1. Introduction

Severe hail accounts for the vast majority of severe weather-induced property loss in the United States, but our understanding of the geographic distribution and spatiotemporal variability of hailfall is limited due to numerous documented biases associated with hail reports (Kelly et al. 1985; Bardsley 1990; Witt et al. 1998b; Fraile et al. 1992; Changnon 1999; Schaefer et al. 2004; Doswell et al. 2005; Trapp et al. 2006; Ortega et al. 2009; Allen and Tippett 2015; Blair et al. 2017; Allen et al. 2017; Witt et al. 2018; Murillo and Homeyer 2019). There are a variety of factors that can introduce errors to hail reports, including (but not limited to) population density near a storm, concurrent weather hazards, and the use of reference objects to estimate size (Kelly et al. 1985; Allen and Tippett 2015; Allen et al. 2017; Witt et al. 2018). The largest and most quality-controlled U.S. hail report database produced by the Storm Prediction Center (SPC) is also limited by an exclusion of small hail [sizes below 19.05 mm (0.75 in.)] and an overall underrepresentation of severe hail frequency (Cecil 2009; Ortega et al. 2009; Allen et al. 2015b; Blair et al. 2017). These limitations prevent testing of objective hail identification methods for the full range of possible hail sizes, which has motivated multiple efforts to fill this data gap (Ortega et al. 2009; Elmore et al. 2014). For a more extensive summary of the myriad limitations and errors associated with hail reports, see Allen and Tippett (2015).

Despite these limitations, hail reports provide one of the most confident ground-truths of hailfall. Thus, numerous studies have evaluated U.S. hail reports from different sources and time periods [e.g., Allen and Tippett (2015) and references therein; Changnon 1999; Allen et al. 2015a,b; Childs and Schumacher 2019]. Overall, these studies have found a hailfall frequency maximum in the U.S. Great Plains, which shifts northward as summer progresses, and a secondary frequency maximum in the Southeast United States (SEUS).

Alternative approaches to producing hail climatologies include estimating hail size from radar and satellite observations, and statistically modeling severe hail occurrence from environmental observations or model analyses. A commonly used metric to estimate hail occurrence and size from radar is the single-polarization parameter maximum expected size of hail (MESH) from Witt et al. (1998a). Briefly, MESH represents the vertically integrated radar reflectivity in a storm to estimate the presence and depth of hydrometeors that have the expected scattering characteristics of hail stones. Previous studies have used MESH as the main hail identification method and compared their results to existing hail reports (Cintineo et al. 2012; Nisi et al. 2016b; Allen et al. 2017; Schlie et al. 2018; Ortega 2018; Nisi et al. 2016a; Murillo and Homeyer 2019; Tang et al. 2019; Allen et al. 2020; Warren et al. 2020). Key MESH-related results from several of these studies indicate that the parameter does not produce reliable hail size estimates [e.g., see detailed analysis by Bunkers and Smith (2013) and Ortega (2018)], and is more skillful at identifying hail occurrence rather than hail size.

One notable difference between the radar-based and report-based frequencies is seen the SEUS, where reports indicate a based frequencies is seen the SEUS, where reports indicate a local maximum that is not replicated in radar-based analyses. Though most other spatiotemporal features are similar between...
the two methods, with the highest hail occurrence observed throughout the Great Plains, the radar-based methods provided better coverage and more confidence in these distributions. Not only is the NEXRAD dataset available for a long time period, but it also provides three-dimensional information of storms in near–real time. However, despite radar-derived hail identification likely being the most documented of the hail report proxy methods, our knowledge regarding the best practices for objective hail identification using radar observations is still limited. In particular, studies using MESH have been carried out over a limited time period, lack extensive comparison with report distributions, or (in some cases) have only leveraged MESH as a secondary source of identification. Murillo and Homeyer (2019) recently evaluated numerous hail identification parameters to assess their ability to discriminate between severe hail producing and nonsevere hail producing storms and found that the original MESH, in addition to two revised MESH calculations, produced skillful storm discrimination. The new MESH configurations utilized a significantly larger number of reports than that used in Witt et al. (1998a) to define the original MESH configuration, warranting further investigation of their utility for larger spatiotemporal evaluations. Additionally, because Murillo and Homeyer (2019) focused on storm discrimination for various events between 2013 and 2017, several aspects of severe hail frequency and how the different MESH configurations represent a variety of events still need to be investigated.

