Deriving Severe Hail Likelihood from Satellite Observations and Model Reanalysis Parameters using a Deep Neural Network

Benjamin Scarino,[a] Kyle Itterly,[a] Kristopher Bedka,[b] Cameron R. Homeyer,[c] John Allen,[d] Sarah Bang,[e] Daniel Cecil[e]

[a] *Analytical Mechanics Associates, Hampton, VA*

[b] *NASA Langley Research Center, Hampton, VA*

[c] *School of Meteorology, University of Oklahoma, Norman, OK*

[d] *Central Michigan University Institute for Great Lakes Research, Mount Pleasant, MI*

[e] *NASA Marshall Space Flight Center, Huntsville, AL*

Corresponding author: Benjamin Scarino, benjamin.r.scarino@nasa.gov

Early Online Release: This preliminary version has been accepted for publication in *Artificial Intelligence for the Earth Systems*, may be fully cited, and has been assigned DOI 10.1175/AIES-D-22-0042.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

ABSTRACT

Geostationary satellite imagers provide historical and near-real-time observations of cloud top patterns that are commonly associated with severe convection. Environmental conditions favorable for severe weather are thought to be represented well by reanalyses. Predicting exactly where convection and costly storm hazards like hail will occur using models or satellite imagery alone, however, is extremely challenging. The multivariate combination of satellite-observed cloud patterns with reanalysis environmental parameters, linked to Next Generation Weather Radar- (NEXRAD-) estimated Maximum Expected Size of Hail (MESH) using a deep neural network (DNN), enables estimation of potentially severe hail likelihood for any observed storm cell. These estimates are made where satellites observe cold clouds, indicative of convection, located in favorable storm environments. We seek an approach that can be used to estimate climatological hailstorm frequency and risk throughout the historical satellite data record.

Statistical distributions of convective parameters from satellite and reanalysis show separation between non-severe/severe hailstorm classes for predictors including overshooting cloud top temperature and area characteristics, vertical wind shear, and convective inhibition. These complex, multivariate predictor relationships are exploited within a DNN to produce a likelihood estimate with a critical success index of 0.511 and Heidke skill score of 0.407, which is exceptional among analogous hail studies. Furthermore, applications of the DNN to case studies demonstrate good qualitative agreement between hail likelihood and MESH. These hail classifications are aggregated across an 11-year GOES-12/13 image database to derive a hail frequency and severity climatology, which denotes the Central Plains, the Midwest, and northwestern Mexico as being the most hail-prone regions within the domain studied.

File generated with AMS Word template 2.0

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-22-0042.1.

1. Introduction

Deep convection is ubiquitous across the globe and integral to the Earth's climate system. It is responsible for redistributing heat, moisture, and momentum, as well as transporting water vapor into the stratosphere. Deep convection produces beneficial rainfall but can also generate severe weather conditions that adversely impact society, such as damaging wind, large hail, tornadoes, lightning, flooding rainfall, and aviation icing and turbulence (Yost et al. 2019; Mecikalski et al. 2021). Hail damage produces approximately 60% of the average annualized loss of the three primary severe weather hazards in the United States – straight line winds, hail, and tornadoes (Gunturi and Tippett 2017). Promoting resilience against such hazards on local and global scales is a primary goal of the NASA Applied Sciences Disasters program, which seeks to encourage use of satellite observations to quantify and mitigate risk (NASA 2021).

Extensive research, using rawinsonde, numerical weather prediction, and reanalysis profiles in environments near severe weather events, indicates that hazards produced by a storm as well as storm structure are a by-product of its thermodynamic and kinematic environments. Parameters such as convective available potential energy (CAPE), vertical wind shear, mid-tropospheric lapse rate, lower tropospheric mixing ratio, and statistical combinations of these and many other predictors have been found useful for developing climatologies of and forecasting severe storms (e.g., Brooks et al. 2003; Johnson and Sudgen 2014; Púčik et al. 2015; Prein and Holland 2018; Taszarek et al. 2020). Severe storm environments are often complex and can change significantly in time and space, with variations that may not be adequately captured by observations and models (Coniglio and Parker 2020; Coniglio and Jewell 2022). These complexities lead to some disagreements in reanalyses such as the Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2; Gelaro et al. 2017) and the fifth-generation ECMWF global atmospheric reanalysis (ERA5; Hersbach et al. 2020), especially near topographic boundaries (Taszarek et al. 2021). All storm cells that form in any given favorable severe storm environment do not necessarily produce severe weather, nor will storms in close proximity produce the same weather hazards, or hazards of the same intensity. Observations of storms from remote sensing instruments, such as ground-based radars, satellite imagers, and lightning detection networks, are therefore needed to improve differentiation of severe from non-severe storm cells.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-22-0042.1.

Studies have shown that severe storms exhibit a variety of signatures in satellite observations near in time to reported severe weather events. Cold, textured cloud tops embedded within a warmer, smooth cirrus anvil indicate the presence of strong updrafts that have penetrated through the level of neutral buoyancy, and often into the stratosphere. These overshooting cloud tops (OTs) can be detected and characterized in terms of their tropopause-relative infrared (IR) brightness temperatures (BT), BT difference (BTD) between the OT and surrounding anvil, OT area, and height using automated methods (e.g., Marion et al. 2019; Khlopenkov et al. 2021). The presence of an OT is statistically correlated with severe weather reports (Dworak et al. 2012; Bedka and Khlopenkov 2016; Bedka et al. 2018; Khlopenkov et al. 2021). Additionally, ice particles within intense convection and hailstorms scatter microwave radiation emitted by the surface before it can reach a satellite sensor, leading to BT depressions observed by passive microwave imagers such as the Tropical Rainfall Measurement Mission (TRMM) Microwave Imager (TMI), the Global Precipitation Measurement Mission (GPM) Microwave Imager (GMI), and the Advanced Microwave Scanning Radiometer 2 (AMSR2) onboard the Global Change Observation Mission – Water (GCOM-W1) satellite. Hailstorms generate notable BT depressions in lower frequency (10-37 GHz) passive microwave observations. These multispectral signals have been recently combined with human spotter hail reports to derive hail probability for convective feature detections (Bang and Cecil 2019; Bruick et al. 2019; Bang and Cecil 2021).

The Geostationary Operational Environmental Satellite (GOES) series has observed IR BT at 4-km/pixel nadir resolution, sufficient for resolving and characterizing OTs, since the launch of GOES-8 in 1994. The GOES-8 to -15 series have collected imagery, typically at 15-minute intervals over North America, useful for determining the climatology of severe storms throughout the diurnal cycle over much of the Western Hemisphere at much higher spatial resolution than is possible from reanalyses. OT detection climatologies have been previously used to define hailstorm risk for application within the reinsurance industry (e.g., Punge et al. 2017, 2023). The GOES-R satellite series, beginning with GOES-16 that provided initial pre-operational imagery in 2017, observes IR BT at 2-km/pixel nadir resolution and 5-minute intervals over North America, and 10- to 15-min intervals over a hemispheric (i.e., "full-disk") field of view with its Advanced Baseline Imager (ABI). This improved ABI image quality represents a discontinuity in the GOES climate data record because spatial resolution has a significant impact on cloud top appearance and apparent intensity that can affect severe storm detection algorithm performance (Khlopenkov et al. 2021; Cooney et al. 2021).

Although severe weather can be generated by storms with OTs, many OT-producing storms do not generate such weather because the environment is not conducive due to deficient or excessive wind shear, unfavorable hodograph shape, and insufficient moisture or instability. Nevertheless, studies analyzing limited samples of data suggest that OTs with colder IR BT and larger area are more likely to be associated with heavy rainfall and severe weather (Dworak et al. 2012; Marion et al. 2019; Sandmael et al. 2019; Khlopenkov et al. 2021). Knowing the extent to which these findings hold, however, requires more rigorous analysis, e.g., examining the influence of the regional meteorological environment. This suggests that a combination of storm environmental conditions and satellite observations could be used to better depict the climatology of severe storms (Cintineo et al. 2018; 2020).

In addition to complexities associated with severe storm detection, knowledge of exactly when and where severe weather has occurred is often lacking. Deep convection can evolve rapidly, and uncertainties in human spotter report timing and location can confound our ability to develop statistical relationships between cloud properties and severe weather occurrence. Spotter hail reports, for example, are known to be biased toward higher population density, daytime hours, and regions with greater road density (Allen and Tippett 2015; Ortega 2018; Elmore et al. 2022). Hail sizes are frequently underreported due to rapid melting, are mis-reported due to perceived size correlation with objects of known size (e.g., golf ball), and hail events coincident with tornadoes are often omitted (Allen and Tippett 2015). The use of hail diameter as a proxy for storm intensity, which fails to sufficiently describe the mass or density of hailstones and is difficult to measure accurately, presents another fundamental uncertainty in hail reports (Doswell et al. 2005; Allen and Tippet 2015). To offset the limitations in observations, the Maximum Expected Size of Hail (MESH), derived from a vertical integration of radar reflectivity at horizontal polarization above the freezing level, is a commonly employed metric for estimating hail occurrence and size that can be used to mitigate uncertainties with reports (Witt et al. 1998; Murillo and Homeyer 2019). MESH has been used to derive high spatial resolution hail climatologies over the contiguous United States (CONUS; Cintineo et al. 2012; Murillo et al. 2021; Wendt and Jirak 2021), consistent with best-available ground report climatologies (Allen and Tippet, 2015; Taszarek et al. 2020).

