest binned hail sizes with these differences statistically significant. However, more overlap exists among the three smallest binned hail sizes compared to $Shear_6$, especially with 0.75–1.25-in (19–32-mm) cases versus larger hail sizes.

Examination of $Shear_{6–EL}$ (not shown) along with SRW_{EL}, and SRW_{10} in Fig. 5 further display this relationship. Events ≥3.5 in (89 mm) partially occupy different parameter space starting at values around 8 m s^{-1} for $Shear_{6–EL}$ and 16 m s^{-1} for SRW_{EL} and SRW_{10}. In contrast, considerable overlap exists among the three smallest hail groups with any differences not statistically significant. These results suggest the sphere of influence of storm-depth shear or SR winds on maximum potential hail size may be limited to extremely large hail. Nonetheless, results for ≥3.5-in (89 mm) hail lends credence to the assertion made by Knight and Knight (2001) connecting lower beneficial competition in extremely large hail reports to stronger storm-depth shear.

e. Lower- and mid-tropospheric SR wind speed

Through analysis of over 250 observed proximity soundings, Maddox (1976) found tornadic storms to display similar lower-to-mid-tropospheric SR winds compared to nontornadic storms. Later research by Brooks et al. (1994a) and Beatty et al. (2009) discovered a modest relationship between stronger mid-tropospheric SR winds and supercells favoring an LP and forward reflectivity mode. Other studies such as Rasmussen and Straka (1998) found the opposite to be true while Bunkers et al. (2006) found little connection with either lower or mid-tropospheric SR winds and supercell longevity. Further, Thompson (1998) revealed midtropospheric SR winds stronger in tornadic supercells compared to nontornadic. However, research by Thompson et al. (2003) uncovered the difference in the mean mid-tropospheric SR winds was too small to make it a suitable parameter for discriminating significantly tornadic and nontornadic supercells.

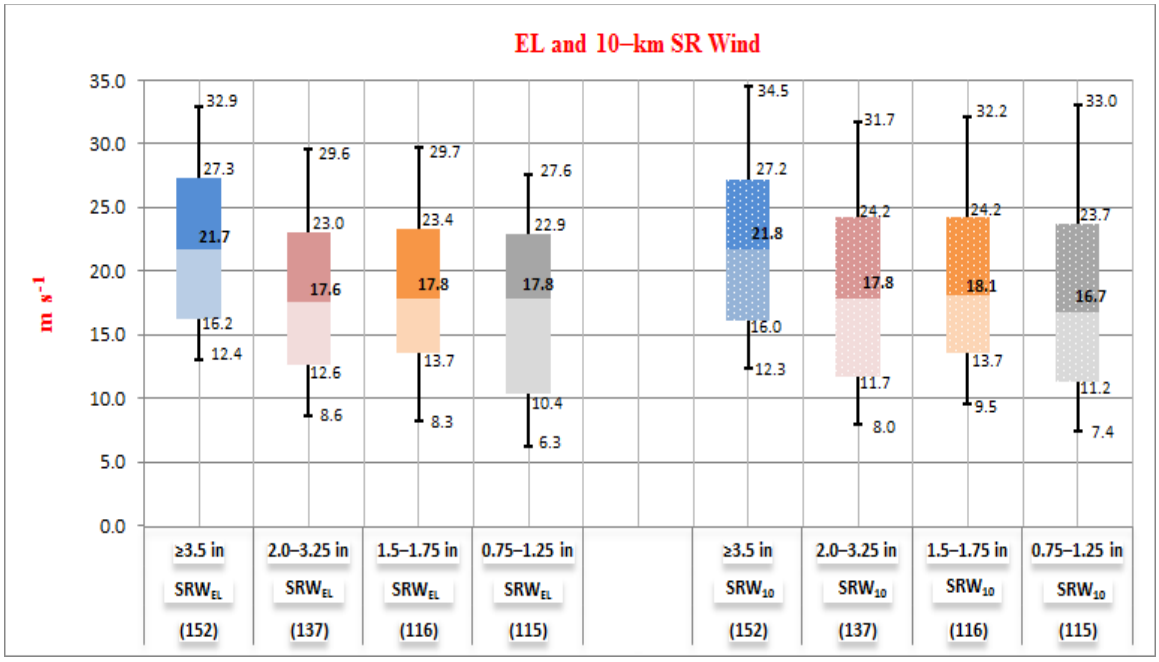

Figure 5: As in Fig. 3 except EL and 10-km SR wind magnitude.

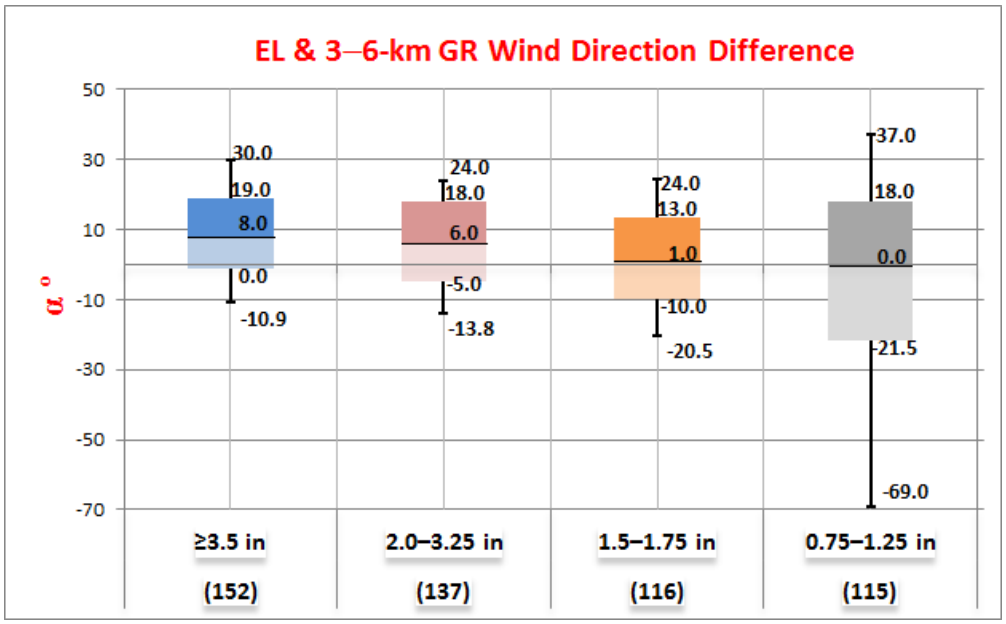

Figure 6: As in Fig. 3, but EL GR wind & 3–6-km GR wind direction difference.

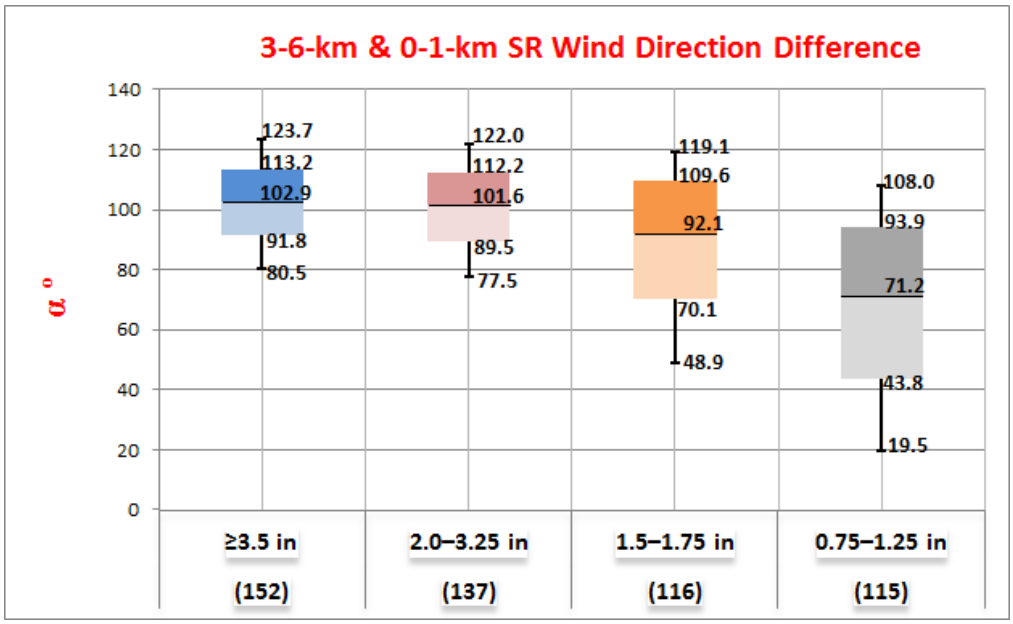

Figure 7: As in Fig. 3, but 3–6-km SR wind & 0–1-km SR wind direction difference.

The 0–1 km and 3–6-km SR wind magnitude (hereafter, SRW_1 and SRW_{3-6} respectively) are calculated but not shown. SRW_{3-6} reveals little, if any, parameter spacing among all bins. In contrast, a modest relationship exists in SRW_1 between the two largest hail-size bins and the 0.75–1.25-in (19–32-mm) cases as they occupy a different portion of the parameter spaceat around 11–12 m s^{-1}. However, extensive overlap exists among the three largest groups with no statistical significance existing between the two largest size ranges.

This relationship is not only similar to that shown by $Shear_6$ and $Shear_{Eff}$ at separating obvious supercell reports (≥2.0 in, 51 mm) from more mixed-mode cases (≤1.25 in, 32 mm), but also highly correlated (Table 1). It may initially

appear odd that SRW_1 is highly correlated to a shear value but near-surface SRW layers are highly influenced by storm motion. With the ID technique dependent on winds in the 0–6-km layer, a near duplication of wind layers used either directly with $Shear_6$ and $Shear_{Eff}$ or indirectly with SRW_1 creates similar results. Further, SRW_1 values occur in a narrow range well within the margin of error of NWP. Both of these latter relationships leave SRW_1 with limited potential as a hail-forecast tool.

f. Lower and upper tropospheric wind direction

As briefly discussed in section 3a, notable wind direction differences exist among the hail-size bins, especially above 6 km AGL. However, most studies looking at the vertical wind structure in relation to severe convection generally focus on parameters providing the magnitude of either the vertical wind shear or SR wind. Browning (1977) found that it may be critical for the wind direction difference, α, between the SR low-level inflow and mid-level flow, be at least perpendicular in order to deliver hail embryos to a location where they can be ingested into the updraft.

Figure 6 further examines α above 6 km AGL, by using a simple subtraction of the GR EL wind direction and GR 3–6-km wind direction (hereafter, GRWα_{EL}). Evaluating Fig. 6 from a parameter space perspective displays significant overlap among the four binned hail-size groups. However, similar to the tendencies seen in the composite hodographs, extremely large hail reports (≥3.5 in, 89 mm) display a propensity toward a slight veering wind profile (positive values) above 6 km. Although far from being a clear line of delineation between reports ≥3.5 in and smaller sizes, <25% of the former cases exhibit even a slight backing profile. Nonetheless, the relevance of this subtle relationship is unknown.

In testing the Browning (1977) conceptual model for hail embryo trajectories, Fig. 7 depicts α of the SR 3–6-km and 0–1-km wind (hereafter, SRWα_{Mid}). A modest amount of interquartile separation exists in SRWα_{Mid} between the two largest hail-size bins and the 0.75–1.25-in (19–32-mm) cases at α of 80–90˚. Further, this parameter spacing occurs in a larger range of values than SRW_1 but still with no statistical significance existing between the two largest hail-size ranges. In addition, much like SRW_1, duplicate wind layers used by SRWα_{Mid} and $Shear_6$ result in very similar skill in separating more obvious supercell cases from mixed-mode cases.

4. Results: Thermodynamic-related

a. MUCAPE and MLCAPE

CAPE is a measurement routinely used by operational forecasters to estimate environmental thermodynamic instability, where larger values correlate with the potential for greater updraft velocity. However, this method is very inaccurate at predicting maximum updraft velocity and hail size as discussed by Doswell and Markowski (2004). As detailed with EL height, computed CAPE uses a vertical thermal profile with a virtual temperature correction applied (Doswell and Rasmussen 1994).

As previously detailed in Section 2, the majority of cases studied have little to no difference between SBCAPE and MUCAPE. However, for the ≈15% of cases that occur with a relatively stable near-surface layer, MUCAPE provides some estimation of elevated instability encountered by these storms that is not well-resolved in a surface value. Nonetheless, rather than examining only the highest CAPE in the lowest 500 hPa without any consideration for entrainment of more stable air, evaluation of CAPE also includes a mixed-layer (ML) CAPE that uses a parcel with a uniformly mixed equivalent potential temperature in the lowest 50 hPa. The 50-hPa mean parcel layer was chosen over the commonly used 100-hPa depth (Craven et al. 2002b) to better represent shallow moisture layers (<500 m thick) in the RUC analysis that follow similar findings by Allen et al. (2011). When attempting to use a deeper mixed column, these shallow moisture layers commonly would mix with enough drier air to create little CAPE.

MUCAPE and MLCAPE in Fig. 8 display a tendency to occupy different parameter space as hail size increases, but still with considerable overlap similar to Jewell and Brimelow (2009) and Edwards and Thompson (1998). Further, values are slightly lower in comparison to Jewell and Brimelow (2009) despite all differences being statistically significant. This may be the result of Jewell and Brimelow (2009) exclusively using reports and associated soundings from the late afternoon to evening, that contain much of the daily peak heating thermal profile at the

surface. Further, their boundary layer corrective scheme also uses the maximum regional surface temperature and dewpoint value found in the inflow air but also results in the upper limit of expected values. Which of these potential bias issues or a combination results in this difference is unknown, yet a signal toward a MUCAPE or MLCAPE ≥2000 J kg^{-1}, is seen in hail reports ≥3.5 in (89 mm) compared to smaller sizes.