Satellite observations have also been leveraged to develop global hail climatology. Several studies have established hail occurrence metrics utilizing spaceborne passive microwave imagery (e.g., Cecil 2009; Cecil and Blankenship 2012; Ferraro et al. 2015; Punge et al. 2017; Ni et al. 2017; Mroz et al. 2017, 2018; Bedka et al. 2018; Brueck et al. 2019). Earlier work found similar distributions to those diagnosed from reports and radar over the United States, though they noted that their methods likely produced overestimates in tropical regions and underestimate in midlatitudes globally. These biases were a result of differing tropopause heights between the two environmental regimes. Midlatitude overshooting convective storms often reside in the warmer, stratospheric environment as opposed to the colder, upper troposphere. More recent studies have made efforts to mitigate these biases and result in more consistent results with other hail climatologies (Bang and Cecil 2019). Although satellite observations display realized convection and typically feature large spatial domains, the low temporal and horizontal resolutions, coupled with the two-dimensional nature of microwave imagery make it difficult to discern vertical and small-scale characteristics of storms in these data. It is also difficult to establish a long-term record of satellite imagery due to the unique challenges associated with spaceborne platforms, especially the uncertain duration that a given platform remains operational.

Due to the limitations of previously mentioned direct and indirect methods of hail identification, environmental parameters from model output have been leveraged as a proxy for hail occurrence in past and future climates. Environments with favorable ingredients for convective storms and hail reaching the surface include various combinations of vertical wind shear, storm-relative helicity (SRH), and convective available potential energy (CAPE) (e.g., Johnson and Sugden 2014; Allen et al. 2015b; Barrett and Henley 2015; Mohr et al. 2015a,b; Prein and Holland 2018; Gensini and Tippett 2019; Tang et al. 2019). For example, Allen et al. (2015b) developed the first model to predict hail frequency over the United States by combining convective precipitation, 0–3-km SRH, 180-hPa mixed-layer CAPE, and mean specific humidity in the lowest 90 hPa of the atmosphere as hail identifiers and showed similar patterns to those seen in reports, albeit with some timing errors in the seasonal cycle. A more recent study by Prein and Holland (2018) focused on observable environmental parameters, including most unstable CAPE, melting level altitude, 0–3-km vertical wind shear and 0–3-km SRH. The overall U.S. spatial distribution and seasonal cycle of observed hail frequency was reproduced; however, the SEUS and regions on the lee side of mountains were biased low, and the Central Plains was biased high. Compared to satellite-based approaches, hail climatologies derived from environmental proxies benefit from the similarly large spatial domains with the added benefit of higher temporal resolution. One unique limitation, however, is the lack of known realization of these environments (i.e., how often they result in severe hail events).

When comparing the various methods for the objective of establishing a U.S. severe hail climatology with the greatest possible level of confidence, ground-based radar observations would appear to provide considerable strengths compared to other datasets that outweigh its weaknesses. Due to the extensive time period of data availability from the operational network in the United States and its ability to provide three-dimensional hydrometeor information of an existing storm in near–real time, we seek to quantify features of severe hailfall from radar using a markedly longer time period than previous studies to produce a climatology of 23 years. In particular, radar observations during the period 1995–2017 are used to create multiple MESH climatologies, which are compared to hail reports. In doing so, we seek to address three primary research questions: 1) how reliable is a radar-only hail climatology when compared to a report-based method? 2) what additional information (i.e., environmental variables) could improve accuracy of a radar climatology? and 3) what new information can we learn about severe hailfall over much of the contiguous United States (CONUS) from a long-term, well-calibrated, radar-based climatology? Potential severe hail occurrences are identified using radar observations, while environmental characteristics provide information regarding the likelihood that hail reaches the surface before melting.

2. Data and methods

a. Environmental data

Numerical model analyses provide environmental quantities for radar metric calculations and statistical analyses carried out in this study, for which we employ the Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2; Gelaro et al. 2017) for the 23-yr period. MERRA-2 is available from 1979 to the present with a 3-hourly temporal resolution, a 0.5° × 0.625° longitude–latitude resolution, and with 72 vertical
model levels with a model top of 0.01 hPa. Environmental parameters used for analysis include the altitudes of multiple isotherms, relative humidity (RH), low- to midtropospheric temperature lapse rates, vertical wind shear, and precipitable water (PWAT).