Long-term GOES satellite imager and MESH climatologies over the CONUS provide a new opportunity to explore satellite-derived statistical properties of hailstorm cloud tops, the environmental characteristics supportive of hailstorms depicted by reanalyses, and the detectability of hailstorms using GOES satellite, reanalysis, and a combination of the two datasets. In this paper, we analyze an 11-year record of GOES-12/13 data, as well as a period during the 2017 warm season (April-August) when GOES-13 and GOES-16 provided nearly coincident measurements during the GOES-16 pre-operational checkout period. We focus on the GOES-12/13 data here to demonstrate how hailstorms could be detected using the nearly 24-year GOES-8 to -15 data record – satellites that carried imagers with consistent spatio-temporal resolution. The NASA MERRA-2 and ECMWF ERA5 reanalyses are analyzed together to determine the sensitivity of hailstorm detectability to the choice of reanalysis, as well as other input variations. GOES and reanalysis data are combined and aligned with MESH-identified convective cells within a deep neural network (DNN) to determine the feasibility of discriminating potentially severe hailstorms from sub- (or non-) severe hailstorms. The relative importance of various GOES and reanalysis parameters for severe hailstorm discrimination was assessed, and these form the basis of the final DNN. With emphasis on satellite-identified storm signatures and their properties, this approach to estimating likelihood for severe hail builds on similar deep learning/machine learning efforts that tied radar- and model-identified cell features and lightning detection to hail reports, MESH, severe weather warnings, or a simulated hail size reference through use of a regularized regression model, a convolutional neural network, or Random Forest (RF) classifier and RF regression models (Gagne et al 2017; Czernecki et al. 2019; Gagne et al. 2019; Burke et al. 2020; Gensini et al. 2021; Mecikalski et al. 2021). We show that a likelihood for potentially severe hail can be estimated with relatively high accuracy using a combination of GOES and reanalysis data, exceeding the predictive skill of a DNN model that uses reanalysis or GOES inputs alone. Note that terms like *predictive*, and different forms of the word, are in the context a DNN output, which is commonly called a *prediction*, and should not be confused with a meteorological *forecast*. This work aims to develop a DNN model that can estimate severe hail likelihood for satellite-identified convective signatures as an initial step toward global application, which would especially benefit regions without a long-term database of weather radar observations.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-22-0042.1.

2. Data

a. GOES convective cloud top characteristics

Automated detections of OT and anvil characteristics are derived from 10.7-µm GOES-12/13 and 10.3-µm GOES-16 IR BTs using algorithms described by Khlopenkov et al. (2021) and validated/analyzed by Cooney et al. (2021). GOES IR BT observations are first resampled to a grid and then converted to tropopause-relative values (IR–Tropopause) by differencing the IR BT and a tropopause temperature derived from the MERRA-2 reanalysis. This measurement is referred to as a “BT-score”. Tropopause temperature is spatially smoothed and interpolated in time/space to the satellite image time and pixel grid. The IR–Tropopause values up to 35 K warmer than the tropopause are further analyzed as possible anvils and OTs. Local BT-score maxima within anvils, referred to as embedded cold spots (ECSs) throughout this paper, are then identified and serve as OT candidates. Note that the absolute BTD between the ECS and the anvil (ECS–Anvil BTD) can be 1 K at minimum (ECS colder than the anvil), so hailstorms without an ECS detection have little to no IR BT variation within their anvils. Derivations of IR anvil rating, anvil cloud extent surrounding each ECS, ECS area, and the ECS OT Probability are further explained by Khlopenkov et al. (2021). ECS area is defined here as the cumulative area of pixels corresponding to a unique ECS ID after parallax correction, with the following additional condition: All ECS pixels must be colder than a factor of 0.75 of the ECS–Anvil BTD added to the coldest pixel's IR BT, which ensures that only relatively prominent portions of each ECS are considered in the area calculation.

GOES-12/13 OT detections during 2007-2017 from 23-50º N and 65-115º W, a region that includes most of the CONUS, sub-tropical convectively active regions over northern and western Mexico, the Gulf of Mexico, and the Gulf Stream, are matched with convective cells defined using hourly NEXRAD MESH. GOES-12/13 scanned the CONUS at approximately 15-min intervals during normal operations and at 7.5-minute intervals when operated in rapid scan mode during high-impact weather events. As described in Section 3.b below, GOES-12/13 data can be matched with MESH convective cells when collected within ±7.5 minutes from the hour. No matches were possible when full-disk scanning limited imaging frequency to every 30 mins, which occurred every three hours (00, 03, 06, ..., 21 UTC). GOES-16 full-disk scans from April-August 2017, during its pre-operational period, were also analyzed at 15-minute scan intervals to provide consistency with typical GOES-12/13 operational

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-22-0042.1.

scanning. Additionally, GOES-16 1-min Mesoscale Domain Sector data were subsampled to emulate the CONUS 5-min resolution for case study analyses. GOES data were acquired from the University of Wisconsin-Madison Space Science and Engineering Center via the McIDAS-X software package (Lazzara et al 1999).

b. Maximum Expected Size of Hail

The NOAA NEXRAD network collects volumetric scans of horizontal reflectivity factor at many elevation angles, which are combined to generate the MESH product. Percentiles of the unfiltered MESH are derived by combining isotherm heights from the National Centers for Environmental Prediction Rapid Refresh model with echo top heights from the 3-D Gridded NEXRAD WSR-88D radar (GridRad) dataset, which are then scaled to observed hail size reports using an empirically derived power law fit. Specifically, observed 40+ dBZ echo top heights above the freezing level are integrated and scaled to derive a Severe Hail Index (SHI). Then, a power-law fit is empirically formulated to fit the SHI using the 75th and 95th percentile of ground hail reports, resulting in the 75th and 95th percentile MESH (MESH75 and MESH95; Homeyer and Kumjian 2015; Homeyer and Bowman 2017; Murillo and Homeyer 2019).

Localized MESH maxima associated with convective cells (see Section 3.a) are identified from hourly, 2-km GridRad MESH over the CONUS (25-49° N; 69-115° W; Bowman and Homeyer 2017). These cells are used as the basis for aggregating GOES and reanalysis data for training and validating the DNN. Case studies shown in this paper take advantage of a recently available 5-minute “GridRad Severe” MESH dataset (University of Oklahoma 2021).

c. Ground spotter storm reports

Reports of hail diameter exceeding 1 inch (25.4 mm) were obtained from the SPC Severe Weather Database (https://www.spc.noaa.gov/wcm/) from 2007 January through 2017 December (42,040 total reports). We examine GOES and reanalysis properties as a function of reported hail size, along with MESH and passive microwave hail probabilities from TMI, GMI, and AMSR2. As noted above, hail size reporting suffers from several inadequacies, compounded with uncertainty regarding the locations of sub-severe or non-hailstorms. These shortcomings inspire use of GridRad MESH as a hail proxy.

d. Passive microwave hail detections

Another proxy for hail occurrence is derived from satellite passive microwave radiometer (MWR) data. The presence of ice particles in an atmospheric column scatters outgoing microwave radiation emitted by the Earth's surface, the magnitude of which is dependent on the particle size, number concentration, and the wavelength band being observed (Cecil 2009). Such scattering results in a lowering of MWR multispectral BTs. Precipitation features, i.e., a set of contiguous MWR pixels with low BTs thought to be caused by convection, are first constructed (Nesbitt et al. 2000; Cecil et al. 2005; Liu et al. 2008). The features are defined as contiguous areas with an 85- (TMI) or 89- (GMI and AMSR2) GHz BT below 200 K. A probability of severe hail algorithm was developed based on an empirical relationship between SPC severe hail reports and the multispectral MWR BT (Bang and Cecil 2019; 2021). The precessing orbits of the TRMM and GPM satellites enable observations of deep convection at all hours, thereby mitigating observation time biases, whereas the GCOM-W1 AMSR2 instrument flies in a sun-synchronous orbit and collects observations at only 01:30 AM/PM local time. Microwave hail detections have been used to construct global climatologies using a 2ºx2º grid, normalized by sampling opportunities, which reveal generally good spatial agreement with climatologies derived from MESH and ground hail reports over CONUS (Bang and Cecil 2019; 2021). These MWR hail measurements provide an alternative perspective of hailstorm intensity that is useful for comparing to the relationships found from GOES/reanalysis and hail size metrics produced from MESH and the SPC.

e. Global reanalyses

Two state-of-the-art atmospheric reanalyses are employed to explore hail likelihood estimation sensitivity to meteorological variables derived from models with distinct parameterizations, means of data assimilation, and spatio-temporal resolutions. Despite the rather coarse resolution relative to convective-scale processes, global reanalyses have been used in numerous studies to identify large-scale environmental conditions favorable for severe weather, and to assess predictability of severe weather (Brooks et al. 2003; Brooks 2009; Punge et al. 2014; Allen et al. 2015; Punge et al. 2017; King and Kennedy 2019; Taszarek et al. 2020; Gensini et al. 2021). Validation of these reanalyses against observed soundings suggests strong performance over the CONUS for both thermodynamic and kinematic parameters, particularly outside of the boundary layer (Taszarek et al. 2021).