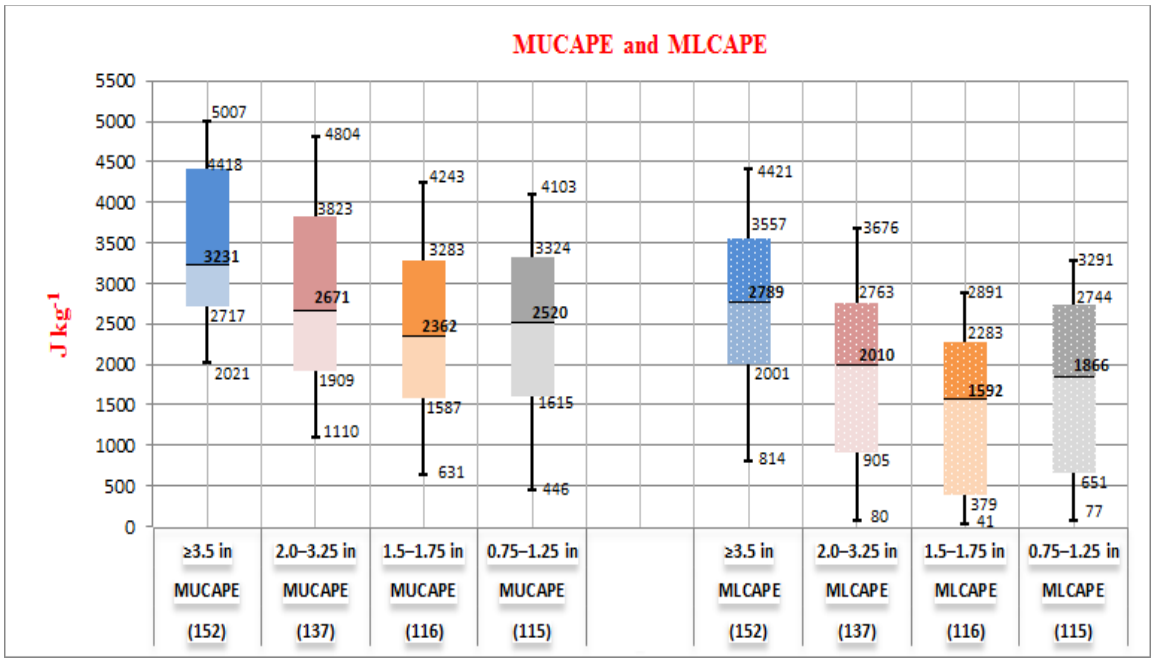

Figure 8: As in Fig. 3 except total MUCAPE and MLCAPE.

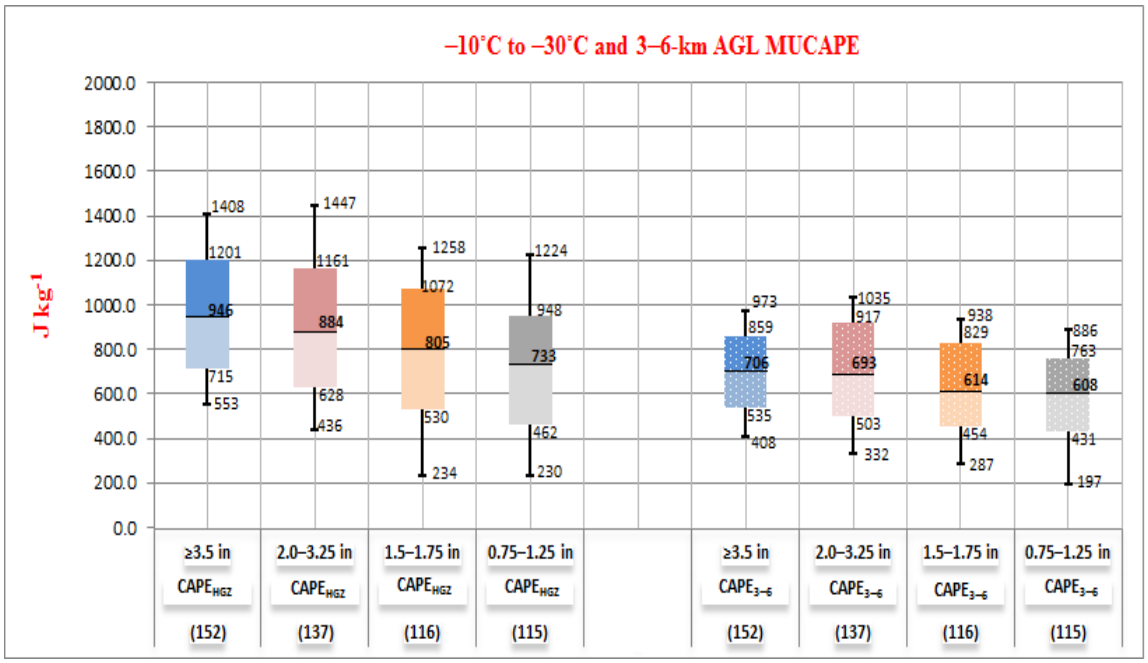

Figure 9: As in Fig. 3 except –10˚C to –30˚C and 3–6-km AGL MUCAPE.

b. Hail-growth-zone parameters

The hail-growth zone (HGZ) as defined by studies such as Nelson (1983), Foote (1984), Miller et al. (1988), and Knight and Knight (2001), occurs within a layer bound by –10˚C and –30˚C. Most forecast techniques looking at this layer are generally radar-based applications such as from Blair et al. (2011). Nonetheless, a few studies such as Fawbush and Miller (1953), Miller (1972), and Moore and Pino (1990), evaluate various measures of instability generally below this layer as predictors for hail size. However, techniques developed from this literature display marginal skill at forecasting maximum hail size with results frequently overestimating diameters (Doswell et al. 1982).

Figure 9 investigates instability within the HGZ and in a region typically just below by looking at MUCAPE in the –10˚C to –30˚C and 3–6-km AGL layers (hereafter, $CAPE_{HGZ}$ and $CAPE_{3–6}$ respectively). $CAPE_{3–6}$ depicts little interquartile separation among the groups although only the two largest hail-size bins have differences that fail in statistically significance. $CAPE_{HGZ}$ displays much of the same statistical significance and slight tendency to occupy different parameter space as hail size increases but still with substantial overlap. Further, comparing $CAPE_{HGZ}$ to MUCAPE and MLCAPE reveals the former displaying less interquartile separation than the total CAPE parameters.

A method to view this relationship is to compare the percentage of total MUCAPE existing only in the HGZ (hereafter, $\%CAPE_{HGZ}$) against the four hail-size groups (Fig. 10). Although little signal is shown for 0.75–1.25-in (19–32-mm) reports versus larger sizes, a noticeable tendency is seen when looking at reports in the three largest hail-size bins. Specifically, comparing the three largest groups against each other reveals an inverse relationship with a lower percentage of total MUCAPE residing within the HGZ as hail sizes become larger. Another method to examine this finding is to compare the HGZ thickness (hereafter, THK_{HGZ}) to the hail-size bins. Figure 11 reveals a similar inverse tendency in THK_{HGZ} versus hail size, especially with ≥3.5-in (89-mm) diameter hail. In addition, the differences in the means among the three largest hail-size bins all are statistically significant for THK_{HGZ}.

c. Lapse rates

Craven and Brooks (2004) have shown that significant severe weather events have steeper 700–500-hPa lapse rates (hereafter, $LR_{7–5}$) than marginal severe or non-severe events. However, that study combines hail and wind reports into one category, with any hail signal possibly obscured by wind reports. Figure 12 in this study provides a direct comparison of binned hail size against $LR_{7–5.}$

$LR_{7–5}$ displays considerable overlap similar to $CAPE_{HGZ,}$ and to a lesser degree, other thermodynamic-related parameters discussed previously. Nonetheless, a tendency toward steeper lapse rates exists between the smallest binned hail sizes and the two largest at around 6.5–7.0˚C km^{-1}. However, little interquartile spacing exists between the two larger hail sizes or between the two smaller groups with differences in either pairing failing in statistical significance. Examining lapse rates in layers above 700–500 hPa (not shown) reveal similar results. The 500–300-hPa and HGZ lapse rates (hereafter, $LR_{5–3}$ and LR_{HGZ} respectively), display a similar tendency toward slightly steeper lapse rates as binned hail size increases but with only minor interquartile separation.

d. Significant height levels

Rasmussen and Blanchard (1998), Markowski et al. (2002), Thompson et al. (2003), and Craven and Brooks (2004) found the lifted condensation level (LCL) height discriminates well between significant tornadic supercell events and those that are only weakly tornadic or nontornadic. However, aside from the significant tornadic cases, these studies also reveal that the LCL height has little additional utility among all other groups, including little skill at differentiating severe and non-severe environments. Evaluation of MULCL in this study (not shown) reveals little, if any, interquartile spacing with any differences failing in statistical significance.

Environmental freezing level (FZL) and wet-bulb zero (WBZ) heights have become common tools when combined with radar for hail prediction (e.g., Donavon and Jungbluth 2007). Nonetheless, these techniques provide limited value prior to storm formation. Miller (1972) attempted to use environmental FZL information with various measures of buoyancy to create a

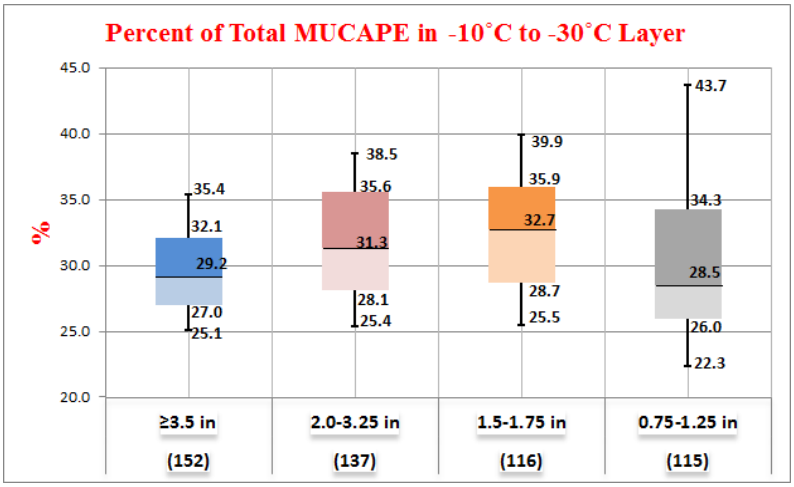

Figure 10: As in Fig. 3 except percent of total MUCAPE in HGZ.

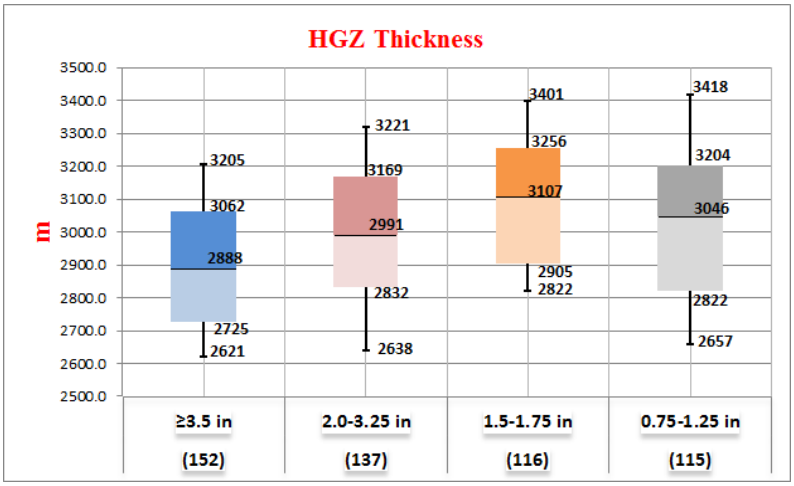

Figure 11: As in Fig. 3 except HGZ thickness.

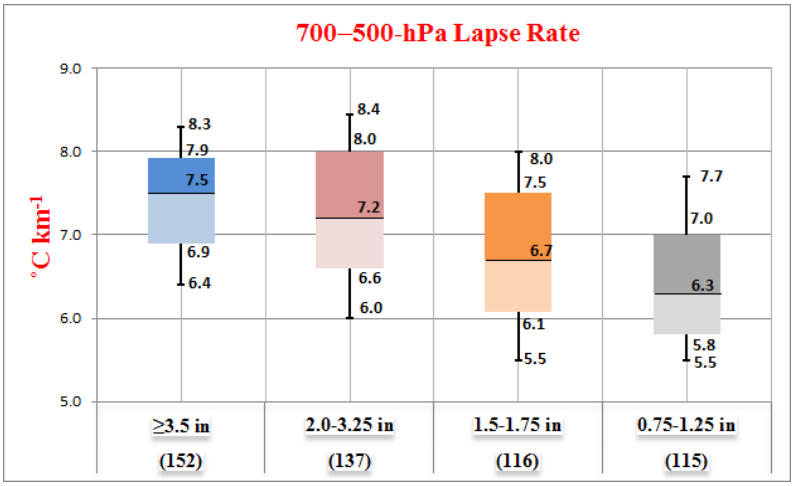

Figure 12: As in Fig. 3 except 700–500-hPa lapse rate.

predictive tool for hail size. However, application of these techniques typically reveal very limited success, and attempting to use FZL and/or WBZ alone as a predictive tool results in even less skill (Kitzmiller and Briedenbach 1993; Edwards and Thompson 1998). Examining FZL and WBZ heights against the hail-size bins (not shown) only further reinforce results from these earlier studies. Specifically, little interquartile separation exists between the hail groups with any differences failing in statistical significance.

5. Results: Parameter combinations

a. Significant severe parameter (SSP)

Various combinations of instability and shear have been created over the years to assist in forecasting severe-convective potential. Popular parameter combinations such as the energy-helicity index (EHI; Hart and Korotky 1991; Davies 1993), vorticity generation parameter (VGP; Ramussen and Wilhelmson 1983), and both the supercell composite parameter and significant tornado parameter (SCP and STP, respectively; Thompson et al. 2002a,b, 2003) have focused on varying degrees of supercell and/or tornado prediction. However, most of these parameters exist with little, if any, literary attention given toward hail prediction although a few exceptions exist.

Craven and Brooks (2004) evaluate SSP against significant tornado events, significant and non-signficant wind and/or hail cases, and general- or no-thunder cases with the index formula following:

$$SSP = (MLCAPE\ J\ kg^{-1})*(Shear_6\ m\ s^{-1})$$

As the product of MLCAPE and $Shear_6$, Craven and Brooks (2004) SSP displays some skill at distinguishing between cases with ≥2.0-in (51-mm) hail and/or ≥65-kt (33.4-m s^{-1}) winds versus reports with smaller hail and/or lesser wind. However, hail-only categories were not constructed. This leaves any potential hail signal blurred with wind reports that may have little significant hail.

A similar CAPE-shear combination called the energy shear index (ESI; Brimelow et al. 2002) is part of the HAILCAST model algorithm (Jewell and Brimelow 2009), and is the product of SBCAPE and magnitude of the 850 hPa–6-km AGL bulk vector shear. However, neither of these studies evaluates ESI directly against binned hail sizes but rather as a measure of updraft duration in HAILCAST.

Yet another CAPE-shear combination called the significant hail parameter (SHIP; SPC 2014) is an index developed in-house at the SPC. Unlike SSP and ESI, SHIP includes more than just CAPE and shear. Three additional variables are part of the formula and include the mixing ratio of a MU parcel, the $LR_{7-5,}$ and 500-hPa temperature that appear to help SHIP delineate between ≥2.0-in (51-mm) hail and smaller sizes. While the SHIP formula specifically targets hail discrimination rather than other severe-convective elements, evaluation of SHIP has not undergone a prior literary review with performance against multiple binned hail sizes unavailable. Further, a skewing of results may exist (SPC 2014) by the removal of all 1.75–2.0-in (45–51-mm) reports in order to magnify interquartile separation.