### b. Radar observations

To identify potential hail events, this study employs hourly, four-dimensional space-time composite NEXRAD data known as GridRad (Bowman and Homeyer 2017) that are available over most of the CONUS (the West Coast is not included, which rarely experiences severe hail), spanning the time period 1995–2017. Spatial and time weighted binning is used to merge individual NEXRAD scans in the common GridRad volume (described in greater detail in Homeyer and Bowman (2017) and references therein). These weights are such that shorter ranges and temporal proximity to the radar are prioritized, leading to retaining of data with the highest quality and resolution. Volumes are available on a ~2-km horizontal resolution and 1-km vertical resolution grid. The same or similar GridRad data have been used for hail detection in several recent studies (e.g., Murillo and Homeyer 2019; Tang et al. 2019; Jeong et al. 2020).

Quality control is applied at several steps from the native GridRad volumes. One of the more substantial filters focuses on the removal of low-confidence echoes consistent with the recommendations outlined in Homeyer and Bowman (2017). These data quality measures ensure that GridRad data retained for analysis contain high-confidence observations of storms. In addition, manual identification of failed radar scans from individual radars (and consequently, erroneous or non-physical echoes merged into the GridRad volumes) was carried out for the entire record to exclude such data from analyses in this study. Instances of uniform or linearly increasing radar reflectivity in range from a given radar are egregious examples of such erroneous scans. Subjective evaluation of maximum annual MESH was often the best indication of these events, which are more common in the earlier part of the record. In total, 200 694 hourly GridRad volumes are retained for analysis after completing these quality control measures (930 out of 201 624 volumes had erroneous scans that could impact MESH analysis, 78% of which were found in years prior to 2006).

MESH is a power law relationship of observed hail sizes to the radar reflectivity-based severe hail index (SHI), both of which are defined in Witt et al. (1998a). The original MESH relationship (hereafter MESHWitt) was empirically fit to the 75th percentile of 147 hail reports and has been broadly used in prior studies. However, Murillo and Homeyer (2019) refit the power law relationship between SHI and MESH to the 75th and 95th percentiles of observed hail size from nearly 6000 hail reports using GridRad data (hereafter MESH95 and MESH95, respectively). MESH75 and MESH95 are more sensitive to lower SHI values, which results in overall smaller SHI values satisfying a 1-in. size threshold (Murillo and Homeyer 2019). In this study, we generate a hail climatology using all three MESH configurations to test their performance with respect to hail reports. Based on the results of Murillo and Homeyer (2019), we expect that MESHWitt might underestimate hail occurrence in comparison to the new configurations because it was not calibrated using smoothed radar observations and was developed using a small sample of 147 hail reports, which were geographically confined to Oklahoma and Florida. MESH calculations using any of these fits may be affected by radar reflectivity biases from radar miscalibration, which have been documented to commonly reach 1 dB for NEXRAD radars (OFCM 2005) and may lead to an overestimate of MESH by up to 20% (Warren et al. 2020).

### c. Hail reports

The SPC Severe Weather Database (available at http://www.spc.noaa.gov/wcm/) provides the time, location, and size of severe hail reports in the United States (NOAA/NWS/SPC 1955). Though the severe hail criterion changed in 2010 from a diameter of ≥19.05 (0.75 in.) to ≥25.4 mm (1 in.), comparisons between the two severe hail day frequencies showed small differences in spatiotemporal distributions (Allen and Tippett 2015). Therefore, we utilize the current severe hail criterion of a diameter ≥ 25.4 mm (1 in.) for the full time period evaluated in this study. The criterion for significant severe hail for this period remained consistent and is defined as a diameter of ≥50.8 mm (2.0 in.). In total, 272 921 reports were used in this study.