Additionally, the continuous global coverage offered by reanalyses provides a framework to apply the model to regions beyond the CONUS. A set of environmental parameters thought to be useful for identifying hailstorm environments, described in Section 3.c, were derived from the ERA5 and MERRA-2 reanalyses using the Python package, xcape (Lepore et al. 2021; https://github.com/xgcm/xcape).

- **ERA5**: The latest global reanalysis product from the European Centre for Medium Range Weather Forecasting (ECMWF) with horizontal grid spacing of 0.25ºx0.25º, 137 vertical levels, and a 1-hour output interval (Hersbach et al. 2020)
- **MERRA-2**: The latest global reanalysis product from NASA's Global Modeling and Assimilation Office (GMAO) with horizontal grid spacing of 0.5ºx0.625º, 72 vertical levels, and a 3-hour output interval (Gelaro et al. 2017)

3. Methods

An 11-year record of GOES-derived cloud top properties for ECSs, spatially and temporally aligned with MESH convective cells, model reanalysis parameters, SPC hail size reports, and MWR hail probability, was assembled using methods described below. We seek to identify cloud-top patterns and environmental parameters consistent with potentially severe hailstorms. A DNN allows for discovery of less intuitive, multi-dimensional input combinations of such parameters, which leads to enhanced predictive capability. These predictors and their roles in various DNN input sensitivity experiments are defined, as are the DNN architecture and evaluation metrics.

a. MESH cell feature database

A database of cell object characteristics is assembled by optimizing the open-source Python package, Tracking and Object-Based Analysis of Clouds v1.2 (tobac), to detect local MESH95 maxima exceeding 10 mm (Heikenfeld et al. 2019). The location and maximum value of MESH95 for each feature is stored. Feature maxima must be spaced by at least 28×28 km^2 (7x7 GOES-12/13 pixels) to facilitate extraction of GOES parameters around each MESH95 maximum. All settings used within tobac v1.2 are listed in Table 1.

tobac feature detection setting	Value used
Minimum distance between features	28,000 m
Gaussian smoothing	None
Minimum pixels per object	1
Position threshold	Extreme
Target	Maximum
Erosion threshold	0
MESH95 thresholds	10, 25, 50 mm
2D watershed segmentation thresholds	10, 25, 50, 75 mm

Table 1. Feature detection and watershed thresholds used for MESH cell detection in tobac.

The ECS matching requirement is expected to exclude less confident MESH detections while anchoring DNN predictions to satellite-observed storm cells. Non-severe MESH95 cells (<1.5-inch, or 38.1-mm, diameter; see Section 3.d.2) match with an ECS about 25% of the time, indicating that such cells either have warm cloud tops or tops with little IR BT variation (Table 2). Cells with higher (1.5+ inch) MESH95 diameter, a value corresponding to *potentially* severe hail, have a 59% match rate with an ECS. The match rate increases to 73% for 2+ inch (50.8+ mm) MESH95, indicating that cells with larger estimated hail size have greater BT variation in their anvils. Furthermore, despite a ~5:1 dominance of warm season to cold season matches, the rather consistent matching frequencies suggest that the relationship between MESH95 and ECS occurrence is largely independent of seasonality, except for a slight increase in incidents of significantly large hail in the warm season.

MESH95 diameter	ECS match frequency		
	GOES-12/13 all seasons	GOES-12/13 cold season	GOES-12/13 warm season
Non-severe, <1.5 inches	25%	24%	25%
Potentially severe, 1.5+ inches	59%	59%	59%
Potentially significant severe, 2+ inches	72%	69%	73%

Table 2. Matching frequencies of non-severe, potentially severe, and potentially significant severe MESH95 signals with ECSs detected by GOES-12/13.

b. GOES convective cloud top property database

A set of GOES-12/13 cloud top properties is produced at ECS locations from every available scan (371,489) during 2007-2017 (Table 3). The same parameters are derived from GOES-16 data during the 2017 warm season (April-August). Properties, such as Mean Anvil Height and Area of Cold Cloud, are computed using pixels around each ECS. Pixels are parallax-corrected based on a cloud top height retrieval derived from matching the ECS IR BT with the co-located MERRA-2 temperature and geopotential height profiles. Only data for the most intense GOES-12/13 pixel (defined by minimum IR–Tropopause) is stored for cases in which parallax causes two pixels to be assigned to the same gridded satellite pixel. GOES ECS regions detected within ±7.5 minutes and 28×28 km^2 (30×30 km^2 for GOES-16) of each MESH95 cell maxima are compiled. Note that during periods of GOES-12/13 rapid scan operation, it is possible to have more than one observation within each 7.5-min window. The peak intensity of the multiple matches is used to accumulate every predictor in Table 3. The satellite and reanalysis parameters (Section 3.c) are stored alongside MESH data and other positional variables.

Predictor short name	**Description**	**Units**
OT Probability	Overshooting top probability (Khlopenkov et al. 2021)	unitless (0.0-1.0)
ECS–Anvil BTD	Brightness temperature difference between coldest pixel and mean anvil background	K
ECS Area	Area of pixels for each ECS region	km^2
IR–Tropopause	GOES IR brightness temperature – MERRA-2 Tropopause temperature	K
Anvil Height	Average cloud top height of pixels with IR anvil rating ≥ 20 within ~30 km^2	km
Area of Cold Cloud	Average area of pixels with IR BT< 225 K within ~30 km^2	km^2
Anvil Frequency	Percentage of pixels with IR anvil rating ≥ 20 within ~30 km^2	%
Cloud Top Height	Derived from IR BT matched with MERRA-2 temperature profile. OT regions are height-assigned using a constant lapse rate assumption from Griffin et al. (2016)	km

Table 3. Satellite IR derived cloud top parameters evaluated in this study.

A snapshot of several GOES-12/13 and MESH properties are shown in Fig. 1 for a severe storm outbreak on 16-17 May 2017 that produced widespread large hail, damaging winds, and tornadoes. At 03:55 UTC, MESH95 values exceeded 50 mm, especially within the regions where MESH areas surpassed 250-300 km^2 along a squall line in Texas and Oklahoma (Figs. 1a and 1e). The coincident GOES-12/13 scan shows cloud tops 5-10 K colder than the tropopause (Fig. 1b) and highly probable OTs (Fig. 1d) aligned with cells containing MESH95 > 30-50+ mm. Several larger-area ECSs (e.g., Fig. 1e grey arrows) exhibit greater prominence relative to the background anvil temperature (Fig. 1c), thereby indicating that wider and colder updrafts were more likely to produce severe hail as observed by radar. Note that the northern-most indicated cell in western Oklahoma has rather weak MESH95 despite relatively strong temperature difference, OT Probability, and ECS area signatures, which suggests that satellite observations alone may not be sufficient for consistently quantifying active hail events, and can perhaps benefit from complementary knowledge of the broader storm environment.