Not all parameters needed for calculation of ESI and SHIP are examined in this study. However, those needed for SSP are available with results following below. Unlike the Craven and Brooks (2004) study blurring hail and wind reports together, SSP results for hail only cases (Fig. 13) reveal overall better interquartile separation among the four original binned hail sizes (left side of Fig. 13) than any of the individual variables. The best tendency in interquartile separation among the four groups exists between two smallest ranges and the largest hail-size bin at a value of 40 000–45 000 $m^3\ s^{-3}$. On the other hand, marginal severe cases ≤1.25 in (32 mm), and reports deemed "significant" from other literature in the 2.0–3.25-in (51–83-mm) range, display a more modest level of interquartile separation, with little improvement over $Shear_6$ or $Shear_{Eff}$. Further, similar to $Shear_6$ and MLCAPE, considerable overlap exists between the two larger hail-size bins although unlike $Shear_6$ differences are statistically significant.

In order to make a more direct comparison to the figures given by SPC (2014) for the SHIP index, two additional hail-size bins are included on the right side of Fig. 13 comparing reports ≤1.25 in (32 mm) with those ≥2.0 in (51 mm). Although with slightly more overlap than SHIP, SSP displays very similar results with good interquartile separation among ≤1.25-in (32 mm)

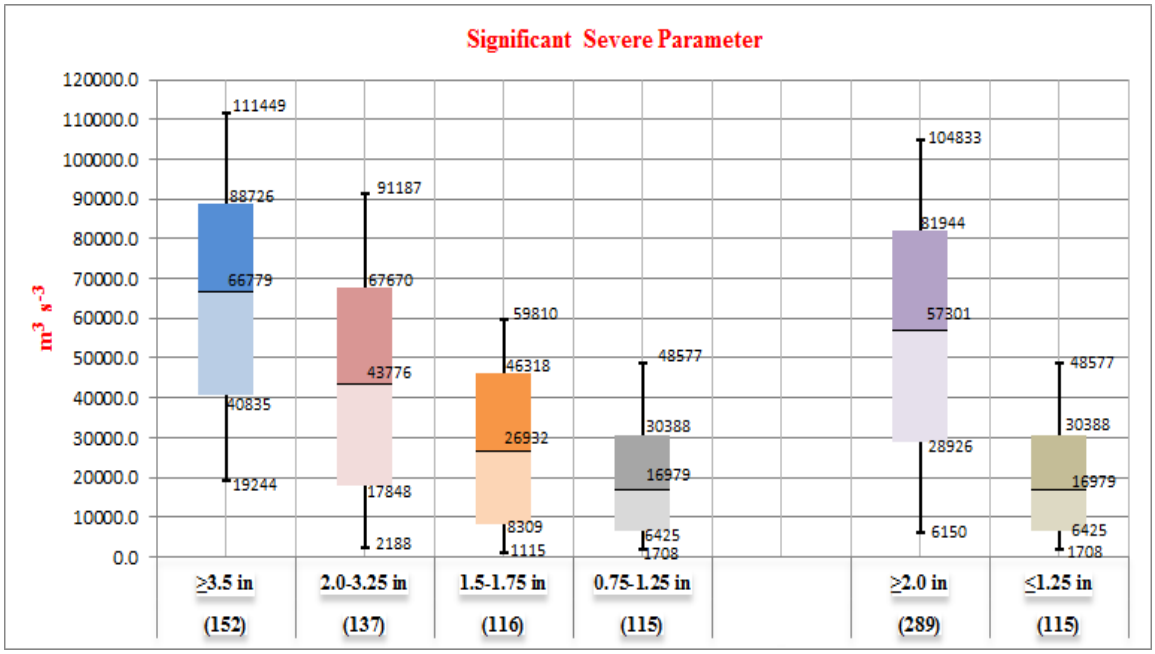

Figure 13: As in Fig. 3 except SSP and inclusion of two extra binned hail sizes.

reports and ≥2.0-in (51-mm) cases. The nearly identical results should not be surprising since both SHIP and SSP use $Shear_6$ and some variation of CAPE. However, it becomes apparent when reviewing the left side of Fig. 13 that most of the interquartile spacing existingwith SSP is not from delineating environments producing ≥2.0-in (51-mm) hail versus smaller sizes, but only from environments that will produce ≥3.5-in (89-mm) hail. The fact that the 2.0–3.25-in (51–83-mm) range still modestly overlaps the two adjacent hail-size bins along with nearly 50 percent of cases in the smallest hail range raises questions on the validity of forecasting ≥2.0-in (51-mm) hail from smaller sizes using SSP or SHIP. The question then becomes whether adding any additional parameters from sections 3 and 4 to a simple CAPE-shear combination would improve interquartile separation and subsequent predictive capability.

b. Large hail parameter (LHP)

One of the first steps in evaluating additional parameter combinations is selecting from those variables displaying at least some interquartile separation, as discussed in previous sections. However, these same parameters must also have strong independence (low correlation coefficient) among the variables included, to avoid overly strong influence from just one set of related parameters. Table 1 displays a correlation matrix for most of the variables listed in sections 3–4.

As expected, wind-related variables in Table 1 have strong independence with the thermodynamic-related variables. Given this reality, use of any parameter from one of these groups depends more on selecting those variables displaying at least some interquartile separation along with strong independence among related variables. Most of the shear and storm-relative wind variables have a strong interdependence (higher correlation coefficient) among each other, especially when comparing layers of similar depth (e.g., 0.94 for $Shear_{Eff}$ and $Shear_6$). Further, GRWα_{EL} and to a lesser degree SRWα_{Mid} display independence in comparison to most other wind variables while also exhibiting some notable parameter spacing differences as discussed in section 3.

Among the thermodynamic parameters displaying strong interquartile separation, most of the MUCAPE variables have high interdependence. This relationship only further reinforces results from section 4 indicating measurements of CAPE in or near the HGZ provide little improvement over CAPE. In contrast, MUCAPE has a strong independence with 1) $LR_{7–5}$ and 2) unconventional measures such as THK_{HGZ}. Further, combinations of these thermodynamic indices with the aforementioned wind parameters may yield some improvement over a simple CAPE-shear combination such as SSP.

Table 1: Correlation matrix for 17 of the 26 parameters evaluated in the present study. The digits across the top row correspond to the numbered fields on the side. See sections 3 and 4 for a definition of the parameters.

	1	2	3	4	5	5	7	8	9	10	11	12	13	14	15	16	17
1) $Shear_{6}$	1.0																
2) $Shear_{Eff}$	0.94	1.0															
3) SRH_{1}	0.50	0.50	1.0														
4) SRH_{3}	0.50	0.50	0.87	1.0													
5) $Shear_{10}$	0.77	0.75	0.31	0.33	1.0												
6) $Shear_{EL}$	0.75	0.73	0.32	0.33	0.89	1.0											
7) SRW_{10}	0.38	0.37	0.01	0.04	0.84	0.69	1.0										
8) SRW_{EL}	0.30	0.29	–0.02	0.00	0.67	0.80	0.80	1.0									
9) $SRW\alpha_{Mid}$	0.62	0.63	0.33	0.45	0.51	0.51	0.26	0.22	1.0								
10) $GRW\alpha_{EL}$	0.21	0.21	0.19	0.17	0.20	0.25	0.00	0.02	–0.01	1.0							
11) MUCAPE	–0.28	–0.19	–0.12	–0.13	–0.28	–0.23	–0.20	–0.11	–0.12	–0.06	1.0						
12) MLCAPE	–0.26	–0.18	–0.17	–0.16	–0.25	–0.20	–0.18	–0.11	–0.13	–0.03	0.88	1.0					
13) $CAPE_{HGZ}$	–0.20	–0.12	–0.12	–0.11	–0.18	–0.13	–0.09	–0.01	–0.04	–0.06	0.94	0.82	1.0				
14) $CAPE_{3–6}$	–0.22	–0.15	–0.16	–0.18	–0.18	–0.13	–0.08	0.00	–0.08	–0.07	0.87	0.76	0.92	1.0			
15) $\%CAPE_{HGZ}$	0.24	0.22	0.09	0.12	0.31	0.25	0.28	0.19	0.19	–0.03	–0.36	–0.34	–0.16	–0.19	1.0		
16) $LR_{7–5}$	0.24	0.26	0.10	0.18	0.26	0.29	0.19	0.21	0.29	0.12	0.19	0.20	0.34	0.28	0.25	1.0	
17) THK_{HGZ}	–0.03	–0.04	–0.04	–0.05	0.00	–0.06	0.06	–0.02	–0.04	–0.10	–0.10	–0.14	0.06	0.04	0.25	–0.10	1.0

Before discussing creation of a new parameter (LHP), a review on the limitations of parameter combinations is necessary. Doswell and Schultz (2006) address how most indices are diagnostic rather than prognostic in nature. In particular, these indices only identify the current state of the atmosphere with no established predictive quality of future weather. Like many other composite parameters, LHP and SSP are not exempt from being diagnostic. With no established predictive quality outside of an NWP-derived value, LHP and SSP cannot make predictions of future weather without regard to the state of the atmosphere at that forthcoming time (Doswell and Schultz 2006). However, by understanding these limitations an observed or NWP-derived LHP, along with deeper analysis, can provide a full picture of potential hail sizes.

To evaluate performance of including the aforementioned variables in a CAPE-shear combination, construction of a new parameter is required. The unitless LHP includes CAPE and shear but also a combination of other thermodynamic and wind-related variables displaying some predictive skill. The LHP formula is as follows:

If Shear_6 magnitude < 14 m s^{-1} OR MUCAPE < 400 J kg^{-1}:

LHP = 0

If Shear_6 magnitude ≥ 14 m s^{-1} AND MUCAPE ≥ 400 J kg^{-1}:

LHP = (**Term A** * **Term B**) **+ 5**

Term A = (((MUCAPE – 2000)/ 1000) +
((3200 – THK_{HGZ})/500) +
((LR_{7-5} – 6.5)/2))

Term B = ((Shear_{EL} – 25)/5) +
(($\text{GRW}\alpha_{\text{EL}}$ + 5)/20) +
(($\text{SRW}\alpha_{\text{Mid}}$ – 80)/10))

An explanation of each portion of the parameter is necessary to understand reasoning behind the formula. Unlike most traditional CAPE-shear combinations using either Shear_6 or $\text{Shear}_{\text{Eff}}$ as a CAPE multiplier, results from section 3b indicate Shear_6 is better suited as a simple toggle for supercell potential and subsequent threat of ≥2.0 in (51 mm) hail. Use of 14 m s^{-1} as a lower limit correlates to the median value for 0.75–1.25-in (19–32-mm) reports and is slightly below the 10th and 25th percentile values for the three largest hail-size bins shown in Fig. 4. This lower limit not only provides a small buffer below the ≈17-m s^{-1} threshold for the three largest hail-size bins, but also fits well with the Thompson et al. (2002a,b, 2003) findings pointing to most supercell cases with values of Shear_6 ≥15–20 m s^{-1}.

Given this Shear_6 limit, the formula checks if values are ≥14 m s^{-1} and if false, the LHP sets to zero with no further calculations made. Further, to avoid unnecessarily high LHP values in weakly unstable environments, the formula also checks if MUCAPE values are ≥400 J kg^{-1} and if false, LHP sets to zero. A MUCAPE of 400 J kg^{-1} was chosen as a lower limit since this correlates to the 10th percentile value for 0.75–1.25-in (19–32-mm) reports shown in Fig. 8 and is slightly below the 5th percentile for the two largest hail-size bins (not shown).

If both of these checks pass then five new variables along with MUCAPE create the LHP. These variables are broken down into a thermodynamic and wind-related component. The product of these two elements along with a small addition term, create the parameter value. The thermodynamic component (Term A) is composed of MUCAPE, THK_{HGZ}, and LR_{7-5}. Each of these variables displays modest interquartile separation and strong independence among each other. The wind-related component (Term B) is composed of a Shear_{EL}, $\text{GRW}\alpha_{\text{EL}}$, and finally $\text{SRW}\alpha_{\text{Mid}}$. If $\text{GRW}\alpha_{\text{EL}}$ is >180°, this portion of Term B is set to –10 as to avoid incorrectly labeling a small backing wind profile as strongly veering. Further, a value of 5 added to the end of the LHP formula creates more separation for index values near zero that were derived after the initial Shear_6 check passes versus leaving LHP at zero when the initial checks fail. Additionally, if Term A and B are both negative the LHP is set to zero to avoid creating a positive value by multiplying two negative terms. Finally, if values are still not above zero after the previous addition term, all negative values of the LHP are set to zero.

Evaluation of LHP in Fig. 14 reveals improvement over SSP, especially between hail reports in the 2.0–3.25-in (51–83-mm) range versus the smallest hail-size bin. At a value of 4–6, these two ranges now appear to occupy a different portion of the parameter space in contrast to SSP. Further, improvement also exists over SSP between the two smallest groups and the largest hail-size bin at an LHP value of 7–8.

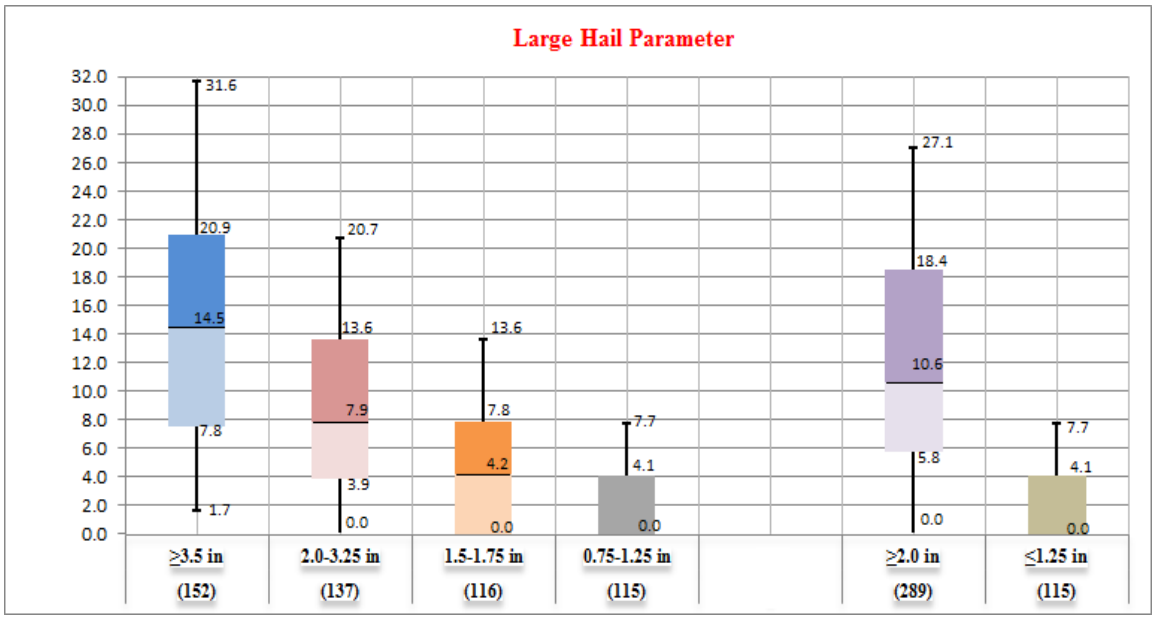

Figure 14: As in Fig. 3 except LHP and inclusion of two extra binned hail sizes.