### d. Data synthesis and analysis techniques

All datasets are analyzed on an 80 km × 80 km grid with 1° Gaussian smoothing applied spatially. The resolution was chosen based on the lowest-resolution dataset, the hail report database, following the same procedure as Brooks et al. (2003) and Doswell et al. (2005). This allows for spatial inhomogeneities of the population biased reports to be somewhat smoothed out for a climatological evaluation. This methodology allows for all data to be evaluated on a common grid, while appropriately utilizing the report observations. We then structure all data such that days are evaluated for the 24-h period from 1200 to 1100 UTC to be consistent with the SPC definition of a severe weather day, allowing for direct comparison between MESH and report distributions. For analysis, we retain the climatological maximum MESH size, and counts of severe and significant severe hail days, along with the hour of each severe and significant severe hail occurrence. These distributions are calculated over the full GridRad domain with reports and each MESH configuration to produce yearly and seasonal distributions. A report hail day is defined as at least one severe or significant severe hail report in a grid box during a 24-h SPC severe weather day. A MESH hail day is only defined when there is at least one severe or significant severe MESH observation in a grid box during a 24-h SPC severe weather day, and GridRad data are available for at least 20 of those hours to ensure that the comparison between radar and reports is fair. MESH days identified over the ocean are discarded due to reduced radar coverage and lack of validation (report) data. To be consistent with the radar analysis, report data analyzed was limited to ±10 min from each GridRad analysis time to account for potential/expected errors in report timing. There are two types of thresholds that we use to define hail days for MESH: 1) conventional thresholds for severe and
significant severe hail days, and 2) threshold values corresponding to the peak critical success index (CSI) for detecting severe and significant severe hail-producing storms, as found in Murillo and Homeyer (2019, see their Figs. 9a and 11a). This second method removes sensitivity of the result to the MESH relationship, essentially relaxing the identification to the integrated radar reflectivity (i.e., SHI value)—the resulting climatologies should be insensitive to the MESH fit used. Table 1 lists the MESH thresholds for all methods used in this study, for which a hail event is identified as a MESH value that reaches or exceeds the given threshold.

To address the second research question, we evaluate the ability of several environmental parameters in correctly distinguishing identified and false alarm hail days (see section 3a for the motivation to incorporate environmental characteristics). We focus on environmental discriminators that could produce storms with similar MESH values to those observed during hail events but dissimilar hail likelihood due to increased opportunity for melting before reaching the ground. The parameters we test are as follows: melting level (0°C) altitude, low-level (altitudes below the melting level) RH, low-level (altitudes below the melting level) lapse rate, 2–4-km lapse rate, PWAT, and 0–6-km bulk wind shear. To do so, we first group hail events into two categories: small hail (sizes of at least 0.75 in. and less than 1.5 in.) and large hail (sizes of at least 1.5 in.). Each of these two categories are further divided into two populations: 1) the “correct” hail identification group, defined at a gridbox level as observing both a MESH value and a hail report within the given size range within 10 min of each hourly GridRad analysis and 2) the “false alarm” group, defined as observing a MESH value within the given size range but no corresponding hail report. These size categories were chosen because false alarms are considered most likely in the lower range [given their relatively slow fall speeds—for hail—and as a result, greater times spent in the warm cloud layer (Heymsfield et al. 2014)]. Note that due to the aforementioned limitations of hail reports, these populations are not expected to be solely comprised of true hits or false alarms, but it is the best available approach for statistical evaluation. The correct and false alarm populations of MESH events are then used to 1) identify environmental factors that discriminate best between correct identifications and false alarms, and 2) use the leading discriminants as input for Fisher’s linear discriminant analysis (LDA) for two groups (Wilks 2006). Briefly, Fischer’s LDA solves for a function (or two-dimensional boundary in discriminant space) that best separates potential hail events into the two target populations (correct and false alarm) using linear combinations of the corresponding environmental quantities.

3. Results
a. Report and radar-based climatology

Though hail size comparisons can be useful for identifying large hail risk areas, there are several other limitations when diagnosing hail size extrema from radar or hail report observations. In addition to the reporting biases previously mentioned in section 1, the rarity of large hail, coupled with the unknown hail size distribution produced from a storm, pose unique challenges when interpreting hail size extrema (Cheng et al. 1985; Bardsley 1990; Fraile et al. 1992; Blair and Leighton 2012; Allen and Tippett 2015; Blair et al. 2017). The sensitivity to errors in radar observations, though minimized by the extensive quality control, could still result in an overestimation or underestimation of the maximum hail size. Given these limitations, hail magnitude comparisons are focused on assessing broad spatial maximums, rather than event specific hail sizes. The climatology from reports indicates that the largest hail sizes occur over the Central Plains (Fig. 1), as routinely documented in previous work. Using the 23-yr GridRad dataset, MESH$_{\text{Wet}}$ broadly underestimates the reported hail sizes, which are known to generally underrepresent true sizes (Fig. 2a). MESH$_{95}$ better represents reported hail sizes in most areas excluding the Central Plains, where it likely underestimates hail size with respect to reports (Fig. 2b). MESH$_{95}$ provides more reliable estimates over the Central Plains, but generally is calibrated slightly high otherwise, with more extreme overestimates across the Ohio Valley and in southern Florida (Fig. 2c).