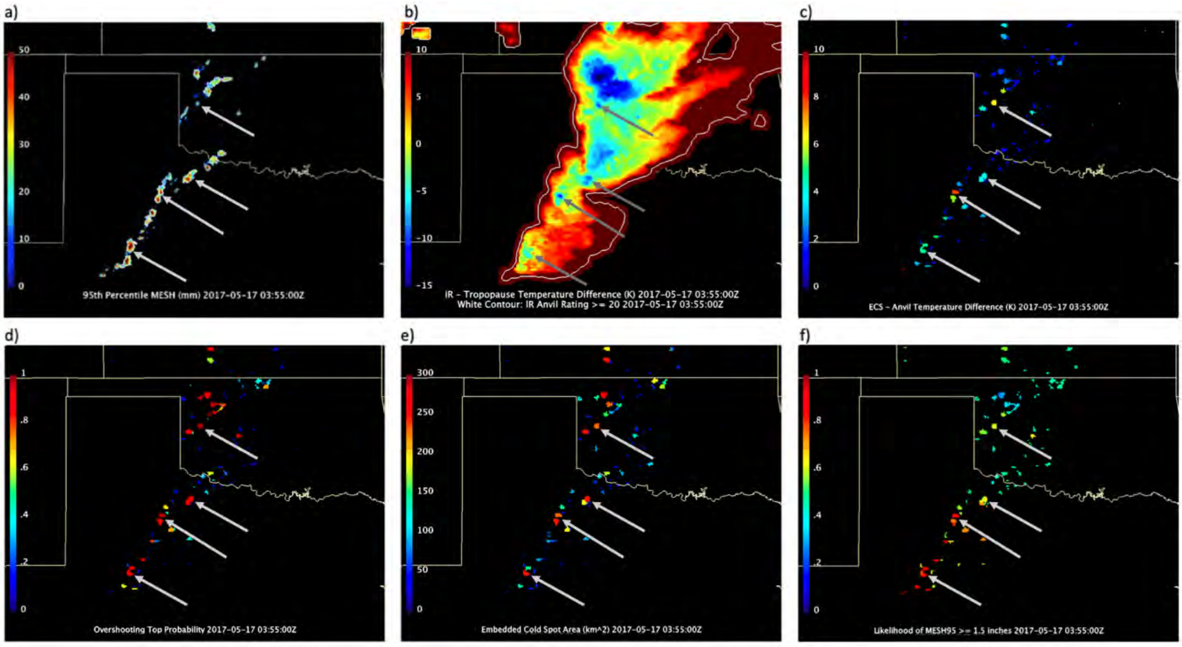
a)
50
40
30
20
10
0
95th Percentile MESH (mm) 2017-05-17 03:55:00Z
b)
10
5
0
-5
-10
-15
IR - Tropopause Temperature Difference (K) 2017-05-17 03:55:00Z
White Contour: IR Anvil Rating >= 20 2017-05-17 03:55:00Z
c)
10
8
6
4
2
0
ECS - Anvil Temperature Difference (K) 2017-05-17 03:55:00Z
d)
1
.8
.6
.4
.2
0
Overshooting Top Probability 2017-05-17 03:55:00Z
e)
300
250
200
150
100
50
0
Embedded Cold Spot Area (km^2) 2017-05-17 03:55:00Z
f)
1
.8
.6
.4
.2
0
Likelihood of MESH95 >= 1.5 inches 2017-05-17 03:55:00Z

Fig. 1. Snapshot of select properties from 03:55 UTC on 17 May 2017 during a severe hail, tornado, and wind event in the southern Great Plains. a) MESH95 and cell areas (white contour thickness is proportional to cell area), b) IR–Tropopause (GOES–MERRA-2) and IR Anvil rating ≥ 20 (white contours), c) ECS–Anvil BTD, d) OT Probability, e) ECS area, and f) DNN-predicted likelihood of severe hail. Grey arrows highlight particularly prominent ECSs, with strong temperature signatures, high OT Probability, and the largest areas.

c. Reanalysis matching with MESH cells and predictor importance analysis

A suite of convective environmental parameters, listed in Table 4, is derived using MERRA-2 and ERA5. The list includes common parameters for evaluating severe weather. Also included are indices and covariate predictors derived in combination, such as the energy helicity index (EHI; Brooks et al. 1994), which combines CAPE and storm-relative helicity SRH into one index for identifying environments favorable for rotating updrafts, and the significant hail parameter (SHIP), which is derived from most unstable CAPE (MUCAPE), mixing ratio, mid-level lapse rate, 500-hPa temperature, and 0-6-km vertical wind shear as follows (NOAA 2022):

$$SHIP = \frac{\left[\left(MUCAPE\frac{J}{kg}\right) \times \left(Mixing\ Ratio\ of\ MU\ Parcel\ \frac{g}{kg}\right) \times \left(MIDLAPSE\ \frac{°C}{km}\right) \times (-T500\ °C) \times \left(SHEAR06\ \frac{m}{s}\right)\right]}{44000000} \qquad (1)$$

Several steps are performed to resample these parameters from the reanalysis resolution to the GOES satellite grid. To avoid impacts of contamination from parameterized convection and to capture the environment that is forcing the convection, the most intense absolute model parameter values over a ~1.5°×1.8° latitude by longitude area (3 grid boxes × 3 grid boxes for MERRA-2 and 6 grid boxes × 7 grid boxes for ERA5) are found for each associated ECS (Brooks et al. 2003). Next, these reanalysis parameters are temporally interpolated to the satellite pixel scan time. Note that the Tropopause Height parameter is a standard output of the Khlopenkov et al. (2021) OT detection software, which always relies on MERRA-2. The other resampled reanalysis convective environmental parameters described in this section, aside from total precipitable water, are derived from either MERRA-2 or ERA5. Total precipitable water is always derived from MERRA-2.

Predictor short name	Description	Units
SURFCAPE	Surface Convective Available Potential Energy	J kg^{-1}
ML1000CAPE	1000-m Mixed-Layer CAPE	J kg^{-1}
ML500CAPE	500-m Mixed-Layer CAPE	J kg^{-1}
MUCAPE	Most Unstable CAPE	J kg^{-1}
SURFCIN	Surface Convective Inhibition	J kg^{-1}
ML1000CIN	1000-m Mixed-Layer CIN	J kg^{-1}
ML500CIN	500-m Mixed-Layer CIN	J kg^{-1}
MUCIN	Most Unstable CIN	J kg^{-1}
FZLV	Freezing Level	m
SHEAR01	0-1-km Vertical Wind Shear	m s^{-1}
SHEAR06	0-6-km Vertical Wind Shear	m s^{-1}
T500	500-hPa Temperature	K
MIDLAPSE	700-500-hPa Lapse Rate	ºC km^{-1}
SHIP	Significant Hail Parameter	unitless
STPLM	Significant Tornado Parameter Left-mover	m^{2} s^{-2}
STPRM	Significant Tornado Parameter Right-mover	m^{2} s^{-2}
SCPLM	Supercell Composite Parameter Left-mover	m^{2} s^{-2}
SCPRM	Supercell Composite Parameter Right-mover	m^{2} s^{-2}
EHI01LM	0-1-km Energy Helicity Index Left-mover	m^{2} s^{-2}
EHI01RM	0-1-km Energy Helicity Index Right-mover	m^{2} s^{-2}
EHI03LM	0-3-km Energy Helicity Index Left-mover	m^{2} s^{-2}
EHI03RM	0-3-km Energy Helicity Index Right-mover	m^{2} s^{-2}
SRH01LM	0-1-km Storm Relative Helicity Left-mover	m^{2} s^{-2}
SRH01RM	0-1-km Storm Relative Helicity Right-mover	m^{2} s^{-2}
SRH03LM	0-3-km Storm Relative Helicity Left-mover	m^{2} s^{-2}
SRH03RM	0-3-km Storm Relative Helicity Right-mover	m^{2} s^{-2}
LAPSE3	0-3-km Lapse Rate	ºC km^{-1}
LAPSE24	2-4-km Lapse Rate	ºC km^{-1}
THGZ	Thickness of Hail Growth Zone	m
SBLCL	Surface-based Lifted Condensation Level	m
LCLML500	500-m Mixed-Layer Lifted Condensation Level	m
TPW	Total Precipitable Water (from MERRA-2)	mm
Tropopause Height	Smoothed tropopause height from MERRA-2 (Khlopenkov et al. 2021)	km

Table 4. Reanalysis model (both ERA5 and MERRA-2 except where otherwise indicated) derived convective parameters evaluated for this study.

Predictor short name	Source
OT Probability[1]	Satellite
SHEAR01[2]	Reanalysis
SHIP[3]	Reanalysis
MUCIN[4]	Reanalysis
TPW[5]	Reanalysis
IR–Tropopause[6]	Satellite/Reanalysis
Tropopause Height[7]	Reanalysis
Anvil Frequency[8]	Satellite
Cloud Top Height[9]	Satellite
Anvil Height[10]	Satellite
ECS–Anvil BTD[11]	Satellite
ECS Area[12]	Satellite
T500[13]	Reanalysis

Table 5. Predictors found to be optimal for general detection of potentially severe hail ranked by order of their importance as determined by recursive feature analysis.

Because several input parameters are composites or multiple variables, it is likely that there would be some degree of input redundancy in a DNN that uses all or most of these selections. Furthermore, overuse of redundant parameters or those with rotation direction dependency (i.e., all STP, SCP, EHI, and SRH parameters) heighten risk of overfitting to a specific environment, thereby limiting general application. As such, the experiments introduced here leverage parameters that were selected (Table 5) based on recursive assessment of model performance while also minimizing use of inputs with strong regional dependency that would most inhibit global applicability (e.g., left-mover and right-mover dynamics and elevation-influenced lapse rates). The order in which predictors were recursively included as model features, determined by their overall contribution to the model's critical success index, are shown as superscripts in Table 5. Including latitude and longitude as predictors may help address such unwanted regional influences, but doing so is undesirable for creation of a generalized model. Although the selected parameters likely nevertheless contain some amount of regional bias, the dominant use of satellite inputs hopefully works to dampen such effects. Eight experiments are considered, defined in Table 6, with each using only the inputs listed in Table 5.