Examining the right side of Fig. 14 for reports also depicts improvement in interquartile spacing of ≥2.0-in (51-mm) versus ≤1.25-in (32-mm) cases over SSP. Nonetheless, with only minor improvement among the three larger hail sizes, the increased parameter spacing in LHP between the two smallest groups and the largest hail-size bin is likely coming through better delineation of the 2.0–3.25 in (51–83-mm) and 0.75–1.25-in (19–32-mm) ranges mentioned earlier. The use of $Shear_6$ as a simple toggle for larger hail potential rather than as a CAPE multiplier appears to improve hail predictive capability via this two-step process.

Analysis of parameters so far in this study provides only a perceived skill by examination of interquartile spacing in box and whisker plots. To bridge this evaluation gap, Table 2 displays the Heidke's skill score for selected parameters (HSS; Doswell et al. 1990). An HSS value of 1.0 is a "perfect" forecast while values <0.2 imply little skill. To provide a more direct comparison of the tendencies discussed in previous sections, HSS results in Table 2 are broken into separate categories.

When including all reports in calculation of HSS for ≥2.0-in (51 mm) hail versus smaller sizes (column a), the LHP and SSP exhibit the highest skill. However, SSP displays only slightly better skill over some individual variables including $Shear_6$ and $SRW\alpha_{Mid}$. Further, MUCAPE reveals a relatively low HSS and given similar skill between SSP and $Shear_6$, CAPE apparently adds little predictive skill at this range of hail sizes.

As intriguing as LHP skill initially appears in this size range, skill at such a sharp delineation in hail sizes is misleading. By examining the other columns (b–d), skill gained in separating ≥2.0-in (51 mm) hail from smaller sizes is not by differentiating the two middle hail-size groups (column b). On the other hand, given the parameter-space overlap between these binned hail sizes, simply replacing 1.5–1.75-in (38–45-mm) reports with 0.75–1.25-in (19–32-mm) cases (column c) allows skill to increase for some parameters. LHP, $Shear_6$ and $SRW\alpha_{Mid}$ display the highest HSS while other parameters including SSP are lower in skill.

Despite the lack of predictive skill at such narrow ranges of typical supercell hail sizes, the interquartile separation shown between reports 1.5–1.75 in (38–45 mm) and those ≥3.5 in (89 mm), still hints at some skill existing, but at a broader range of hail sizes (Figs. 13 and 14). By excluding the other two hail-size bins, LHP and SSP skill (column d) increase slightly at this larger delineating size in comparison to the smaller size comparison (column c). However, skill of the individual parameters reverses roles in many cases. Further, unlike all other columns, MUCAPE along with non-traditional parameters such as the THK_{HGZ} and $GRW\alpha_{EL}$ increase to a value indicating at least slight skill.

Table 2: a) Parameters ranked according to Heidke's skill score for all reports based on the threshold value given for ≥2.0 in (51 mm) hail size prediction. b) Same as (a) but with the largest and smallest hail size reports excluded. c) Same as (a) but with the largest bin and 1.75–2.0-in (45–51-mm) reports excluded. d) Same as (a) but with the smallest bin and 2.0–3.25-in (51–83-mm) reports excluded and based on the threshold value given for ≥3.5-in (89-mm) hail-size prediction.

		(a)	(b)	(c)	(d)
		All Reports	**Excluded**	**Excluded**	**Excluded**
			0.75–1.25 in	1.5–1.75 in	2.0–3.25 in
			≥3.5 in	≥3.5 in	0.75–1.25 in
Parameter	**Threshold Value**	**HSS**	**HSS**	**HSS**	**HSS**
	≥2.0 in | ≥3.5 in	≥2.0 in	≥2.0 in	≥2.0 in	≥3.5 in
LHP	5 | 7	0.47	0.26	0.49	0.52
SSP (m^3 s^{-3})	30,000 | 40,000	0.39	0.23	0.38	0.41
$Shear_6$ (m s^{-1})	20 | 22	0.32	0.19	0.41	0.24
$Shear_{EL}$ (m s^{-1})	24 | 29	0.28	0.11	0.27	0.30
MUCAPE (J kg^{-1})	1,850 | 2,700	0.18	0.12	0.07	0.24
$LR_{7–5}$ (°C km^{-1})	6.5 | 7.0	0.33	0.22	0.32	0.35
THK_{HGZ} (m)	3,100 | 3,000	0.17	0.14	0.03	0.31
GRWα_{EL} (°)	–5 | 0	0.17	0.14	0.14	0.23
SRWα_{Mid} (°)	80 | 90	0.38	0.27	0.47	0.28

6. Conclusions and summary

A large-hail report database created in this study uses *Storm Data* reports from the years of 2003, 2004, 2010, and 2011. Through various quality-control measures, 520 remaining hail reports sort into one of four hail-size categories. From these cases, extraction of numerous RUC sounding-derived fields creates a parameter-based climatology of hail reports.

Initial examination of mean composite hodographs indicate a weaker vertical wind structure with the smallest hail-diameter range in comparison to the three largest bins. However, among the three largest hail-size bins, the vertical wind structure is nearly indistinguishable in the lowest 3–4 km but then displays differences above 6 km AGL. The ≥3.5-in (89-mm) cases reveal the most notable difference via a slightly more veering profile along with a greater wind magnitude.

Additional inspection of wind-related parameters simply reinforce tendencies shown in the mean hodograph, as lower shear layers such as $Shear_1$ or SRH_1 reveal little if any parameter spacing among all hail sizes. Surface-to-mid-level shear layers such as $Shear_6$ reveal modest skill at discriminating the smallest hail-size bin from the two largest, but with little ability to distinguish between the three largest hail-size bins. In contrast, much deeper shear such as $Shear_{EL}$, display less of a signal at the lower end of hail sizes yet better skill at delineating ≥3.5-in (89-mm) hail from smaller sizes. Although overall skill of GRWα_{EL} remains low, results indicate that for ≥3.5-in (89-mm) hail to occur, winds in the upper half of a storm's depth should veer or back only slightly with height. In the lower half of a storm's depth, SRWα_{Mid} veers by no less than 80–90° in most ≥2.0-in (51-mm) hail-size reports.

MUCAPE and MLCAPE along with $LR_{7–5}$ display a slight upward tendency in values and resultant skill as hail size increase but still with a considerable amount of overlap. Shallower layers of instability such as $CAPE_{HGZ}$ and $CAPE_{3–6}$ reveal less interquartile separation than total MUCAPE. Further, $\%CAPE_{HGZ}$ and THK_{HGZ} display an inverse relationship with hail size with values decreasing as diameters increase. Significant height levels such as FZL, WBZ, and LCL reveal little if any interquartile spacing or skill to discriminate among the hail-size bins. The poor performance of FZL and WBZ is similar to Edwards and Thompson (1998) and raises questions on the validity of using them in severe-convection forecasting despite common use in radar-based interrogation.

The new parameter combination LHP, and to a lesser degree SSP, show the strongest ability to discriminate between hail-size bins. When delineating ≥2.0-in (51-mm) hail cases versus

smaller sizes, LHP displays improved skill and interquartile separation over all individual parameters, including SSP. However, when comparing 2.0–3.25-in (51–83-mm) reports to the smallest hail-size bin, LHP and SSP skill is similar or lower than SRWα_{Mid} and Shear$_6$. In contrast, skill for SRWα_{Mid} and Shear$_6$ drop when comparing ≥3.5-in (89 mm) reports to 1.5–1.75-in (38–45-mm) cases, while SSP and LHP skill increase. An area of future work that may result in additional skill for the LHP is to switch from Shear$_6$ to Shear$_{Eff}$ while adjusting SRWα_{Mid} to use winds from this same effective layer. To compare successfully an effective layer LHP to the fixed layer version requires a refinement and subsequent binning of reports emanating from elevated convection. However, this level of convective detail is unknown in this study.

The implication of these results is that once the environment becomes favorable for rotating updrafts, the role of traditional supercell-based indices such as Shear$_6$ to dictate production of ≥3.5-in (89-mm) hail diminishes. In contrast, the role of items associated with storm-precipitation efficiency, updraft strength, and storm longevity such as Shear$_{EL,}$ THK$_{HGZ}$, GRWα_{EL}, and MUCAPE, increase. Essentially, once the supercell box toggles to "yes" results outlined suggest a forecast of maximum hail sizes at any diameter ≥1.5 in (38 mm) is within the bounds of reality. However, the next step involving various non-traditional parameter combinations assists with discriminating ≥3.5-in (89-mm) hail from smaller sizes.

The counter intuitive results of parameters in or near the HGZ suggest the depth of this layer becomes shallower for larger binned hail sizes but with no significant increase in layer instability. In addition, with CAPE increasing as hail size becomes larger, the lack of a similar change in instability within and below the HGZ imply higher CAPE is occurring above this layer. Although the exact mechanism resulting in this relationship is unknown, Knight and Knight (2001) discuss the impacts on hail growth by modification of storm-precipitation efficiency and resultant beneficial competition. This type of setting may produce a higher percentage of nascent hail embryos prematurely ejecting into the anvil due to inadequate growth time and higher parcel ascent above the HGZ. Conceptually, this would leave only a few remaining hail embryos with unfettered access to supercooled water and increased ability to grow quickly into extremely large hail sizes. Nonetheless, this is only one of many mechanisms that can result in lower beneficial competition needed for larger hail growth. In addition, some of this relationship might also relate to nonhydrostatic vertical pressure gradients induced by the rotating updraft as Blair et al. (2011) reveal a strong correlation between mid-level rotational velocity and hailstone size.

Results do not suggest any reasonable capability through parameter analysis alone to discriminate 2.0–3.25-in (51–83-mm) reports from the adjacent larger and smaller hail-size bins. These narrow ranges of hail size are nearly indistinguishable from each other with low predictive skill existing. Even when broadening a range to better detect ≥3.5-in (89-mm) hail, upstream convection may create anvil ice seeding that ultimately suppresses large hail development due to increased competition for supercooled water (Knupp et al. 2003). Additional refinement or narrowing of results would have to be accomplished through prediction of convective mode and duration, hail growth models such as HAILCAST (Jewell and Brimelow 2009), and/or other methods yet to be deployed.

This study raises many questions as to the physical explanation behind some of the parameter trends shown. The answers to these questions are well beyond the scope of this study without extensive modeling and/or analysis of storm microphysics in field projects. However, predictability problems will likely always exist. Further, hail embryo trajectories and subsequent growth into a hailstone is complex and dependent on processes not easily resolved in operational meteorology or modeled soundings. Nonetheless, future projects may build on this study's climatology to create new predictive tools for hail-size discrimination.

ACKNOWLEDGMENTS

The authors wish to express their appreciation to Larry Ruthi and Mike Umscheid (NWS Dodge City, Kansas) for technical editing and assistance with computer scripting. We would also like to thank Roger Edwards (SPC), John T. Allen (IRI, Columbia University), Scott Blair (NWS), and Jeff Manion (CRH) for their helpful reviews and thorough copyediting. The views expressed within this manuscript are those of the authors and do not necessarily represent those of the National Weather Service.

REFERENCES

Allen J., D. Karoly, G. Mills, 2011: A severe thunderstorm climatology for Australia and associated thunderstorm environments. *Australian Meteor. Oceanographic J.*, **61**, 143–158.

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

Baumgardt, D., cited 2014: Hail estimation: How good are your spotters? [Available online at http://www.crh.noaa.gov/arx/hail_size_MSP.pdf.]

Beatty, K., E. N. Rasmussen, J. M. Straka, and L. R. Lemon, 2009: The supercell spectrum. Part II: A semi-objective method for radar classification of supercell type. *Electronic J. Severe Storms Meteor.*, **4** (1), 1–30.

Benjamin, S. G., G. A. Grell, J. M. Brown, T. G. Smirnova, and R. Bleck, 2004: Mesoscale weather prediction with the RUC hybrid isentropic–terrain following coordinate model. *Mon. Wea. Rev.*, **132**, 473–494.

Blair, S. F., and J. W. Leighton, 2012: Creating high-resolution hail datasets using social media and post-storm ground surveys. *Electronic J. Operational Meteor.*, **13** (3), 32–45.

——, 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.

Bluestein, H. B., and C. R. Parks, 1983: A synoptic and photographic climatology of low precipitation thunderstorms in the southern Plains. *Mon. Wea. Rev.*, **111**, 2034–2046.

——, and G. R. Woodall, 1990: Doppler-radar analysis of a low-precipitation severe storm. *Mon. Wea. Rev.*, **118**, 1640–1664.

Brimelow, J. C., G. W. Reuter, and E. R. Poolman, 2002: Modeling maximum hail size in Alberta thunderstorms. *Wea. Forecasting*, 17, 1048–1062.

Brooks, H. E., C. A. Doswell III, and R. B. Wilhelmson, 1994a: The role of midtropospheric winds in the evolution and maintenance of low-level mesocyclones. *Mon. Wea. Rev.*, **122**, 126–136.

——, ——, and J. Cooper, 1994b: On the environments of tornadic and nontornadic mesocyclones. *Wea. Forecasting*, **9**, 606–618.

——, J. Lee, and J. Craven, 2003: The spatial distribution of severe thunderstorm and tornado environments from global reanalysis data. *Atmos. Res.*, **67**, 73–94.

Browning, K. A., 1963: The growth of large hail within a steady updraught. *Quart. J. Roy. Meteor. Soc.*, **89**, 490–506.

——, 1977: The structure and mechanisms of hailstorms. *Hail: A Review of Hail Science and Hail Suppression, Meteor. Monogr.*, No. **38**, Amer. Meteor. Soc., 1–43.

——, and G. B. Foote, 1976: Airflow and hail growth in supercell storms and some implications for hail suppression. *Quart. J. Roy. Meteor. Soc.*, **102**, 661–695.

Bunkers, M. J., B. A. Klimowski, J. W. Zeitler, R. L. Thompson, and M. L. Weisman, 2000: Predicting supercell motion using a new hodograph technique. *Wea. Forecasting*, **15**, 61–79.

——, M. R. Hjelmfelt, and P. L. Smith, 2006: An observational examination of long-lived supercells. Part II: Environmental conditions and forecasting. *Wea. Forecasting*, **21**, 689–714.

Changnon, S. A., 1972: Examples of economic losses, from hail in the United States. *J. Appl. Meteor.*, **11**, 1128–1137.

——, 1999: Data and approaches for determining hail risk in the contiguous United States. *J. Appl. Meteor.*, **38**, 1730–1739.

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

Coniglio, M. C., 2012: Verification of RUC 0–1-h forecasts and SPC mesoscale analyses using VORTEX2 soundings. *Wea. Forecasting*, **27**, 667–683.

Craven, J. P., and H. E. Brooks, 2004: Baseline climatology of sounding derived parameters associated with deep, moist convection. *Natl. Wea. Dig.*, **28**, 12–24.