Rather than evaluating individual hail reports, we analyze the frequency of hail days as suggested in Allen and Tippett (2015), which reduces the sensitivity to report errors, mitigates duplicate reporting, and is more relevant to impacts and losses. There is a maximum in severe hail report days per year over the Central Plains (Fig. 3). When using the severe threshold corresponding to the maximum skill (Table 1), we see similar
distributions and frequencies for all MESH identifications, which broadly agree with the reports (Fig. 4). However, as mentioned above, hail reports likely underestimate true hail occurrence, which indicates that this approach is also likely underestimating severe hail frequency. Thus, a more appropriate approach for capturing the majority of severe hail events and identifying differences in hail climatology due to varying MESH configuration is to apply the conventional threshold [i.e., diameter $\geq 25.4$ mm (1 in.)] to each (Fig. 5). This conventional threshold approach reveals considerable differences in severe hail frequency between the MESH configurations. The conventional MESH$\text{Witt}$ hail day climatology (Fig. 5a) shows similar geographic distribution and magnitude to that seen in the report climatology, with slightly higher frequencies than found in the climatology using the maximum skill threshold. The conventional MESH$\text{75}$ and MESH$\text{95}$ hail day climatologies result in 2–4 times more hail days per year than the corresponding max skill climatologies, with peak occurrence found over the Central Plains and SEUS (Figs. 5b,c).

The considerable differences seen between MESH$\text{Witt}$ and new MESH configurations in the SEUS is likely due to the reduced slopes of the new power law fits. Ultimately, the changes in the new MESH fits result in a greater sensitivity to lower SHI values and smaller sizes than MESH$\text{Witt}$, resulting in more severe hail days as the 25.4-mm (1-in.) diameter conventional severe hail threshold is more easily reached. When evaluating seasonality (Fig. 6), we see the largest discrepancy between the reports and MESH$\text{75}$/MESH$\text{95}$ in the SEUS occurs during the summer (JJA). This large (and presumably false) signal in the SEUS is likely a result of deep, tropical summer storms that achieve similarly high $Z_{HH}$ aloft as compared continental hailstorms. However, these storms typically lack the necessary thermodynamic and dynamic conditions for severe hail production and/or maintenance. From a radar perspective, storms with a large concentration of small, subsevere hail or a small concentration of large hail can produce similarly high radar reflectivity but unique dual-polarization scattering characteristics, which is impossible to assess using MESH-only identification approaches (e.g., see emerging work on storms with large accumulations of small hail; Kumjian et al. 2019). In the SEUS, many storms occur in a dynamic regime that is not supportive for severe hail growth but may be capable of producing small hail aloft. Due to the large warm layer depth (or similarly, high melting level height) in the SEUS during summer, these small hail stones melt before reaching the surface. Thus, leveraging environmental information to establish confidence in a diagnosed MESH severe hail day would help to limit impacts from nonsevere hail producing storms in the analysis, in the SEUS and elsewhere.

To assess diurnal characteristics, we compute the fraction of severe hail events [diameter $\geq 25.4$ mm (1 in.)] that occur before 0000 UTC (the boundary of diurnal transition), removing grid boxes with less than 50 events to ensure a large enough sample.
FIG. 4. The number of severe hail days per year as defined by the corrected threshold sizes for (a) MESH_{Witt} [diameter ≥ 35.56 mm (1.14 in.)], (b) MESH_{75} [diameter ≥ 41.91 mm (1.65 in.)], and (c) MESH_{95} [diameter ≥ 63.25 mm (2.49 in.)].