Experiment name	Description
GOES-13 + MERRA-2	GOES-12/13 and MERRA-2 reanalysis
MERRA-2 Only	Only MERRA-2 reanalysis
GOES-13 + ERA5	GOES-12/13 and ERA5 reanalysis (except Tropopause Height and TPW)
ERA5 Only	Only ERA5 reanalysis (except Tropopause Height and TPW)
GOES-13 Only	Only GOES-12/13
GOES-16 + MERRA-2	GOES-16 and MERRA-2 reanalysis, i.e., GOES-12/13-based training applied to GOES-16
Warm Season	GOES-12/13 and MERRA-2 reanalysis, trained on all seasons and applied only to days-of-year 91–273
Cold Season	GOES-12/13 and MERRA-2 reanalysis, trained on all seasons and applied only to days-of-year 1–90 and 274–366.

Table 6. Summary of input and application constraints for the nine DNN experiments.

d. Deep neural network

Classification of potentially severe hail is accomplished through the training and application of an artificial neural network (ANN) – specifically a DNN. Figure 2 roughly outlines the basic DNN structure from input data *X*, through hidden layers with activations *A*, to the final layer output signal $\hat{Y}$. Further explanation of the DNN setup and architecture is provided below.

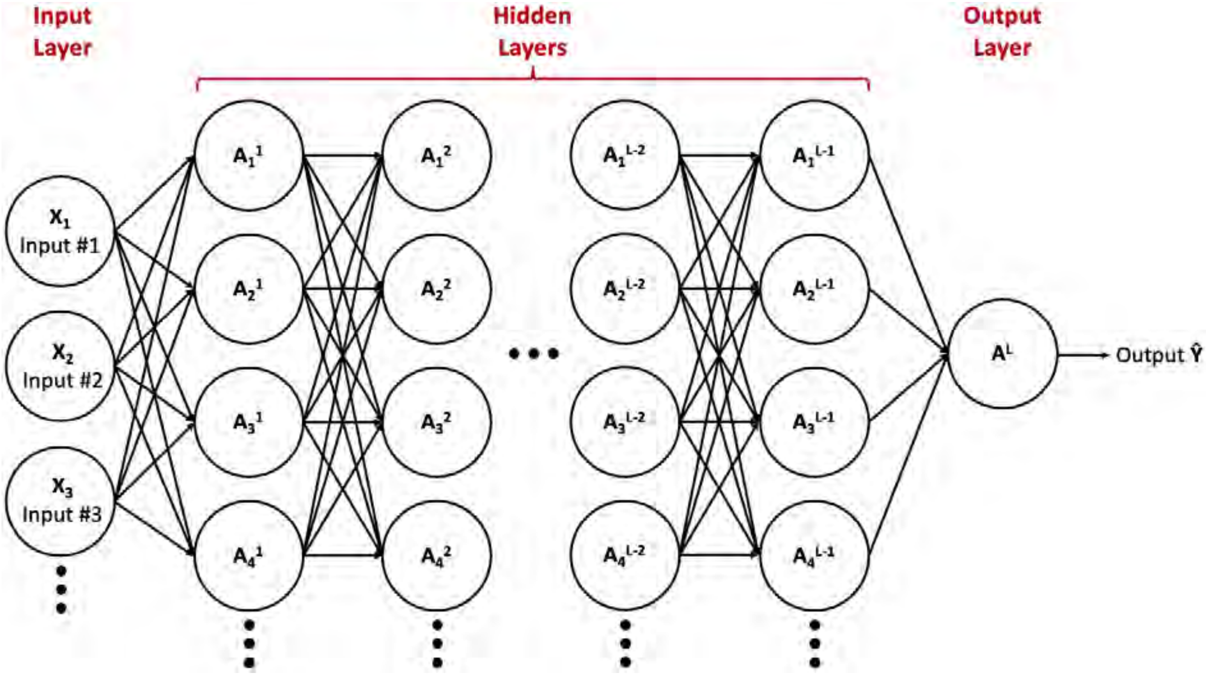

Fig. 2. Basic illustration of an *L*-layer neural network.

1) Architecture

The DNN output is a likelihood for potentially severe (MESH95 ≥ 1.5 inches) at each ECS detection. As such, a sigmoid activation function is used in the output layer neuron, which is sensible for binary classification problems (Goodfellow et al. 2016). Sigmoid activation in the output layer produces a final prediction in terms of a probability ranging from 0.0 to 1.0. Model performance is assessed at some probability threshold *p*, at which value either a negative, i.e., $\hat{Y}(\hat{Y}<p)=0$, or positive, i.e., $\hat{Y}(\hat{Y}\geq p)=1$, prediction is determined. The means of determining the value for *p* is explored below as well as in Section 4.c. Note that the DNN severe hail likelihoods are linearly scaled such that a pre-scaled likelihood of 0.8 would instead equal 1, which produces clearer visual distinction between lower and higher likelihoods of potentially severe hail.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0042.1.

Determination of the optimal DNN hyperparameters is accomplished through a randomized search of the hyperparameter space. These hyperparameters and their values are briefly discussed here and summarized in Table 7. The DNN structure determined for this study is one with 8 hidden layers, having 75, 38, 17, 12, 11, 8, 7, and 4 neurons in layers 1 through 8, respectively (determined within a random search space of 3-10 hidden layers and number of neurons selected randomly within 10-100 without exceeding the size of the previous layer). In general, it was found that models with less capacity could not learn the task well enough, and larger models tend to overfit to the training set. Neurons of the hidden layers employ the ReLU (Rectified Linear Unit) activation function, which allows for fast training given the relative simplicity of its derivative (Glorot et al. 2010). Adam Optimization is employed using 256 mini batches at a learning rate of 0.001 (Bengio 2012; LeCun et al., 2012; Kingma and Ba 2015; Masters and Luschi 2018). Additionally, early stopping is employed to mitigate possible overfitting (Prechelt 1998). With early stopping and the above learning schedule, the DNN cycles for 10-19 epochs, depending on the fold, saving learned parameters for the last epoch before the skill on the validation set fails to improve over 10 consecutive epochs. Finally, the DNN employs focal loss, which is a modified version of binary cross-entropy loss in which the solution is adjusted by a factor that modulates the contributions of correctly classified and incorrectly classified samples. Implementing focal loss influences the value at which *p* represents the optimized model. As such, focal loss modulating and weighting factors specified as 2.0 and 0.374, respectively, result in an optimized *p* value near 0.50 (Lin et al. 2020).

DNN feature	**Value**
Hidden layers	8
Neurons in layers 1-8	75, 38, 17, 12, 11, 8, 7, and 4
Minibatches	256
Learning rate	0.001
Focal loss modulating factor	2.0
Focal loss weighting factor	0.374

Table 7. Summary of DNN architecture and hyperparameters.

2) Training and validation sets

The DNN is trained using a random selection of ECS and MESH95 matches within groupings of three consecutive days across the 11 years of satellite and reanalysis data. A smaller, similarly random selection is used as a validation set that is monitored throughout the DNN tuning/training process, and an even smaller testing set is put aside for evaluation after learning is finalized to ensure that the model was not inadvertently overfit to the validation set. Only instances where the matched ECS is least two adjacent pixels in size are

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-22-0042.1.

considered in order to minimize introduction of GOES imager noise. Furthermore, following the method of Murillo et al. (2021), Fisher's linear discriminant analysis (LDA, Wilks 2006) is performed on the training and validation sets using TPW and SHEAR06 to reduce the number of MESH false alarms, which tend to occur in high TPW and low SHEAR06 environments (Murillo et al. 2021).

Splitting the datasets at the boundaries of three consecutive days lends confidence that all sets are meteorologically distinct. All datasets are standardized based on the *z*-score of the training set. The training set randomly selects from 55% of the 11-year record in three-day groups, whereas the validation set selects from 36%, and the testing set selects from 9%. These values roughly correspond to the equivalent of 6 years, 4 years, and 1 year of data being used for the training, validation, and testing sets, respectively. We used only 6 years for training because we felt it was desirable to minimize training data while keeping it in the majority as one way to reduce perceived overfitting in the climatology assessment. To demonstrate that the model is insensitive to randomness, i.e., has low variance, the DNN skill is evaluated with *k*-fold cross-validation using six folds. In other words, the DNN is trained and assessed six separate times using six distinct random compositions for the training, validation, and testing sets to determine the average and standard deviation of the skill metrics across all folds. The dataset fold used to train the model that's applied to the climatology and case studies (see Section 4 and Appendix B) was chosen because it uses none of the case study days in the training set.

Cell maxima of MESH95 are used as the truth set (*Y*). A 1.5-inch MESH95 threshold is chosen to mark the boundary of binary classification categories (*Y*=0 or *Y*=1), where MESH95 less than 1.5 inches are considered non-severe, and MESH95 of 1.5 inches or more are considered potentially severe. The distribution of the categories is summarized in Table 8. The reason the positive class is labeled as *potentially* severe is because of a high bias in MESH data, which means that 1.5-inch (1-inch) MESH95 (MESH75) is more comparable to 0.75-inch hail reported at the ground – falling short of the National Weather Service definition of severe hail (Murillo et al. 2021). Hail this size, however, is still considered part of the small hail category and remains of significant interest to the insurance industry (Murillo et al. 2021; North American Hail Workshop Panel Discussion 2022).