——, R. E. Jewell, and H. E. Brooks, 2002b: Comparison between observed convective cloud-base heights and lifting condensation level for two different lifted parcels. *Wea. Forecasting*, **17**, 885–890.

Davies, J. M., 1993: Hourly helicity, instability, and EHI in forecasting supercell tornadoes. Preprints, *17th Conf. on Severe Local Storms*, St. Louis, MO, Amer. Meteor. Soc., 107–111.

Davies-Jones, R. P., D. W. Burgess, and M. Foster, 1990: Test of helicity as a tornado forecast parameter. Preprints, *16th Conf. on Severe Local Storms*, Kananaskis Park, AB, Canada, Amer. Meteor. Soc., 588–592.

Donavon, R.A., 2010: Updating the Donavon and Jungbluth 2007 hail criteria to one inch. *14th Annual Northern Plains Convective Conference*, Sioux Falls, SD. [Abstract located at http://sio.midco.net/weather/workshop2010/abstracts/08_Donavon.pdf.]

——, 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. III, and D. W. Burgess, 1993: Tornadoes and tornadic storms: A review of conceptual models. *The Tornado: Its Structure, Dynamics, Prediction, and Hazards*, Geophys. Monogr. No. 76, Amer. Geophys. Union, 161–172.

——, and E. N. Rasmussen, 1994: The effect of neglecting the virtual temperature correction on CAPE calculations. *Wea. Forecasting*, **9**, 625–629.

——, and L. F. Bosart, 2001: Extratropical synoptic-scale processes and severe convection. *Severe Convective Storms, Meteor. Monogr.*, **28**, No. 50, Amer. Meteor. Soc., 27–69.

——, and P. M. Markowski, 2004: Is buoyancy a relative quantity? *Mon. Wea. Rev.*, 132, 853–863.

——, and D. M. Schultz, 2006: On the use of indices and parameters in forecasting severe storms. *Electronic J. Severe Storms Meteor.*, **1** (3), 1–22.

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

——, 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.

——, J. T. Schaefer, D. W. McCann, T. W. Schlatter, and H. B. Wobus, 1982: Thermodynamic analysis procedures at the National Severe Storms Forecast Center. Preprints, *Ninth Conf. on Weather Forecasting and Analysis*, Seattle, WA, Amer. Meteor. Soc., 304–309.

Duda, J. D., and W. A. Gallus, 2010: Spring and summer Midwestern severe weather reports in supercells compared to other morphologies. *Wea. Forecasting*, **25**, 190–206.

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.

Esterheld, J. M. and D. J. Giuliano, 2008: Discriminating between tornadic and non-tornadic supercells: A new hodograph technique. *Electronic J. Severe Storms Meteor.*, **3** (2), 1–50.

Fawbush, E. F., and R. C. Miller, 1953: A method for forecasting hailstone size at the earth's surface. *Bull. Amer. Meteor. Soc.*, **34**, 235–244.

Foote, G. B., 1984: A study of hail growth utilizing observed storm conditions. *J. Climate Appl. Meteor.*, **23**, 84–101.

——, and J. C. Fankhauser, 1973: Airflow and moisture budget beneath a northeast Colorado hailstorm. *J. Appl. Meteor.*, **12**, 1330–1353.

Foster, D. S., and F. C. Bates, 1956: A hail size forecasting technique. *Bull. Amer. Meteor. Soc.*, **37**, 135–141.

Frisby, E. M., and H. W. Sansom, 1967: Hail incidence in the tropics. *J. Appl. Meteor.*, **6**, 339–354.

Guyer, J. L., and R. Ewald, 2004: Record hail event—Examination of the Aurora, Nebraska, supercell of 22 June 2003. Preprints, *22nd Conf. on Severe Local Storms*, Hyannis, MA, Amer. Meteor. Soc., 11B.1.

Hales, J. E., 1988: Improving the watch/warning system through use of significant event data. Preprints, *15th Conf. on Severe Local Storms*, Baltimore, MD, Amer. Meteor. Soc., 165–168.

Hart, J. A., and W. D. Korotky, 1991: The SHARP workstation v1.50 users guide. National Weather Service, NOAA, U.S. Dept. of Commerce, 30 pp. [Available from NWS Eastern Region Headquarters, 630 Johnson Ave., Bohemia, NY 11716.]

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.

Kitzmiller, D. H., and J. P. Breidenbach, 1993: Probabilistic nowcasts of large hail based on volumetric reflectivity and storm environment characteristics. Preprints, *26th Int. Conf. on Radar Meteorology*, Norman, OK, Amer. Meteor. Soc., 157–159.

Knight, C. A., and N. C. Knight, 2001: Hailstorms. *Severe Convective Storms, Meteor. Monogr.*, No. 28, Amer. Meteor. Soc., 223–248.

Knight, N. C., and M. English, 1980: Patterns of hailstone embryo type in Alberta hailstorms. *J. Rech. Atmos.*, **14**, 325–332.

Knupp, K. R., S. Paech, S. Goodman, 2003: Variations in cloud-to-ground lightning characteristics among three adjacent tornadic supercell storms over the Tennessee Valley region. *Mon. Wea. Rev.*, **131**, 172–188.

Laing, A. G., and J. M. Fritsch, 1997: The global population of mesoscale convective complexes. *Quart. J. Roy. Meteor. Soc*., **123**, 389–405.

Maddox, R. A., 1976: An evaluation of tornado proximity wind and stability data. *Mon. Wea. Rev.*, **104**, 133–142.

Markowski, P. M., J. M. Straka, and E. N. Rasmussen, 2002: Direct surface thermodynamic observations within the rear-flank downdrafts of nontornadic and tornadic supercells. *Mon. Wea. Rev.*, **130**, 1692–1721.

Marwitz, J. D., 1972: Precipitation efficiency of thunderstorms on the High Plains. *J. Rech. Atmos.*, **6**, 367–370.

Miller, D. J., 2006: Observations of low level thermodynamic and wind shear profiles on significant tornado days. Preprints, *23rd Conf. on Severe Local Storms*, St. Louis, MO, Amer. Meteor. Soc., 1206–1223.

Miller, L. J., J. D. Tuttle, and C. A. Knight, 1988: Airflow and hail growth in a severe northern high plains supercell. *J. Atmos. Sci.*, **45**, 736–762.

Miller, R. C., 1972: Notes on analysis and severe storm forecasting procedures of the Air Force Global Weather Center. AWS Tech. Rep. 200 (Rev.), Headquarters, Air Weather Service, Scott AFB, IL, 106 pp.

Moore, J. T., and J. P. Pino, 1990: An interactive method for estimating maximum hailstone size from forecast soundings. *Wea. Forecasting*, **5**, 508–526.

Nelson, S. P., 1983: The influence of storm flow structure on hail growth. *J. Atmos. Sci.*, **40**, 1965–1983.

NCDC, 2003–2011: *Storm Data*. [Available from National Climatic Data Center, 151 Patton Ave., Asheville, NC 28801–5001.]

Newton, C. W., 1963: Dynamics of severe convective storms. *Severe Local Storms, Meteor. Monogr.*, No. 27, Amer. Meteor. Soc., 33–58.

Orlanski, I., 1975: A rational subdivision of scales for atmospheric processes. *Bull. Amer. Meteor. Soc*., **56**, 527–530.

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

Rasmussen, E. N., 2003: Refined supercell and tornado forecast parameters. *Wea. Forecasting*, **18**, 530–535.

——, and R. B. Wilhelmson, 1983: Relationships between storm characteristics and 1200 GMT hodographs, low-level shear, and stability. Preprints, *13th Conf. on Severe Local Storms*, Tulsa, OK, Amer. Meteor. Soc., J5–J8.

——, and D. O. Blanchard, 1998: A baseline climatology of sounding-derived supercell and tornado forecast parameters. *Wea. Forecasting*, **13**, 1148–1164.

——, and J. M. Straka, 1998: Variations in supercell morphology. Part I: Observations of the role of upper-level storm-relative flow. *Mon. Wea. Rev.*, **126**, 2406–2421.

Renick, J. H., and J. B. Maxwell, 1977: Forecasting hailfall in Alberta. Hail: A review of hail science and hail suppression, Meteor. Monogr., No. **38**, Amer. Meteor. Soc., 145–151.

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

SPC, cited 2014: Significant hail parameter (SHIP): [Parameter description available at http://www.spc.noaa.gov/exper/mesoanalysis/help/help_sigh.html.]

Thompson, R. L., 1998: Eta model storm-relative winds associated with tornadic and nontornadic supercells. *Wea. Forecasting*, **13**, 125–137.

——, R. Edwards, and J. A. Hart, 2002a: Evaluation and interpretation of the supercell composite and significant tornado parameters at the Storm Prediction Center. Preprints, *21st Conf. on Severe Local Storms*, San Antonio, TX, Amer. Meteor. Soc., J11–J14.

——, ——, and ——, 2002b: An assessment of supercell and tornado forecast parameters with RUC-2 model close proximity soundings. Preprints, *21st Conf. on Severe Local Storms*, San Antonio, TX, Amer. Meteor. Soc., 595–598.

——, ——, and C. M. Mead, 2007: Effective storm-relative helicity and bulk shear in supercell thunderstorm environments. *Wea. Forecasting*, **22**, 102–115.

——, ——, J. A. Hart, K. L. Elmore, and P. Markowski, 2003: Close proximity soundings within supercell environments obtained from the Rapid Update Cycle. *Wea. Forecasting*, **18**, 1243–1261.

Weisman, M. L., and J. B. Klemp, 1982: The dependence of numerically simulated convective storms on vertical wind shear and buoyancy. *Mon. Wea. Rev.*, **110**, 504–520.

——, and ——, 1984: The structure and classification of numerically simulated convective storms in directionally varying wind shears. *Mon. Wea. Rev.*, **112**, 2479–2498.

——, and ——, 1986: Characteristics of isolated convective storms. *Mesoscale Meteorology and Forecasting*, P. S. Ray, Ed. Amer. Meteor. Soc., 331–358.

——, and R. Rotunno, 2000: The use of vertical wind shear versus helicity in interpreting supercell dynamics. *J. Atmos. Sci.*, **57**, 1452–1472.

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

Wilson, C. J., K. L. Ortega, and V. Lakshmanan, 2009: Evaluating multi-radar, multi-sensor hail diagnosis with high resolution hail reports. Preprints, *25th Conf. on Interactive Information Processing Systems*, Phoenix, AZ, Amer. Meteor. Soc., P2.9.

REVIEWER COMMENTS

[Authors' responses in *blue italics.*]

REVIEWER A (Roger Edwards):

Initial Review:

Recommendation: Accept with major revisions.

General comments: The modified "Rasmussen table" below summarizes my evaluation of this study. Related general and specific comments follow the table.

Criterion	Satisfied	Deficient, but can be remedied	Deficient; *cannot* be remedied by modifying the paper	Deficient, *not known* if it can be remedied by modifying the paper
1. Does the paper fit within the stated scope of the journal?	X			
2. Does the paper 1) identify a gap in scientific knowledge that requires further examination; 2) repeat another study to verify its findings; or 3) add new knowledge to the overall body of scientific understanding?	X			
3. Is the paper free of errors in logic?		X		
4. Do the conclusions follow from the evidence?	X			
5. How reproducible are the findings given the information presented?				X
6. Are alternative explanations explored as appropriate?		X		
7. Is uncertainty quantified?		X		
8. Is previous work and current understanding represented correctly?		X		
9. Is information conveyed clearly enough to be understood by the typical reader?		X		

The manuscript consists of a statistically driven environmental examination of severe hail, binned by size groupings, for selected years. A new index is proposed for operational use, the large hail parameter (LHP), that appears to outperform individual components and other traditional hail diagnostics, and that is analogous to the supercell composite (SCP) and significant tornado parameters (STP) in the tornadic realm.

My recommendation is "accept with major revision". The main value in this work is in its fresh insights on parametric hail prediction while somewhat incorporating microphysical (cloud-scale) concepts. A sound literature review forms the foundation. The top-level organization is keen. New and potentially useful material appears here, and conclusions generally follow from the analyses. While some important methodological concerns came up (major comments below), the paper is worth publishing pending due revision. I see no show-stoppers in a scientific sense that preclude proceeding, as long as the methodology concerns are clarified and the sampling methods are reconsidered.

Instead, the greatest burden of this manuscript, and what's responsible for most of the "major" characterization, lies in its presentation. The text is excessively verbose, too passive in tone, often redundant, often vague, and quite frequently written awkwardly, with numerous excessively long and/or run-on sentences. To a reader less familiar with the ideas, I imagine this draft would be akin to a mental slog through wet cement. Fortunately, three practices will improve the presentation hugely:

- Before revising, please read chapters 8–10 of *Eloquent Science* (Schultz 2009). They're short and straightforward, with examples. Doing so, along with a tough review to the same effect by Schultz himself, helped me to improve my own wordy excesses in a couple of papers. [Yes, I've been called out by reviewers for being too verbose in formal writing, and in hindsight, rightfully so! I do empathize with your plight here.] Those three chapters *specifically* address the biggest weaknesses of this paper's presentation: effective paragraphs (8), sentences (9) and phrases (10).
- Armed with that understanding, the authors confidently can address all stylistic improvements I suggest, major and minor, and maybe rewrite in *better* ways than I can suggest. To summarize—under the same basic structure, recompose the text in a way that flows. Make it concise, clear and uncluttered: all signal, no noise. Once all these changes have been made, the authors may be surprised at how much shorter the manuscript becomes *with no loss of factual information*.
- Using a proofreader with solid English compositional skills would have helped hugely here. Please consider doing this before resubmitting.

As a reviewer (and editor), my role here is to help the authors to refine their paper into publishable form, if the substance is robust enough. I think the substance already is adequate and can become very solid (with improvements suggested below). Even if my minor, in-line criticisms seem numerous and overwhelming, please accept them in the framework that I am trying to help this paper to be the strongest it can be.

I wish to read the next draft and accompanying point-by-point responses. Contrary to my normal, single-document reviewing style, the technical and copy revisions required of this paper make it more efficient to embed all minor comments in the Word document itself. Please reply to all major comments below. No reply is needed to minor comments in the markup *if the authors incorporate them*. However, in the markup, please do reply to any minor comments if there is a disagreement or alternative, so we all can track them efficiently.

We appreciate the time and effort you spent reviewing our paper and have attempted to address each issue raised as thoroughly as possible. Thank you for the recommended reading and editing practices as these were very useful in composing the revision. You will find several of your suggestions incorporated into the paper, although we do disagree with some of your points.

Substantive (major) comments: Data sampling choice: In one word: insufficient. Why just these four years when so much more data are available?