FIG. 5. The number of severe hail days per year using the conventional threshold [diameter ≥ 25.4 mm (1 in.)] for (a) MESH_{Witt}, (b) MESH_{75}, and (c) MESH_{95}. Note that the ranges in (b) and (c) are 2 and 4 times as large, respectively, as that seen in (a).
Reports over the Central Plains generally tend to be more frequent after 0000 UTC, with a transition to more frequent before 0000 UTC toward the eastern United States (Fig. 7). While MESH85 indicates a fairly uniform fraction of events before 0000 UTC across the CONUS, MESH75 and MESH95 exhibit timing distributions more similar to reports, though a majority of severe hail events in the eastern United States are occurring before 0000 UTC in comparison to reports (Fig. 8). This general pattern seen in reports, MESH75, and MESH95 is consistent with the findings from Krocak and Brooks (2020) that demonstrate similar timing characteristics when evaluating all types of severe weather.

b. Radar climatology with environmental parameter filtering

The results summarized in section 3a highlight that a severe hail climatology based on MESH alone does not adequately represent (or agree with) the frequency and size of severe hail as diagnosed from reports. Here, we address the second research question: What additional information could improve accuracy of radar climatology? Namely, we evaluate the utility of incorporating certain environmental quantities into severe hail diagnosis that have been previously identified as relevant to hail potential: melting level altitude, low-level RH, low-level lapse rate, 2–4-km lapse rate, PWAT, and 0–6-km bulk wind shear (Johnson and Sugden 2014; Allen et al. 2015a, 2020). To do this, we considered two criteria: 1) separation between the correct identifications and “false” alarms in a two-dimensional parameter space and 2) sensitivity to diurnal variability. The environmental parameters that show the greatest separation between the two populations and little sensitivity to the diurnal cycle were PWAT and 0–6-km bulk wind shear (Fig. 9d). Other parameter combinations were tested (Figs. 9a–c and others not shown).

Fig. 6. The number of severe hail days per year using the conventional threshold [diameter $\geq 25.4$ mm (1 in.)] during each season for (left) reports, (center) MESH75, and (right) MESH85. Seasons are represented by rows and labeled with each season’s month abbreviations. Note that the range is different depending on the season and type of hail identification.
shown) but either displayed larger overlap between the two populations or were sensitive to diurnal variations and produced vastly different peak timing of hail frequency.

Performing the LDA using PWAT and 0–6-km bulk wind shear produced greater separation between the two populations, further indicating “false” alarms tend to occur in regions with higher PWAT and lower 0–6-km bulk wind shear. We note that the two populations seen in both the pre and post LDA distributions are not expected to be completely composed of true hits or false alarms due to the numerous report limitations mentioned above. However, utilizing the large sample size and LDA provide a higher confidence in these distributions than that possible using radar data alone.

Applying LDA environmental filtering to the MESH climatologies leads to minimal differences in the maximum sizes. MESH\textsubscript{Witt} is nearly equivalent to the radar-only climatology (not shown). The large maximum for southern Florida in MESH\textsubscript{75} and MESH\textsubscript{95} in the radar-only climatology are suppressed after environmental filtering (not shown). Larger climatological differences are seen in the frequency analysis, with an environmentally filtered MESH\textsubscript{Witt} producing significant underestimates in hail days compared to reports (Fig. 10a). MESH\textsubscript{75} and MESH\textsubscript{95} severe hail day frequencies in the corrected climatology better represent the overall spatial distribution of severe hail days, with the previous presumably false maximum in the SEUS now absent (Figs. 10b,c), a result that was also found in Cintineo et al. (2012). Additionally, the resulting magnitudes are slightly higher than the reports, which is expected and is likely more representative of true hail days given the previously mentioned report biases. When evaluating hail days by season (Fig. 11), we see the expected Central Plains maximum during the spring and summer, which is consistent with the report climatology. MESH\textsubscript{95} better captures extremes, especially in the cool season and for regions with nonzero but rare hail occurrence (e.g., Arizona).

Hail event timing in MESH\textsubscript{Witt} was mostly unchanged after performing the LDA. This indicates that this MESH approach does not accurately capture the diurnal variability of hail events that is apparent in both reports and other MESH configurations (Fig. 12a). The large eastern region of frequent daytime hail events seen in both MESH\textsubscript{75} and MESH\textsubscript{95} decreased after the environmental filtering (Figs. 12b,c) while still agreeing with results from Kroack and Brooks (2020).
Most of the SEUS features a similar number of events before and after 0000 UTC.