Dataset	Total	Warm Season	Cold Season	Non-severe	Potentially Severe	Severe/ Non-severe
Training	131311 (1636)	131311 (1636)	131311 (1636)	80419 (1140)	50892 (587)	0.633
Validation	85943 (2924)	71722 (3216)	14221 (1561)	52736 (1773)	33207 (1208)	0.630
Testing	20488 (2892)	17064 (2994)	3425 (378)	12689 (1636)	7799 (1258)	0.615

Table 8. Average and variation (in parentheses) of total, Warm Season, and Cold Season sampling, as well as the total non-severe and potentially severe samples across the 6-fold training, validation, and testing sets. The last column is the ratio of potentially severe samples to non-severe samples.

Note that for three of the experiments listed in Table 6, i.e., GOES-16 + MERRA-2, Warm Season, and Cold Season, the DNN training and validation conditions are not the same. For the GOES-16 + MERRA-2 experiment, the model is trained with GOES-13 + MERRA-2 data that exclude input from 2017 during the 6-fold random selections (each fold still representing the equivalent of 6 years of data), and the ECS Area parameter is scaled between 0 and 1 based on GOES-12/13 minimum and maximum values. The model is then applied to 6 folds of 2017 GOES-16 + MERRA-2 data (80% validation set, 20% testing set) with ECS Area scaled between 0 and 1 based on GOES-16 minimum and maximum values. The Warm (Cold) Season experiment uses the trained parameters from the GOES-13 + MERRA-2 experiment, but skill is evaluated only for days-of-year 91–273 (1–90 and 274–366).

3) Skill metrics

There are multiple relevant metrics with which the capability of a DNN can be evaluated. These metrics and their relevance to meteorological warnings have been reviewed extensively in previous works and are summarized in this section (Doswell et al. 1990; Schaefer 1990; Barnston 1992; Gerapetritis 1995; Kunz 2007; Barnes et al. 2009; Hyvärinen 2013; Johnson and Sugden 2014; Gensini and Tippet 2019). Skill metrics consider different combinations of *true positive*, *false positive*, *false negative*, and *true negative* classifications, and therefore it is important to diversify one's evaluation criteria to be sure the model is not deficient in any one category. The probability of detection (POD) considers the ratio of true positives to the sum of observed severe events – thereby explaining the rate of correct severe hail predictions for all actual cases of severe hail. Next, there are two metrics for explaining false alarms (FA). One is the FA Ratio, which considers the ratio of false positive predictions to the sum of all positive predictions. The second is FA Rate, which considers the ratio of false positive predictions to the sum of observed non-severe events. The critical success index (CSI) is given by the ratio of true positive predictions to the sum of true positive, false positive, and false negative predictions, and has a range from 0 (no predictive skill) to 1 (perfect prediction). Heidke skill score (HSS) accounts for all four classification outcomes, and can range from -∞ to 1, where values 0 and below indicate no predictive skill and 1 indicates a perfect prediction. The final metric is the Frequency Bias Index (FBI), often simply called *Bias* in the context of categorical forecasts, which is the ratio of total positive predictions to total observed severe events, where 1 is perfect and values less than (greater than) 1 indicate tendency toward under (over) forecasting. The exact formulations for these skill metrics are summarized by Wilks (2006).

e. Climatological aggregation

To compare the frequency of the predicted hail events with various observational sources, daily severe hail event occurrences are counted on an 80×80-km^2 grid for each dataset over CONUS from 2007 through 2017 (Brooks et al. 2003; Doswell et al. 2005; Cintineo et al. 2012; Allen and Tippet 2015; Murillo and Homeyer 2019). Only one detection per day (00:00-23:00 UTC) is required to be considered an “event day,” which thereby limits the impact of spatio-temporal resolution and sampling opportunity differences on the resulting climatologies. Following Murillo and Homeyer (2019), 1-sigma gaussian smoothing is applied. Whereas the training, validation, and testing datasets use data collected within ±7.5 mins from the top of the hour, the climatology is aggregated from satellite files with scans beginning 15 mins after the hour (e.g., 23:15 UTC). Owing to the time resolution of the reanalysis data, however, around 55% of the reanalysis inputs used to train the model also appear in the climatology application, but never in the same exact paring with the satellite data and with slightly different time-interpolated values. Therefore, input combinations for application are distinct from training input combinations, although with 6 of 13 predictor values similar to what may have been used during training within a given reanalysis grid box.

4. Results and Discussion

In this section we discuss results covering five focus areas, beginning broad before narrowing on specific applications. We start by documenting the climatological distribution of MESH, hail reports, and MWR hail probability over the study domain. Second, the parameter space of certain satellite and reanalysis inputs are explored as a function of MESH hail size, and correlations of hailstorm detections with input parameters are examined. Outcomes of the DNN are presented next, including explanation of general predictive capability, demonstration of specific performance metrics for each input sensitivity experiment, and discussion of caveats. Fourth, we highlight the DNN hail likelihood climatology and relate that to the MESH climatology discussed previously. Finally, we end the section with examination of hail likelihood estimates for two case studies (with a third presented in Appendix B).

a. Regional hail distribution

The yearly-average severe hail event days from MESH and SPC storm reports show a maximum frequency in the Central Plains that is shifted west of the maximum frequency for non-severe hail event days (Figs. 3a and 3c). Reduced frequency, however, of ground spotter reports (Fig. 3b), likely due to well-known biases (Allen and Tippett, 2015), is apparent. Note that MWR hailstorm detections are rendered as point observations rather than as a grid because TMI observations do not extend beyond 39° latitude and have greater sampling density at the northernmost extent of the orbit (Fig. 3d). This sampling discrepancy, in addition to time-of-day bias from the AMSR2 sun-synchronous orbit, would bias a gridded climatology relative to patterns depicted in the MESH climatology. Nevertheless, there is a concentration of high likelihood MWR hail detections in the Central Plains (red symbols), consistent with MESH and SPC. Higher densities of MWR hail detections also occur in regions outside of CONUS, including along the Sierra Madre in northwestern Mexico, northeastern Mexico, and the Gulf Stream.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-22-0042.1.

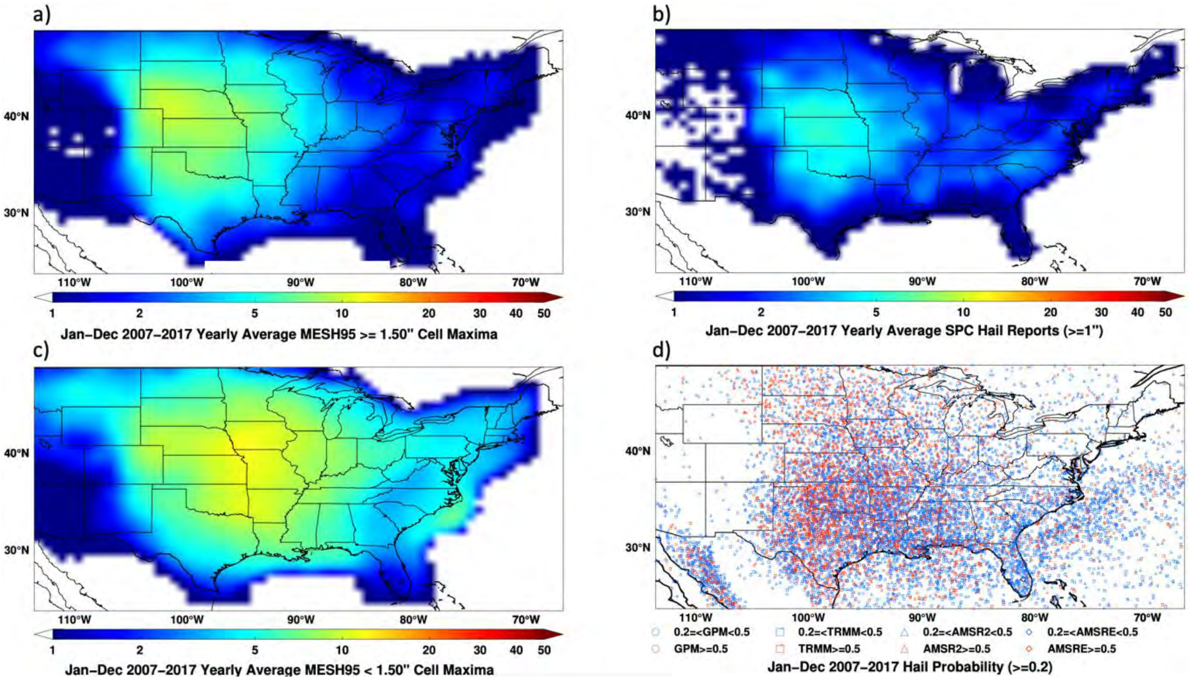
a)
b)
c)
d)
40°N
30°N
110°W
100°W
90°W
80°W
70°W
1
2
5
10
20
30
40
50
Jan–Dec 2007–2017 Yearly Average MESH95 >= 1.50" Cell Maxima
Jan–Dec 2007–2017 Yearly Average SPC Hail Reports (>=1")
Jan–Dec 2007–2017 Yearly Average MESH95 < 1.50" Cell Maxima
0.2=<GPM<0.5
GPM>=0.5
0.2=<TRMM<0.5
TRMM>=0.5
0.2=<AMSR2<0.5
AMSR2>=0.5
0.2=<AMSRE<0.5
AMSRE>=0.5
Jan–Dec 2007–2017 Hail Probability (>=0.2)