While this dataset is small relative to the number of hail reports each year, we disagree with the assertion that our sampling choice results in an insufficient number of events to produce meaningful results. Further, our quality control steps do not prevent us from creating sample sizes sufficiently representative and consistent with other parameter-based climatology literature. For comparison, we provide examples of sample size distributions associated with other parameter-based climatology studies:

- *Maddox (1976)*
  - *62 southwest flow tornado*
  - *19 westerly flow tornado*
  - *19 southerly flow tornado*
  - *23 tornado outbreak*
  - *10 nontornadic severe events*
- *Thompson (1996)*
  - *69 tornadic supercell*
  - *62 nontornadic supercell*
- *Rasmussen and Blanchard (1998)*
  - *51 tornado supercell*
  - *119 nontornadic supercell*
- *Rasmussen and Straka (1998)*
  - *17 classic supercells*
  - *13 HP supercells*

- *12 LP supercells*
- *Evans and Doswell (2002)*
  - *51 weak forcing derechos*
  - *47 strong forcing derechos*
  - *15 hybrid cases*
  - *18 significant tornadic supercell*
  - *34 weakly tornadic supercell*
  - *46 nontornadic supercell*
- *Thompson el al. (2003)*
  - *54 significant tornadic supercell*
  - *144 weakly tornadic supercell*
  - *215 nontornadic supercell*
  - *75 discrete nontornadic*
  - *15 marginal nontornadic supercell*
- *Bunkers et al. (2006)*
  - *184 long-live supercells*
  - *137 moderate-lived supercells*
  - *119 short-lived supercells*
- *Thompson et al. (2007)*
  - *113 significant tornadic supercell*
  - *280 weakly tornadic supercell*
  - *397 nontornadic supercell*
  - *45 elevated supercell*
  - *250 discrete nontornadic*
- *Esterheld and Giuliano (2008)*
  - *18 significant tornadic supercell*
  - *33 weakly tornadic supercell*
  - *16 nontornadic supercell*
- *Jewell and Brimelow (2009)*
  - *490 significant hail*
  - *424 nonsignificant hail*

This study has a sample size varying from 115 to 152 events in each hail-size bin that is well within the range and consistency of these other parameter-based climatology studies. While we would never argue against a goal of having a large sample size, we concur with the assessment made by Thompson et al. (2003) below:

> *"Though even larger sample size would be desirable, our sample sizes are reasonably consistent with comparable groupings in previous proximity sounding studies. Differing sample sizes may also impact the statistical interpretation of our results, but we believe our samples are sufficiently representative to allow comparisons among the storm groups."*

Arguments for or against our quality control procedures are likely to exist no matter how many events remain after completing the process. However, in the end our sample size matches many of these other studies. If our database is insufficient due to nonconsecutive years, relative frequency mismatching, or total sample size, then most of the parameter-based climatology studies over the last several decades are deficient by some or all of these standards as well.

Yes, I see the explanation that the years were chosen based on individual largest-hailstone sizes. That's meaningless. There's no known reason why a single hailstone in a given year, out of thousands of reports, should disqualify other hail-years from a study. Put another way, what is the physical sense in choosing a given year containing thousands of reports, based on the characteristics of one or two reports? Instead, I suggest similarly examining hail data from all years 2003–2011—or better yet, 2003–2012, since those data now are readily available on the SPC WCM page and constitute a big 10-y dataset. Resulting very large sample sizes should yield optimally robust analytic results, even for the giant-hail bin.

The years chosen were driven mostly by data availability on the NOMADS server. In contrast, the reference to the largest verifiable hailstones was a toss-in fact with little relevance to the selection process. Specifically, much of the RUC analysis for 2005–2009 was missing or only partially available during our data-gathering phase. However, availability of RUC files for 2003–2004 and 2010–2011 still allowed us to create a sufficiently representative sample size consistent with other parameter-based climatology studies such as those listed previously. Further, we chose not to expand beyond 2011 as the RAP v.1 had not undergone an extensive comparative analysis of assimilated data versus observed fields similar to the Thompson et al. (2003) RUC study.

We originally omitted specific details on the years chosen for sake of brevity. However, given the questions raised in this instance, we incorporated a brief outline of these details in the manuscript.

Resampling of data groupings: In section 2, the authors state that "a ratio" (whatever that means) of reports was constructed, in order to greatly reduce the sample sizes in the smaller-hail-diameter groupings, in turn making them comparable in number of events to the sample size of the giant hail bin. I understand the reasoning but am not convinced that it is necessary. *Reducing* sample size by an order of magnitude or more seldom is justifiable. Numerous statistical methods exist to normalize across large gaps in sample size; I'll leave it to the authors to investigate literature for which other methods may work best that do not involve blowing away huge chunks of data. My suggestion in major comment (1) above also will boost the giant-hail sample sizes considerably (along with all hail).

We have attempted to provide better clarity with this section of the manuscript. In particular, the process used here is an additional quality control step yet in our haste to shorten the paper, obviously oversimplified it to the point of confusing the reader. While we had what were likely unfounded worries over sample size distribution after the initial quality control steps, we also had lingering concerns with event inclusion at smaller hail sizes. Specifically, even minor size estimation errors were more likely to affect the correct bin assignment of a smaller hailstone unless additional inclusion criteria account for some of this concern. To help mitigate these issues, we developed an additional quality control process that makes event inclusion more stringent at smaller hail sizes while leaving the sample size issue rendered a passive side note. We provide clarification along with specific details on this process in the manuscript.

If the authors insist that the existing method (ignoring big gobs of readily available hail data) is best, convince me—on two fronts:

a. Why *not* just analyze the original groupings as-is and present the results, simply acknowledging the much smaller sample size of the giant-hail bin? Reducing the sample size of the smaller hail-size bins seems to be an exercise in robbing Peter to pay Paul, except without paying Paul.

The previous discussion should ease much of this concern. Since this study addresses the maximum expected hail size for an event, stringent removal of smaller hail reports occurring simultaneously and typically at a much higher frequency than larger hail was necessary. Simply loosening our criteria to include a higher percentage of a smaller hail to accommodate more reports, is a highly flawed argument that strongly blurs any signal in the larger hail-size bins.

b. What specifically is "objective removal"? That's too vague and glib of a descriptor for what is a crucial analytic foundation of the study. Ambiguity is the enemy of understanding. In formally writing about one's scientific analyses, methods need to be documented meticulously at every turn, so that an independent researcher could reproduce the results if he/she had the same data. *Be specific, for the sake of reproducibility*. State exactly what method of "objective removal" was performed, how, and why.

The previous discussion on use of additional quality control criteria should ease much of this concern. We provide clarification along with specific details on this process in the manuscript.

Fuzzy documentation and reasoning for upper-level winds: In section 3d, the equilibrium-level (EL) wind is defined as a "mean wind" through the 1.5-km layer below the EL. This is too vague and insufficiently supported. For reproducibility, specify in the text how this calculation is performed (e.g., pressure-

weighted or not, and based on how many vector-data points in that layer?). Please insert concise statements justifying:

a. That method of computation and
b. The choice of the (EL–1.5-km) thickness threshold.

We have attempted to address this by briefly summarizing the following.

The mean wind approach used here follows results from Rasmussen and Straka (1998). Specifically, they find the strongest signal at discriminating LP versus HP supercells occurs with upper-tropospheric flow typically 1–2-km below the tropopause. While this study also notes most supercells reach heights above the tropopause, the layer approach attempts to sample this zone of winds occurring at or below the EL. This process also permits sampling a broader range of winds near anvil height than a single level.

Representation of some earlier work: The comparison of the Thompson (1998) and Thompson et al. (2003) findings in section 3e is misleadingly worded, as if suggesting a conflict between them. There's no conflict, simply clarification of an earlier idea. Just because the signal didn't get larger doesn't mean there's conflict. Thompson stated that the 4–6-km SR wind speed did not appear to be a suitable parameter to discriminate between significantly tornadic and nontornadic supercells because: 1) mean difference between the two was only 1.4 m s^{-1}, and 2) the difference could be easily overshadowed by errors resulting from storm-motion estimates such as Bunkers'. These are important nuances lost in your current characterization of that literature.

This misleading wording was removed during revision of this section.

Innovative analysis with counterintuitive results (not necessarily a problem): In section 4b, the authors "compare the percentage of total MUCAPE existing only in the HGZ (%MUCAPE therein) against the four hail size groups..." This is a very intriguing, innovative and seemingly physically meaningful measure. At least, I haven't seen it before. Like the authors, I'm surprised at the inverse relationship of CAPE percentage in the HGZ to increasing hail size. To the extent possible, without wandering too far into speculative guesswork, the authors offered some additional elaboration in section 6. Is there any other clue from the hail-growth literature as to why this may be, from a physical perspective [e.g., larger overall CAPE may reduce the percentage in any particular thermal layer and also implies greater liquid-water content with which to manufacture hail, fast lofting of big-CAPE hydrometeors out of that layer (i.e., precip-efficiency arguments), or some other process]?

Unfortunately, there is simply a dearth of any literature specifically addressing this relationship. Logically, it would make sense that if the HGZ is shallow yet instability is no higher, then the hail embryos are ejecting into the anvil due to inadequate growth time. However, this is purely speculative since I have no direct evidence to support this theory.

Overly ambiguous comparison descriptors: In section 5, on a few occasions, the words "good" or "decent" are used to describe interquartile separation or independence of statistically analyzed parameters. Those descriptors are too vague and subjective. What constitutes "decent" versus indecent? What constitutes good vs. no good, at least in a reproducible sense? How do you measure and define goodness of matching in that context? Please quantify this, define such distinctions in a specific way, or choose more precise and literally defensible verbiage.

We have attempted to address this concern throughout the manuscript.

Acknowledging potential weaknesses in an index: As a co-author of the formally vetted SCP and STP, I certainly am in no position to oppose the development of other indices and parameters that have some physical reasoning behind them, and that includes LHP. I actually like the LHP concept in general, and have no substantive qualms with its development or formulation. That said, the concerns elucidated by Doswell and Schultz (2006) need to be addressed in section 5, with short discussion and a citation thereto, because their paper is directly relevant to the advocacy of any index intended for operational use. One of those concerns is that LHP (like SCP and STP) appears to be a diagnostic, not prognostic index. Again,

I'm not saying to dump LHP—far from it—but instead, to address more completely both its strengths and limitations. On a related note, in section 6, please briefly discuss potential areas of improvement or refinement to LHP, as a follow-up to what I requested above for section 5.

Thank you for the referenced literature as this will be a good addition to the manuscript. You are correct in that the LHP is more of a diagnostic variable rather than prognostic. Further, by no means are we suggesting the LHP or other indices use be limited to the only variable analyzed, but more as a tool that quickly finds a target area where these index values in combination with deeper analysis provide a full picture of potential hail sizes. We have attempted to address these concerns in both section 5 and section 6.

[Minor comments omitted…]

Second Review:

Recommendation: Accept with minor revision.

General comments: The modified "Rasmussen table" below summarizes my evaluation of this study. Related general and specific comments follow the table.

Criterion	Satisfied	Deficient, but can be remedied	Deficient; *cannot* be remedied by modifying the paper	Deficient, *not known* if it can be remedied by modifying the paper
1. Does the paper fit within the stated scope of the journal?	X			
2. Does the paper 1) identify a gap in scientific knowledge that requires further examination; 2) repeat another study to verify its findings; or 3) add new knowledge to the overall body of scientific understanding?	X			
3. Is the paper free of errors in logic?	X			
4. Do the conclusions follow from the evidence?	X			
5. How reproducible are the findings given the information presented?		X		
6. Are alternative explanations explored as appropriate?		X		
7. Is uncertainty quantified?	X			
8. Is previous work and current understanding represented correctly?	X			
9. Is information conveyed clearly enough to be understood by the typical reader?		X		

Thank you once again for reviewing our paper and providing a second round of valuable input. All minor comments were incorporated as suggested by the reviewer.

Substantive (major) comments: Data sampling choice: The authors mostly have done a more thorough and reproducibly specific effort to justify their choice of sampling. The previous explanations were insufficient and incomplete; these are much better, e.g.:

- Years of choice being tied to data availability from NOMADS instead of some association with a physically irrelevant peak-hail size report within a given year and
- Specifying the more stringent inclusion criteria at smaller sizes).

As such, the choice of years now is acceptable, and while I'm not entirely comfortable with the small-size criteria, those too are at least marginally defensible. Still, it's unclear from existing documentation if more hail can be used (below).

Regarding the sample sizes used by a number of other studies, the authors offer compelling-sounding points in their review response that are factually true but not very pertinent. The logistical limitation in each of those cases was not the same as that here. As such, using other studies' sampling numbers appears to be a conflation, or "apples-and-oranges" fallacy. For example, in the study by Evans and Doswell, no data were available for additional derechoes. In Thompson et al. (2003, 2007), those were the only cases available to the authors due to their real-time, operationally constrained data-collection methods (data expired from systems used). [As a co-author on the latter studies, I can attest that we settled for the sample sizes involved and were able to justify them to reviewers as sufficient, but as noted in the authors' quoted statement, larger sampling was desirable and certainly would have been beneficial. In short, we gladly would have used more data without hesitation, had it been available!]

Unlike in those cases, full CONUS data east of the Rockies *are available readily* online, even *within those chosen years*. I want to verify with the authors, and they should state so specifically in the text if true, *all possible raw hail data* in those years has been processed through their QC filters. [This wasn't clear in the latter part of section 2 where the methodology and QC are described.] If not, the authors should go ahead and use it (accessed via the SPC WCM page), without regard to sample sizes of other studies that were performed under much more restrictive logistic contexts. More data only can help, not hurt, the analyses.

Yes, all possible raw hail reports process through our quality-control process. We have addressed this question with a brief entry on page 4.

[Minor comments omitted...]

REVIEWER B (John T. Allen):

Initial Review:

Recommendation: Accept with minor revisions.

Overall Review Characteristics:
Scientific Content: Excellent
Organization: Very Good
Impact: Excellent
Writing: Very Good
Figures & Tables: Very Good
(Scale: Very Poor, Poor, Fair, Good, Very Good, Excellent)

Overview: The authors present an interesting and detailed proximity analysis of the conditions associated with large hail, and in doing so begin filling an important gap in existing research. They show that the distribution of many of the commonly applied forecast parameters have little skill for hail prediction. To address this limitation, they also present an analysis of a new large hail parameter that improves on the significant hail parameter and the significant severe covariate. I only have minor comments for this manuscript regarding the formulation of parameters that are used, missing references, discussion of sample size, and the number of figures included.

Thank you. We appreciate the time and effort you spent reviewing our paper. All minor comments were incorporated as suggested by the reviewer, with the exceptions listed below.

[Editor's note: A comment termed "minor" by the reviewer appeared substantive enough to include in the review record.]

With respect to the choice of proximity sounding, the authors utilize a point sounding that is closest to the observed hail event. They might consider the results of Potvin et al. (2010) that discuss the implications of the choice of proximity distance and time on the resulting soundings. It may be that the nearest point in the hour preceding the storm corresponds to a convectively mixed environment in the RUC, as the models timing will not necessarily synchronize mesoscale features with the observed event. The relative position of this sounding point to the observation may also be important, particularly where boundaries are involved. Did the authors consider whether the nearest downstream proximal sounding would be more representative than the closest?