Using this filtered hail climatology, we compare MESH75 and MESH95 to the original report distribution to assess the overall spatiotemporal variability and address our third research question: What new information can we learn about U.S. hailfall from a long-term, well-calibrated, radar-based climatology? Apart from the previously mentioned differences in frequency between the annual filtered MESH and report distributions, there are notable differences in locations of frequency maxima both in the annual and seasonal geographic distributions (Figs. 3 and 10b,c). First, the SEUS frequency maximum (stretching mostly in the east–west direction from northeast Mississippi to the western portions of the Carolinas) is only present in the report climatology, which indicates possible overreporting in this region noted in several previous studies (Cintineo et al. 2012; Allen et al. 2015b; Allen and Tippett 2015). This high bias could be due to various factors such as the high population density in the SEUS, resulting in a higher likelihood of severe hail being reported compared to adjacent locations. Nonstationarities identified by Allen and Tippett (2015) in this region in the spatial frequency of reporting also suggest that this region may be subject to a greater fraction of reports being driven by warning verification. Perceptual biases as to the size of SEUS hail may also be a factor, because this region infrequently receives hail much larger than 25.4 mm (1 in.). Second, the Central Plains frequency maximum extends slightly further into southwest Texas and east Colorado in the filtered MESH than the report climatology indicated. These shifts could be a result of better capturing hail occurrence in regions with low population density and overreporting in this region noted in several previous studies (Cintineo et al. 2012; Allen et al. 2015b; Allen and Tippett 2015).

![Fig. 9](image-url)

**Fig. 9.** Two-dimensional joint frequency distributions of correct identifications (blue) and “false” alarms (red) for precipitable water and 0–6-km bulk wind shear. Populations are split into two size groups as follows: (a),(c) limited to sizes between 0.75 and 1.5 in. and (b),(d) limited to sizes greater than 1.5 in. Populations in (a) and (b) are generated through report verification, and those in (c) and (d) are generated through LDA environmental filtering. The contour values represent the percentage of all correct identifications (blue) or “false” alarms (red) that exist within the parameter space.
density. The independence of the filtered radar-based climatology of this bias indicates greater severe hail occurrence than previously estimated via radar methods.

The filtered seasonal hail climatology seems to indicate that filtered MESH75 produces a more realistic climatological frequency distribution, while filtered MESH95 generates frequencies closer to a worst case scenario (Fig. 11). There is also a clear seasonal shift in the peak frequency of hail. The winter is characterized by predominantly SEUS events, while the rest of the year sees a shift toward different regions of the Central Plains. There are also frequency maxima in Idaho and upstate New York during winter when using MESH95, which are mostly a result of false detections of a few heavy winter precipitation events. The higher sensitivity to lower hail sizes in MESH95, combined with bright band contamination during these winter events, likely contributes to the false signal over these regions.

To briefly evaluate the yearly changes of the multiple MESH configurations, we accumulate all the grid points with severe hail days for each year (Fig. 13). Data were quite sparse prior to 2000 and showed increasing trends in frequency until 2006, after which data availability is roughly constant from the NEXRAD network. From that point forward, natural variability dominates, and frequency on the national scale appears to be fairly stable, consistent with findings from Allen and Tippett (2015). The relative severe hail day frequency between each unfiltered and filtered MESH also stays rather constant, indicating that the filtering process works uniformly. When assessing each MESH configuration, we find that MESHWitt is the least useful, since it is insensitive to filtering, does not adequately demonstrate observed spatial patterns, and underestimates the frequency of severe hail days. Conversely, both MESH95 configurations likely overestimate the frequency, given the considerable increase from that seen in reports. MESH75 is likely the most useful for a climatological approach, as it displays a moderate increase in frequency with respect to reports and demonstrates reasonably consistent interannual variability (after 2006). Because of the lower radar data quality/completeness before 2006, frequency analyses (Figs. 10 and 11) likely underrepresent the overall true frequency of hail events diagnosed by radar over the 23-yr period. However, restricting MESH analyses to 2006–17 where data quality are much more consistent reveals consistent spatial distributions compared to the full period (not shown).