Fig. 3. Yearly average severe hail event days observed on an 80×80 km^2 grid (1-sigma Gaussian smoothing applied) during 2007-2017 matched within ±7.5 minutes and 28×28 km^2 of a GOES-12/13 ECS detection. a) MESH95 ≥ 1.5 inches, b) SPC hail reports, c) MESH95 < 1.5 inches, and d) MWR hail detection probability exceeding 0.2. The satellite MWR is plotted as symbols and colored blue (red) if exceeding a low (high) threshold. Note that TMI observations do not extend beyond 39° latitude and have greater sampling density at the northernmost extent of the orbit.

b. Parameter sensitivity to hail size and reanalysis

The satellite and reanalysis parameter distribution as a function of hail size is complex due in part to 1) challenges in depicting the spatio-temporal evolution of severe storm environments by reanalysis, 2) the fact that severe hailstorms can be caused by a broad spectrum of environmental conditions (i.e., low CAPE and high wind shear, high CAPE and low wind shear; Johnson and Sugden, 2014; Gensini et al. 2021; Zhou et al. 2021), and 3) the imprecise process of spatio-temporally matching satellite-based cloud top properties with hail reports and MESH cells. Nevertheless, value distribution plots stratified by MESH hail size categories and correlation analyses can reveal patterns indicative of severe hailstorms.

Box and whisker distributions of selected GOES-12/13 IR-based parameters show increasing satellite-derived intensity as hail size increases, with varying degrees of overlap among neighboring bins (Fig. 4). For example, the median OT Probability within the smallest MESH95 bin (0.5-1.0 inches) is ~0.33, increasing to ~0.8 for 3+ inch MESH95 (Fig. 4a). ECS Area is 2-3× larger (Fig. 4b) and tropopause-relative IR BT (Fig. 4d) is ~5 K colder for 3+ inch MESH95 relative to < 1-inch MESH95, indicating that colder and wider updrafts are more likely to generate larger hail. Marion et al. (2019) computed OT Area using a similar

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-22-0042.1.

approach and found significant correlation between OT Area and tornado intensity. Some parameters like ECS–Anvil BTD have more gradual and subtle sensitivity to MESH95 size, for which a ~2 K median increase in BTD from the lowest to highest MESH95 bin was found (Fig. 4c). Sensitivity of these relationships to imager resolution is explored during the 2017 warm season using GOES-13 and GOES-16 in Appendix A, showing that GOES-16 metrics are largely consistent with GOES-13, but GOES-16 records slightly colder temperatures, smaller ECS area, and greater sensitivity to hail size due to the higher pixel resolution. It is also important to note that large hail has occurred when satellite parameters are weak, which could be caused by factors such as 1) time differences between GOES and MESH data, 2) uncertainty in MESH data where, in reality, a cell may not truly be as intense as depicted by MESH, and 3) the ability for large hail to be produced in a cloud without notable BT variations at cloud top.

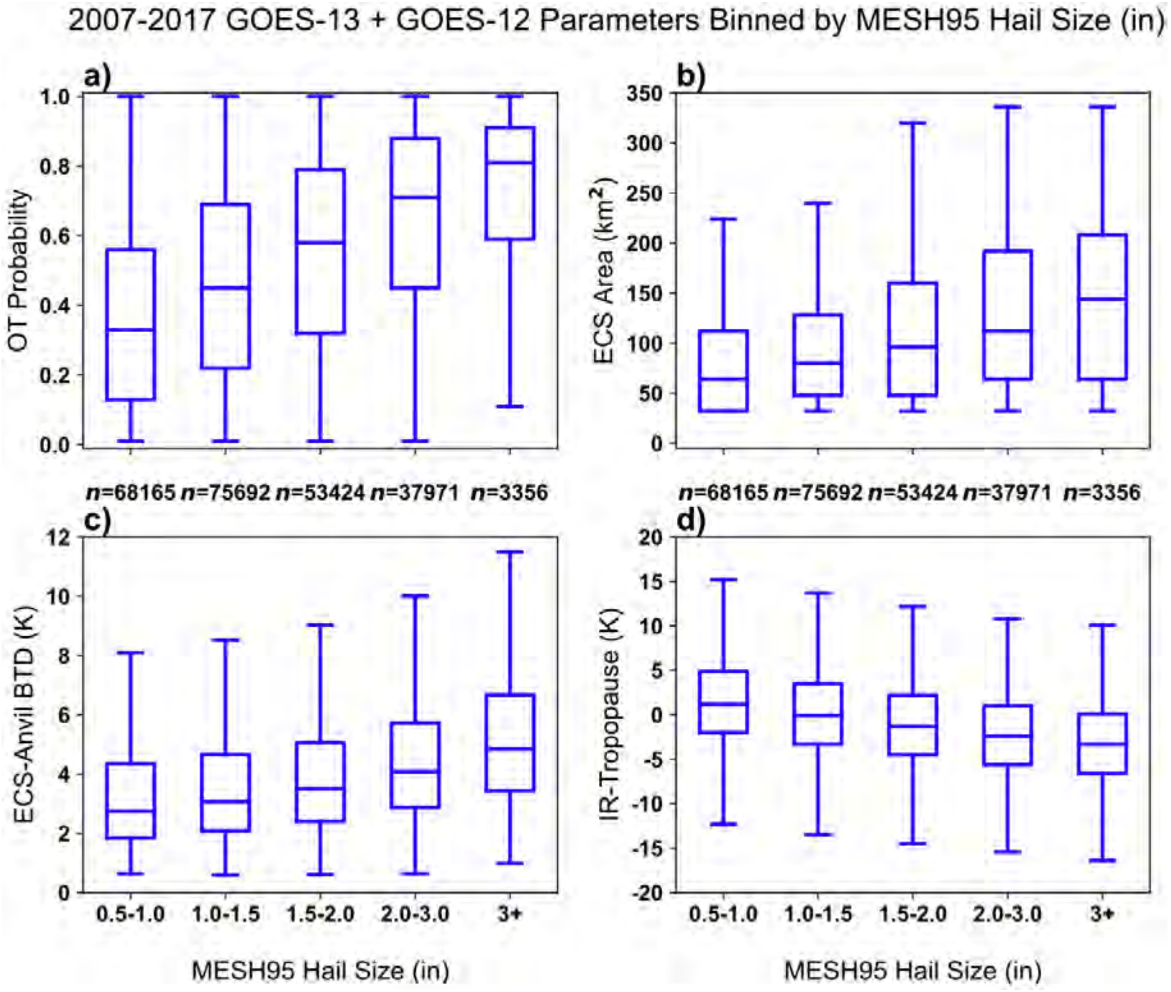

Fig. 4. Box and whisker distributions (center line indicates median value, rectangles encompass the IQR, and whiskers extend to 1.5×IQR, outliers are not shown) of select GOES-12/13 IR cloud top parameters stratified by MESH95 hail size bins (matched within ±7.5 minutes and 28×28 km^2) during 2007-2017. a) OT Probability, b) ECS Area, c) ECS–Anvil BTD, and d) IR–Tropopause.

Some MERRA-2 convective environment parameters also increase in intensity with increasing MESH95 (Fig. 5). Similar such results are also found for ERA5 but are not shown. From the lowest to highest MESH95 bin, the median MUCIN (Fig. 5a) increases by ~40 J kg^{-1} while SHEAR01 (Fig. 5c) decreases by ~5 ms^{-1}. Furthermore, Fig. 5d shows a preference for ~2× larger median values of SHIP for 2+ inch compared to < 1-inch MESH95 bins. It is intuitive to anticipate more vigorous convective updrafts when parcels overcome large CIN in drier environments such as the explosive growth associated with the initiation of dryline thunderstorms over Oklahoma, Texas, and Kansas. The pattern may also be a reflection of there being reduced interaction between convective cells in large CIN environments, i.e., discrete storms favoring larger hail likelihood. Similarly, environments with higher SHIP (>1) show a tendency for producing larger MESH95. Although the negative correlation between MESH and SHEAR01 might seem counter intuitive, it is possibly owed to decreased hailstone residence times in highly sheared updrafts (Dennis and Kumjian 2017). Like the satellite analyses depicted above, large MESH95 does occur in many instances where environmental parameters are more muted than expected, such as in Fig. 5b, which presents a challenge to DNN learning capability. Evaluating such parameters in isolation in this way, however, does not capture possible multivariate contributions that the DNN may detect.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0042.1.