Potvin et al. (2010) specifically addresses questions of representativeness with observed proximity soundings similar to Orlanski (1975) and Brooks et al. (1994b). While not a completely irrelevant issue with temporal items, the use of RUC analysis negates much of this concern, with most events occurring within 30-45 minutes of the analysis data. In terms of spatial concerns, we loaded RUC surface winds, dewpoints and surface-based CAPE in IDV to briefly diagnose any representativeness issues with the analysis data. In a few instances, switching to a different gridpoint within 40–50 km of the event was necessary to sample the inflow sector.

[Minor comments omitted...]

Second Review:

Recommendation: Accept with minor revisions.

Overview: The authors have addressed each of my original concerns and I only have 2 remaining further minor comments. In my opinion the manuscript has benefitted from the Round 1 revisions and copyediting and is much improved, reading as more succinct, precise and manageable for readers.

Thank you once again for reviewing our paper and providing a second round of valuable input. All minor comments were incorporated as suggested by the reviewer.

[Minor comments omitted...]

REVIEWER C (Scott F. Blair):

Initial Review:

Recommendation: Accept with major revisions.

Summary: The authors have done quite a bit of work collecting and analyzing a relatively exhaustive list of environmental parameters, and exploring whether or not individual or combinations of these parameters may be used to predict the potential for large hail in the pre-storm forecast phase. The study attempts to rectify and build upon a knowledge gap in the ability for operational forecasters to accurately predict maximum hail size prior to storm formation and therefore seems well-suited for the EJSSM. The paper overall is very well-written, figures are aesthetically pleasing and clear, and the reference list is extensive.

My primary concerns at this initial stage in the review process are with some of the data and methodology chosen for the research, and the characterization of supercell-produced hail. I have provided several suggestions that might both result in a stronger signal in the data and limit some uncertainty in the results.

We appreciate the time and effort you spent reviewing our paper and have attempted to address each issue raised as thoroughly as possible. You will find several of your suggestions incorporated into the paper, although we do disagree with some of your points.

General and substantive comments: The authors explain that the four years selected (2003, 2004, 2010, and 2011) in the database are based on six of the seven largest verifiable hailstones in *Storm Data*.

However, it is unclear how these singular extreme events have any relationship to the annual hail fall or whether the hail data in these select years are better representative or more accurate than any others in recent history. A quick look at hail reports across the United States from 2003-2011 show minor annual variability.

Year – ≥ 3.50", ≥ 0.75"
2011 – 98, 17312
2010 – 57, 10908
2009 – 54, 13313
2008 – 75, 17760
2007 – 42, 12712
2006 – 62, 16665
2005 – 45, 13721
2004 – 52, 13170
2003 – 66, 13915

I would expect some readers to wonder why non-consecutive years were chosen (and why five years of data were excluded), as opposed to using all available data from 2003–2011. The inclusion of additional years to the database would be very useful to determine whether a stronger signal could be obtained and whether the results found in the paper can be replicated.

The years chosen were driven mostly by data availability on the NOMADS server. In contrast, the reference to the largest verifiable hailstones was a toss-in fact with little relevance to the selection process. While we concur these years are not unique in terms of the total number of hail reports, we disagree with the notion that a temporal gap of a few years in our database is somehow uncommon to the reader or produces inconclusive results. A simple, cursory literature examination returns various severe-convective articles containing databases with similar temporal gaps and themes as this manuscript with a brief list provided below:

- *Bunkers et al. (2000) in evaluating the "internal dynamics method" includes data from five different sources, many containing multiple non-consecutive years between events.*
- *Rasmussen and Straka (1998) evaluate a database of supercells that include several non-consecutive years.*
- *Thompson et al. (2007) detail the use of effective shear by combining two databases of RUC proximity soundings that have a temporal gap of over a year between these databases.*
- *Evans and Doswell (2002) study on derecho and supercell proximity soundings, contains a five-year gap between events.*

None of these articles devotes a section to a detailed explanation on temporal gaps nor do they list any reason why non-consecutive years would harm the results. By no means are we suggesting a database containing only consecutive years is of any lower quality or a bad target. Instead, the concern raised here is not an established practice in creating a sufficiently representative sample size. Further, the temporal gap in this manuscript database does not differ in any substantial way from this other literature nor are results influenced in any undo manner that would invalidate results.

For clarification purposes, the temporal gap relates to missing or only partially available archived RUC files on the NOMADS server. Specifically, much of the RUC analysis for 2005–2009 was missing or only partially available during our data-gathering phase. Similar to the previously listed literature, we omitted specific details on the years chosen for sake of brevity. However, given the questions raised in this instance, we incorporated a brief outline of these details in the manuscript.

The selected years in the database are also problematic when using RUC data for the proximity soundings. The authors acknowledge the change in horizontal grid spacing (20-km to 13-km) between the two subsets of years. However, additional changes were made to the RUC beyond just the horizontal resolution that likely impacted some of the model parameters. The switch to the diabatic digital filter initialization (DFI), assimilation of radar, mesonet, soil moisture, etc., and significant changes to the physics packages

(convective parameterization, radiation, microphysics, etc.) all impacted the 0-h initialization (refer to the briefing documents on ruc.noaa.gov). The authors should acknowledge that these changes could have impacted their data, and should also investigate and disclose if there were any significant differences between the parameters for the 2003–04 data and the 2010–11 data.

We have addressed this concern by mentioning the additional model changes and listing of the online briefing documents. Further, in a similar fashion to model changes in Thompson (1996) and Thompson et al. (2007), we detail no noticeable impact on diagnoses or data differences.

One notable concern I have with the methodology is the objective removal of hail cases to accomplish the author's desire for a similar proportion of cases among all hail size bins. First, the authors must provide sufficient detail on how these objective removals of data were performed, so that these methods may be replicated. As it stands, there is not enough information to understand how and why specific cases were removed, outside of the quality-controlled process described.

We have attempted to provide better clarity with this section of the manuscript. In particular, the process used here is an additional quality control step yet in our haste to shorten the paper, obviously oversimplified it to the point of confusing the reader. While we had what were likely unfounded worries over sample size distribution after the initial quality control steps, we also had lingering concerns with event inclusion at smaller hail sizes. Specifically, even minor size estimation errors were more likely to affect the correct bin assignment of a smaller hailstone unless additional inclusion criteria account for some of this concern. To help mitigate these issues, we developed an additional quality-control process that makes event inclusion more stringent at smaller hail sizes while leaving the sample size issue rendered a passive side note. We provide clarification along with specific details on this process in the manuscript.

Second, the authors state that the objective removal of data was to avoid unequal weighting of smaller versus larger hail events. However, the inclusion of additional data would not prohibit comparison and analysis of the box-whisker plots that were used to examine the data throughout the study. Instead, more cases would only increase confidence that the results represent a real and consistent signal.

The previous discussion should ease much of this concern. However, while we concur that additional data does not prohibit "comparison and analysis", we strongly disagree with the assertion that our sample size is somehow insufficient to produce a "real and consistent signal".

Storms producing hail ≥3.50 in in diameter are a much more infrequent occurrence in nature than storms producing smaller maximum diameter sizes (0.75–1.25 in), and therefore it is reasonable that there be a disproportionate number of cases for comparison.

While we have no disagreement with your statement that ≥3.5-in hail events are less frequent than 0.75–1.25-in hail, the notion that there should be a disproportionate number of cases ignores quality-control steps that can have substantial influence on the distribution of events. There is little argument in a much higher frequency of 0.75–1.25-in hail reports, but for every 1.75-in hail report on a particular day, there may be 25 or more 0.75–1.25-in hail reports occurring in the same geographic region that are excluded through quality-control steps. Ultimately, the goal of this study is predicting the maximum expected hail size for a particular event, not in accessing the relative frequency of any particular hail-size bin.

It is believed that the environment and storm mode responsible for producing extreme events (≥3.50 in) are very different compared to events that produce maximum-sizes of marginally severe hail (1.00 in) or robust multicellular storms (≈1.50 in).

The reviewer comments on storm mode are extremely speculative pertaining to the smaller hail sizes. Specifically, outside of the strong likelihood of supercells for events with ≥2.0-in hail, the assertion that the environment and storm mode for 1.0-in events or 1.5-in hail cases are "multicellular" is unsupportable. A more accurate statement of "mixed-mode" or "multimode" would fit, but labeling anything with <2.0- in hail as "nonsupercell" is a misrepresentation.

Therefore, the environments supporting smaller hail are presumably more diverse since they occur much more frequently, and removing the majority of these cases (3,000 reports reduced to 115 cases for hail diameter ≤1.25 in; approximately 30 cases for each quarter-inch hail size), even objectively, subjects the study to a potential misrepresentation of the wide range of environments in which these smaller-hail sized events may occur.

We strongly disagree with the assertion this study is creating a potential misrepresentation of results by not matching the relative frequency of a particular hail group. While we agree that 0.75–1.25-in hail events display more diversity in environments, this study addresses the maximum expected hail size for an event not the full range of sizes produced in a particular day. This requires stringent removal of smaller hail reports occurring simultaneously and typically at a much higher frequency than larger hail or with questionable size estimation. This allows us to isolate, for example, 0.75–1.25-in hail environments to those that could only produce this maximum size of hail with the term "maximum" still implying any smaller hail size as a likely possibility. Simply loosening our criteria to include a higher percentage of a smaller hail to accommodate a larger diversity of environments is a highly flawed argument that strongly blurs any signal in the larger hail-size bins.

In addition, it is not an established practice in other parameter-based climatological studies to adjust sample size distributions to match event relative frequency. Another simple, cursory literature examination returns several studies containing databases with a sufficiently representative sample size that do not match event relative frequency (brief list provided below).

- *Maddox (1976) in evaluating tornado proximity wind and stability data creates a database of 159 combined tornado events but only 10 nontornadic severe events.*
- *Thompson et al. (2003) create a database with 413 combined supercell events but only 75 discrete nonsupercell events and 15 marginal supercell cases.*
- *Thompson et al. (2007) builds on the previous literature by creating a database with 835 combined supercell events but only 350 nonsupercell events.*
- *Jewell and Brimelow (2009) in evaluating HAILCAST create a database of hail events spanning the years of 1989–2004. They briefly document 72% of all reports are from hail in the 0.75–1.0-in range yet after quality-control, the database contains 490 ≥2.0-in hail events and only 420 <2.0-in cases.*

The smaller sample size of nonsupercell, nonsignificant hail, or nontornadic events relative to the lower frequency supercell, significant hail, or tornadic events does not invalidate these studies. In this same manner, our manuscript database does not differ in any substantial way nor are results influenced in any undo manner that would invalidate results.

I think the authors have encouraging preliminary findings, but their research would be greatly enhanced by the inclusion of additional data to both strengthen their initial results and attempt to smooth any reporting or data inconsistences found with Storm Data. This seems especially important if they wish to further validate the LHP as an operational forecast tool.

While this dataset is small relative to the number of hail reports each year, our quality control steps do not prevent us from creating sample sizes sufficiently representative and consistent with other parameter-based climatology literature. For comparison, we provide examples of sample size distributions associated with other parameter-based climatology studies:
[Editor's note: The same list and succeeding text appears in the Review A response and, for brevity, is not reproduced here.]

If possible, the research would greatly benefit from the inclusion of at least some high-resolution hail data, such as SHAVE. Storm Data may be suitable for research applications if ample cases are utilized and available high-resolution datasets are used for comparison. The authors do acknowledge the quality issues associated with Storm Data and briefly mention the SHAVE project in the data section, but do not attempt to remedy their low-resolution dataset by supplementing it with higher quality data. Comparison and verification of the study results with a subset of SHAVE data would greatly strengthen the paper and provide a degree of improved confidence in the findings.

We briefly investigated use of SHAVE reports but as you likely know, the project did not exist for the 2003–2004 RUC years. Further, with limited resources at their disposal SHAVE functions over a restricted timeframe and geographic region each year. Because of these limitations, reports from 2010–2011 were small enough in quantity to where most were superseded by larger hail sizes from Storm Data. We subsequently did not get enough additional reports to warrant inclusion of another data source.

In addition to expanding the database, the research would benefit from some sort of statistical testing (a Student's t-test or similar) on the various hail parameters explored in the study, in order to determine whether the values for each hail size bin are statistically different. This is especially true for the parameters that are used in the LHP. The manuscript only mentions the interquartile overlap, but struggles to actually quantify the difference in parameters between the bins.

Excellent point as we have addressed this concern with inclusion of a simple t-test when discussing various parameters.

There is some amount of precision implied by using four specific bins for the maximum hail size in a storm—perhaps too much precision, considering the inherent uncertainty found in *Storm Data*. This should be explicitly stated in addition to the *Storm Data* caveat. Reducing the number of bins to three may also strengthen the classification scheme, which is discussed below.

A degree of uncertainty does exist in any hail database and Storm Data *is not exempt, especially at the smaller hail sizes. We attempt to account for some of this with our quality control process but it will likely always be an imperfect system of reporting accurate hail sizes. However, results suggest a notable change between 0.75–1.25-in events and 1.5–1.75-in cases (2.0-in events now placed into adjacent hail bin) and reducing the bins to three groups ignores a real relationship difference.*

It would be beneficial for the authors to explain in the methodology why these specific hail size bins were chosen. For example, the first two bins contain half-inch increments (0.75–1.25 in; 1.50–2.00 in), the third bin is a one-inch increment (2.25–3.25 in), and the final bin is anything >3.50-in diameter hail.

This was originally omitted for sake of brevity but an explanation is now included in the manuscript (section 3, p. 3).

One concern I have with the research is the lumping of 2.00-in hail reports in with the 1.50-in and 1.75-in reports in a singular bin. The authors cited several references in the paper that have shown ≥2.00" to be a good proxy for hail originating from supercell thunderstorms (see additional reference, Duda and Gallus, 2010, *Wea. Forecasting*). As it currently stands, it is probable that the 1.50–2.00-in bin encompasses a variety of storm modes, which likely dampens the signal within this particular size bin. In fact, the two bins on the right side of [then] Figs. 17 and 23 illustrate well the difference between storm modes [supercells (hail ≥2.25 in) versus non-supercells (hail ≤1.25 in)].

As you note, there is a long literary record of separating 2.0-in events from 1.75-in cases as a threshold for supercell versus mixed mode events. Due to this literary record, we hesitantly move the 2.0-in cases into the adjoining hail-size bin but also note the information below as a precautionary statement. Luckily, this accounts for only a few events in the original 1.5–2.0-in hail size bin and none of these cases results in removal of a 0.75–1.25-in case in our quality control process.

Although we moved 2.0-in events into the larger hail size bin, we still have reservations about separating 2.0-in events from 1.75-in cases. In particular, Baumgardt (2014) reveals a negative bias toward 2.0-in hail sizes reported with a 1.75-in diameter.

Similar to previous reviewer comments on storm mode, these labels for convective mode are extremely speculative and unsupportable. Further, even with the revised distribution of hail sizes, the signal among the various parameters typically associated with supercell diagnosis changes very little. This actually provides a slight argument toward 1.5–1.75-in hail being the lower limit for supercell events rather than 2.0-in hail.