4. Discussion and conclusions

This study developed a new U.S. severe hail climatology based on 23 years of ground-based radar observations and environmental characteristics from reanalysis data in an effort to answer our three research questions: 1) How reliable is a radar-only hail climatology when compared to a report-based method? 2) What additional information (i.e., environmental variables) could improve accuracy of a radar climatology? 3) What new information can we learn about contiguous United States (CONUS) severe hailfall from a long-term, well-calibrated, radar-based climatology? Our conclusions are as follows:

![Fig. 10. The number of severe hail days per year, after environmental filtering, using the conventional threshold (diameter \( \geq 25.4 \text{ mm (1 in.)} \)) for (a) MESHWitt, (b) MESH75, and (c) MESH95. Note that the ranges in (b) and (c) are 2 and 4 times as large, respectively, as that seen in (a).](image-url)
1) Hail climatologies based on radar observations alone are insufficient for reproducing the broad characteristics of hail report distributions, and likely also true hailfall distributions as well.

2) MESH configurations developed to represent a large number of reports are more sensitive to marginal severe hail-producing storms. This results in large frequency maxima across the SEUS, particularly during JJA, which likely indicate strong, tropical summer convection that have similar $Z_H$ to continental hailstorms, but lack environmental conditions necessary for severe hail occurrence.

3) PWAT and 0–6-km bulk wind shear show the best discrimination between correct identifications and “false” detections of severe hail occurrence.

4) Performing LDA to filter out environments that are not favorable for severe hail occurrence improves the agreement between radar climatology and report climatology.

5) U.S. severe hail climatology based on well-calibrated radar and environmental characteristics shows overall larger frequencies and one single hail frequency maxima through the Plains and Ozarks.

6) Hail events throughout the Central Plains are more frequent before 0000 UTC, while those in the SEUS are similarly as frequent before and after 0000 UTC.

Though various methods exist for using radar observations to identify likely areas of severe hailfall, they alone do not provide enough information to accurately establish a severe hail climatology. $MESH_{W_{min}}$, the commonly used radar-derived parameter, was fit to only 147 reports, which resulted in an
unreliable relationship between hail size and SHI [see Fig. 13 of Murillo and Homeyer (2019)]. The increased sensitivity of MESH\textsubscript{75} and MESH\textsubscript{95} to lower SHI values, though more accurately representing the reported hail size-SHI relationship, results in greater “false” detection in regions of deep, tropical convection with environments that are not supportive for severe hailfall. However, it remains that one of the greatest challenges for objective severe hail identification is the uncertainty associated with “false” detection. This limits the ability to accurately measure the skill of any parameter, given the unknown number of hail occurrences left unreported. Even so, certain techniques such as LDA can be adopted to help minimize potential bias. There are additional challenges associated with radar-derived products, including the variable number of operational radars throughout the time period, potential erroneous radar scans missed during quality control, the unknown time- and space-varying calibration biases of the radars, application of Gaussian smoothing to regridded data, and the resolution of the radar observations. For example, high MESH values could be spatially overestimated due to nonuniform beam filling of high reflectivity, which was also discussed in Cintineo et al. (2012). The overall observational record length, stability in radar data availability since 2006, reduction in spatial resolution, and insensitivity to sigma smoothing are all aspects of the data or analysis that serve to mitigate these issues. It is also worth noting that we tested the sensitivity of isotherm altitude used in MESH calculations to different reanalysis temperature fields and found that MESH was insensitive to these changes.

The U.S. severe hail climatology established in this study produces a solid foundation that can be leveraged for a variety of applications, including its utility as a validation dataset, the ability to evaluate how hail events covary with climactic variability, and the opportunity to apply similar LDA evaluations to improve model output-based hail proxy methods. Although we only utilized single-polarization NEXRAD observations here, these results provide valuable information for the next
decade as a sufficient database of dual-polarization observations becomes established.

Acknowledgments. This work was supported by the National Aeronautics and Space Administration (NASA) under Awards NNX15AV81G and 80NSSC19K0347.

Data availability statement. All GridRad radar data used for analysis are publicly available at the National Center for Atmospheric Research Research Data Archive (Bowman and Homeyer 2017). Report data are publicly available from the SPC (NOAA/NWS/SPC 1955). MERRA-2 reanalysis data used are publicly available from NASA (Global Modeling and Assimilation Office (GMAO) 2015). Derived quantities based on these datasets are available from the authors upon request.

REFERENCES