I believe it would be beneficial for the authors to slightly adjust their hail size bins to a storm mode-based classification scheme, which will not only provide a sound, physical basis for the bins chosen, but may also improve the signal in their results. One option is to reduce the bins to three: non-supercell (1.00–1.75 in), supercell (2.00–3.25 in), extreme event (≥3.50 in); or if four is still preferred: marginally severe (0.75–1.00 in), non-supercell (1.25–1.75 in), supercell (2.00–3.25 in), extreme event (≥3.50 in). [Editor's Note: since supercells can produce 1.00–1.75-in hail, maybe "non-supercell" isn't the most appropriate name, but the reviewer's point still is a very good one with the substitution of a word such as "multimodal" for "non-supercell".] Either way, it seems advantageous for 2.00 in hail events to be grouped with other presumed supercells; and this would fall in line with previous work and the operational/literary nomenclature "significant hail." A realignment of the bins may also help the authors better explain environmental differences between the bins found in the results and conclusions, as the narrow range of sizes may have been hindering their analysis.

As noted previously, we moved 2.0-in cases into the adjacent larger hail-size bin. However, given the reasoning behind the chosen hail-size bins (section 3, p. 3) along with results shown in this study, there is little argument toward adopting the other suggestions. Specifically, the three bin suggestion arbitrarily dumps 0.75–0.88-in events (dangerous precedent for events within 0.25-in of the popular quarter size threshold), improperly assigns 1.0–1.25-in events into 1.5–1.75-in group despite notable parameter differences, and provides a strongly speculative and misrepresentative naming scheme of "nonsupercell" for <2.0-in events. In addition, the four hail-size bin suggestion makes a better argument but still improperly assigns 1.25-in events into 1.5–1.75-in bin despite the noted bias in reporting of 1.25-in events as 1.0 in. Further, it also continues the practice of a strongly speculative and misrepresentative naming scheme of "nonsupercell" for <2.0-in events.

There are several occurrences in the paper where the author's description of a particular hail size is confusing, and this presumably a result of the classification scheme; e.g., "significant hail (>2.0 in)" and "marginal severe hail events (<1.5 in.) from larger hail events (>2.0 in)." The reader may interpret that "significant hail" begins at 2.01 in or "marginal severe" is 1.49 in. This is an artifact of the selected hail bins. Please change these values to match your bins where you have known data; for instance, "marginal severe hail events" would be (≤1.25 in).

Thank you. We have addressed this concern in the revised manuscript.

The authors state several times that maximum diameters of hail ≥1.50 in emanate from supercells. However, the literature referenced early in the paper instead states that hail ≥2.00 in in diameter is mainly produced by supercells, not hail as small as 1.50 in. I am unaware of literature that explicitly states that hail with a diameter of 1.50 in results solely from the supercell storm mode, but I am much more familiar with those studies that utilize 2.00 in as a good proxy for hail size generated by supercells.

We have addressed this concern as most of this emanates from our original mixing of 2.0-in cases with 1.5–1.75-in events.

Additional testing of the LHP is needed to truly evaluate the skill of the parameter as an operational forecast tool. It is not surprising the LHP showed some skill with the dataset, since the parameters chosen for the LHP were based off higher-performing individual parameters of the same dataset. Consider testing its skill for the years that were not included in the study, or for high-resolution data (see above).

While we understand the nature of this statement, this inquiry appears well beyond the scope of this study and for the matter most other literature unveiling a new index. Specifically, through a cursory review of severe-convective literature detailing new indices (e.g., Thompson et al. 2003; Craven and Brooks 2004), it is not common practice for authors to create multiple databases for numerous rounds of testing and evaluation. While this concept is not a bad idea, this inquiry is best suited for subsequent studies focusing on hail size prediction. As Rasmussen and Blanchard (1998) evaluate tornado and supercell predictive skill of the energy-helicity index derived by Hart and Korotky (1991) and Davies (1993), future hail-based studies are best suited for further evaluation of LHP.

[Minor comments omitted...]

Second Review:

Recommendation: Accept with major revisions.

Summary: The revised manuscript has been much improved overall. The study illustrates well the potential, along with the challenges, in anticipating specific maximum diameter hail sizes from diagnosing the environmental conditions prior to storm formation. While I believe the study as a whole is well conducted, the primary nagging issue stems from the methodology, which draws some concern to the results presented. Beyond this issue, I have only minor comments as most of my previous concerns were addressed.

Thank you for reviewing our paper and providing a second round of valuable input. We have attempted to address each issue raised as thoroughly as possible. As before, you will find several of your suggestions incorporated into the paper, although we do disagree with some of your points.

Substantive comments: The written methodology used for obtaining individual hail events ultimately placed in one of four hail size bins remains confusing and a problematic portion for replicating the research.

"Only the largest hail report was initially included for a particular event."
How is an event defined in this study? Based on the information available, it does not appear radar imagery was examined for each event as there is no mention of radar data in the paper and the authors state that storm mode was unknown for each event.

We have addressed this concern as our use of the word "event" is leading to a misinterpretation that we group all reports emanating from one storm or a cluster of storms into "events" prior to quality control. However, we examine each report as a separate occurrence independent of the density of surrounding reports prior to quality control. Further, given the pre-storm focus of this manuscript where specific convective details are unknown, clustering of events would not match the research goals of this paper. The revised manuscript replaces this confusing wording with a simple "report" reference where appropriate.

"Further, any subsequent hail reports of similar size were only included if they were beyond 6 h or 250 km from the other report. Identical conditions apply to inclusion of smaller hail except they must also pass the criteria against any of the larger hailstones removed in the previous step."
I interpret this as once the largest sized stone is identified in the previous step by "event", then additional maximum-sized events are allowed to be included into the database, assuming they fit the criteria of "beyond 6 h or 250 km" from the first event selected.

The previous discussion should assist with this question. In particular, the interpretation is correct but replace "event" with "report" as all hail reports process through this quality-control step.

According to the information in the first draft, this yielded 5400 hail events that fit into one of the four hail size bins. Presumably the data at this stage (5400) should provide a good database that is representative of each storm of interest and its environment. Otherwise, it is not clear why the "6 h or 250 km" methodology would have been chosen.

While we understand the confusion emanating out of the first draft's quality-control description, this section already underwent significant editing in the previous revision as reviewer response drove a more detailed description. Specifically, the previous revision contains no reference to the 5400 toss-in value as the >6 h or >250 km thresholds are only one-step in the quality control process that also require reports <250 km apart be no more than one bin size in difference if the reports are <12 h apart. Further, a final round of quality control steps examine reports for questionable timing or location issues and these steps together leave the 5400 value with little relevance to the final report number.

In the initial submission, the authors had concerns with a disproportionately higher number of smaller sized hail storms compared to larger hail storms. As I stated in my previous review, I'm unclear why this would be an issue as the QC methodology utilized should avoid any duplication of individual storms or their nearby environment. If the authors were concerned with the inclusion of too many storms using the 6-h and 250-km requirement, then the initial spatiotemporal range should have been altered to something greater.

Even though we had what were likely unfounded worries over sample size distribution after the initial quality control steps, we also had lingering concern with report inclusion at smaller hail sizes. Specifically, even minor size estimation errors were more likely to affect the correct bin assignment of a smaller hailstone unless additional inclusion criteria account for some of this problem. A quality-control step aimed at mitigating the bin assignment errors with smaller hailstones indirectly narrows the sample size distribution. A final round of quality control steps examining hail report integrity combine with the previous QC step to result in the current distribution without simply dumping a report.

At this stage, the methodology remains very unclear, especially when comparing the initial and second draft of the manuscript. The authors initially stated that "to avoid unequal weighting of smaller versus larger hail events, utilization of a ratio of reports in the ≥3.5 in. group relative to those in each of the three smaller groups allow objective removal of cases until similar proportionality exists among all groups." This process/ratio is not described in the second draft of the methodology, but is critical since this objective procedure excluded a substantial number of cases. A detailed description (at least to this reviewer) of the process that allowed this objective removal of reports until proportionally similar groupings existed must be disclosed prior to publication.

We echo your concern with the quality-control description in the original manuscript. However, we are unclear why the reviewer refers to the defunct first draft since the original misleading wording was eliminated in the previous revision. As detailed in the discussions above, additional quality control steps make report inclusion more stringent at smaller hail sizes and along with a final round of quality-control steps, leaves the sample size issue rendered a passive side note (e.g., reports are not simply dumped). We provide clarification along with specific details on this process above and in the manuscript.

Further, eliminating 90% of events (5400 to 520) that passed the initial primary quality control process is significant, and as described in the previous review, seemingly does not benefit the research providing fewer data. Having a disproportionate number of events in the selected four bins should be expected, and assuming the initial QC procedure is sufficient to isolate conditions for each event, adding these cases would further strengthen the research. Forcing all four size bins to similar sample sizes brings up additional questions how this was objectively accomplished and its reproducibility.

As stated previously, the "5400" number is ultimately a toss-in fact with little relevance to the final report number. However, the notion that there should be a disproportionate number of reports ignores quality control steps that can have substantial influence on the distribution. Ultimately, the goal of this study is predicting the maximum expected hail size for a particular event, not in accessing the relative frequency of any particular hail-size bin. Simply loosening our criteria to include a higher percentage of a smaller hail to accommodate a larger disproportionate number of reports, is a highly flawed argument that strongly blurs any signal in the larger hail size bins.

Based on the explanation in the previous review, the authors mentioned they had concerns with smaller hail sizes and potential minor size estimation errors that may affect the smaller bins. The revised manuscript cites an informal study by Baumgardt, suggesting that spotters may report hail sizes to the nearest common object, which could affect the two smallest bins. The authors write: "Essentially, at smaller sizes the odds increase this study could unknowingly sort some hailstones into the wrong bin. To address this issue, reports passing the initial criteria but occurring within short distances of larger hail were a logical target." This needs more explanation. Does this mean that specific hail sizes were eliminated from the study because the authors believe they might have been incorrectly reported by a spotter due to the stone being close in size to a common object? If so, the specific sizes removed should be stated.

We have addressed this concern as individual hail sizes are not completely eliminated from the study but rather reports occurring <250 km apart must also must be no more than one bin size in difference if the reports are <12 h apart. Essentially, if any ≤1.75-in report fails this quality-control step then the report is excluded from the database yet any other hailstone in the same range that passes the criteria is included. Given the level of confusion, some additional clarification follows in the manuscript.

With *Storm Data*, there is inherent uncertainty in all estimated maximum-diameter hail sizes, regardless if the report is the same size as a common object or not. This includes reports of larger hail sizes (baseball, softball) [the authors mention the uncertainty of reported softball sizes in the paper and likewise adjust their bin to 3.5 in]. If the authors are legitimately worried that near common object-sized hail reporting might compromise a signal following their initial QC procedure, why not change the range/number of small hail-sized bins in the study instead of removing more data? (The 1.5–1.75-in bin is particularly precise and presumably could be more sensitive to errors; I understand the attempt to capture these types of storms, but not at the cost of eliminating data from speculation.)

We agree that a level of uncertainty exists in any hail database and Storm Data is not exempt. We attempt to account for some of this with our stringent quality-control process but Storm Data remains an imperfect system of reporting accurate hail sizes. We understand that by reducing the hail-size bins to three groups it potentially permits elimination of the more stringent quality-control criteria given the larger range in that hail-size bin. However, results suggest a notable change between 0.75–1.25-in reports and 1.5–1.75-in reports with reduction of the hail bins to three groups ignoring a real relationship difference. Further, these results support Donavon (2010) who notes a difference in radar-based thresholds and potential environments that support 1.75-in hail compared to 1.0-in sizes.

Lastly, further confusion arises with the additional QC process introducing a "6-12 h" range. "Building on the initial thresholds, this study defined events passing the first quality control steps but within 250 km and 6–12 h of the larger hail as those requiring additional inspection. Further, if any of these events were more than one bin size different then the group containing the largest hailstone, they were questionable enough to exclude." Does this mean that storms occurring 6 h apart, yet within 250 km of each other, were then reexamined, and for example if one storm produced 2-in hail and another 1.25 in, then the smaller storm was eliminated because it was more than one bin apart? How does the 12-h range fit into this and why wasn't this utilized in the initial QC? If I'm interpreting this correctly, I don't understand what made hail data questionable enough to exclude just because different storms 6 h apart produced different sized hail (bins)?

The interpretation is only partially correct. The existence of much larger hail (two or three bin sizes larger) within a short distance and time of a smaller report raises serious questions as to the uniqueness of the environmental setting associated with the much smaller report. As noted previously, some mitigation of this problem occurs through the initial quality control step. However, a narrower range of values exist with the two smallest hail bins and given the common object reporting bias noted by Baumgardt (2014), even small errors in hail size are more likely to result in environmental data from a smaller hailstone being incorrectly assigned to a particular bin. Based these issues, we had to strengthen the inclusion thresholds for the two smallest hail-size bins to maintain some level of uniqueness for the environments connected with the database reports.

As it stands, the current methodology remains insufficient in clarity and reproducibility, and at times is unnecessarily complex. I thank the authors in advance for addressing these questions and clarifying the quality control process, and considering changing their current methods of event inclusion and data elimination.

We believe the discussions above and extra clarity found in the revisions reduce much of the confusion.

The inverse relationship of increasing diameter hail size to decreasing MUCAPE in the HGZ described on P.12, P.19, and in Fig 10 may be partially explained by internal processes such as the effects of storm rotation. The vertical pressure gradient forces induced by this rotation enhance accelerations in the updraft

beyond environmental buoyancy alone, and Blair et al. (2011) showed increasing mid-level rotational velocity favored increasing hail sizes in the supercell storm mode.

The reviewer may be confusing this with THK_{HGZ} or $\%CAPE_{HGZ}$ that display an inverse relationship with hail size while in contrast $CAPE_{HGZ}$ increases only slightly as the hail-size bins increase. Nonetheless, the point is a good one as a strong and persistent rotating updraft with a non-hydrostatic vertical pressure gradient is essential for parcel accelerations that assist with size sorting and subsequently unfair competition for supercooled water. The problem from an environmental parameter standpoint is that shear and CAPE layers below 6 km that correlate well with stretching of ambient vorticity into the vertical, reveal little to no difference between a 1.75-in hail report and a 4.5-in hail report. This leaves very little in the pre-storm perspective that can help insinuate the ferocity of rotation and subsequent non-hydrostatic parcel accelerations of an updraft.

[Minor comments omitted...]

Third Review:

Recommendation: Accept.

General comment: I read through the author's comments and I'm fine with giving the paper my blessing for publication. The methodology from the start was unnecessarily complex and convoluted, but from what I can decipher, it has led to a manuscript sufficient for operational use. I think its reproducibility will still be challenging, but manageable. Ultimately, I believe it's in the best interest of everyone to push the paper through and let other scientists examine their results and effectiveness in an operational setting. Thanks again for the opportunity to review the paper!