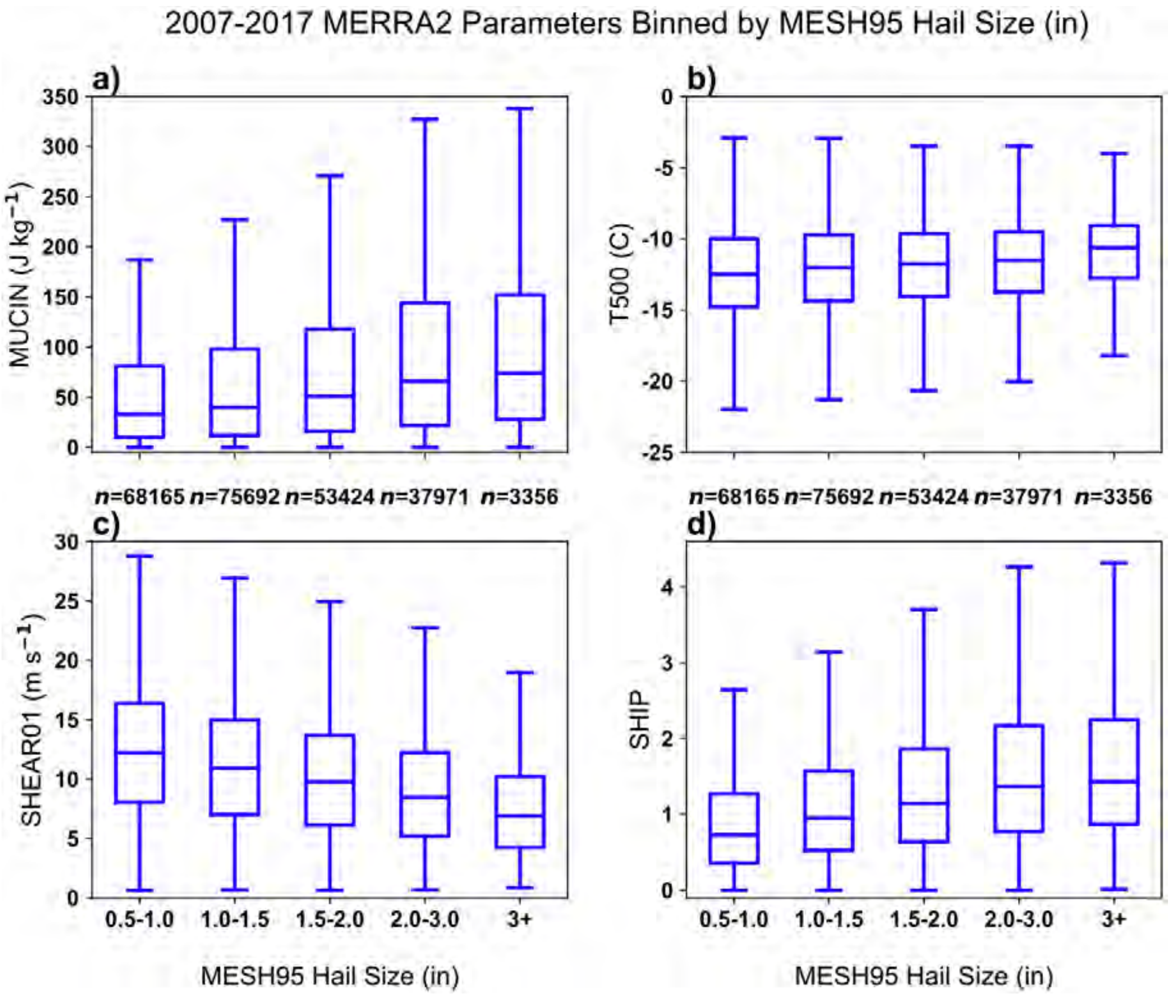

Fig. 5. Same as Fig. 4 except for select MERRA-2 convective parameters binned by MESH95: a) MUCIN, b) T500, c) SHEAR01, and d) SHIP.

A correlation matrix quantifies the relationships between hailstorm detections and satellite/reanalysis parameters (Fig. 6). Specifically, MESH95 maxima, MWR hail probability, and ground reported hail size are correlated with one another and with the GOES-12/13 and MERRA-2 parameters from Tables 3 and 4. OT Probability, ECS Area, SHIP, and IR–Tropopause are the most strongly correlated with MESH95 with Pearson correlation coefficients (*r*) in the range 0.27-0.38. Ignoring the expectedly correlative height variables, and despite a few instances of exceptionally strong correlation among other predictors, e.g., OT Probability vs. ECS Area (*r*=0.75) and ECS Area vs. IR-Tropopause (*r*=-0.61), most predictors are relatively independent of one another. While MWR hail probability correlation with the selected parameters is overall comparable to that of MESH95, with some notable exceptions, e.g., SHEAR01, significantly weaker relationships are found between reported hail size and reanalysis/satellite parameters, likely a result of uncertainty of hail size reports (Allen and Tippet 2015) and the inherent absence of sub-severe hail reporting. Lastly, and still with acknowledgement of the same caveats for hail reporting, we see that MESH and MWR hail probability are more than twice as correlated as MESH and SPC hail size, which quantifies the relative patterns seen in Fig. 3.

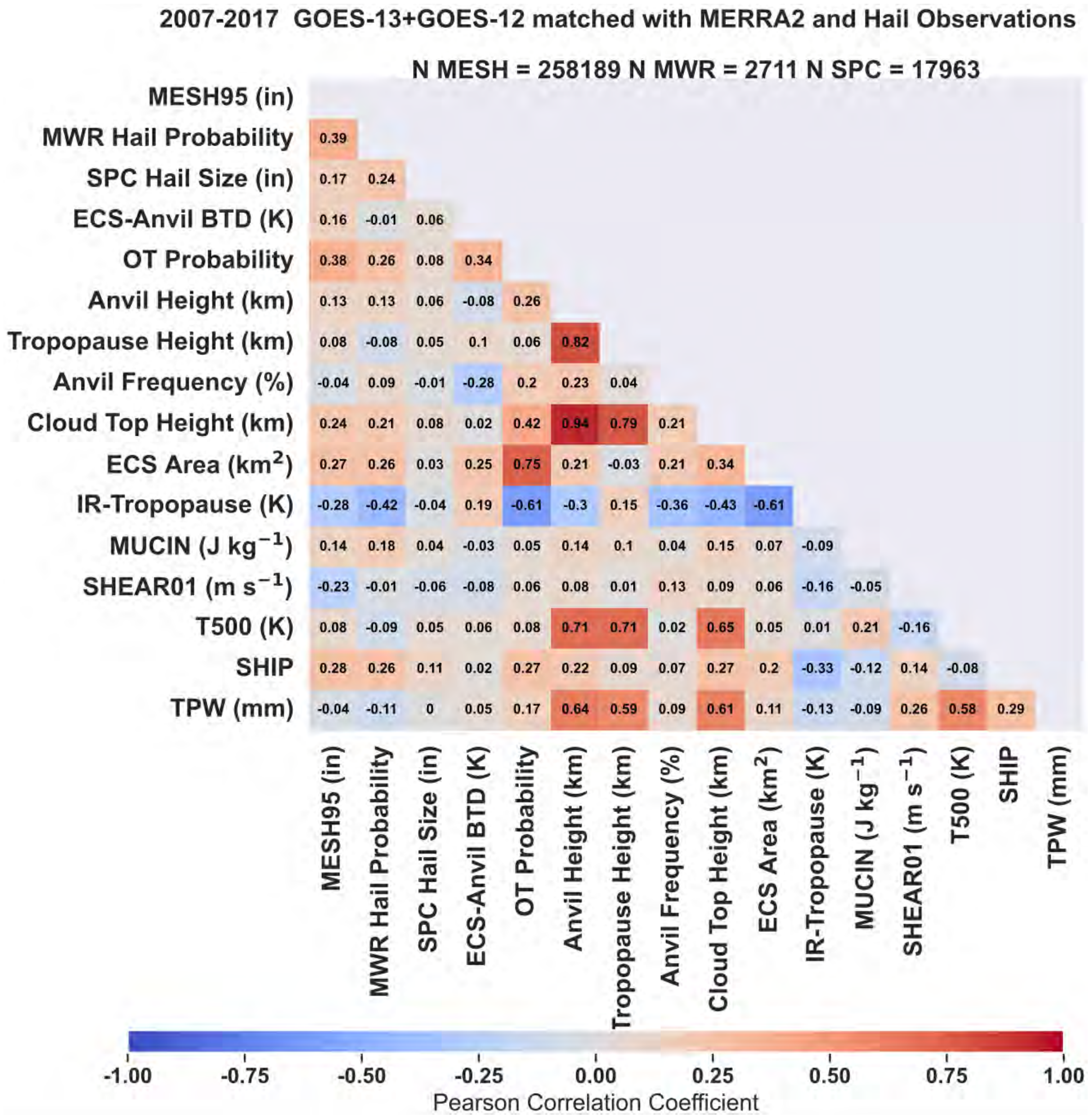

Fig. 6. Correlation matrix of select GOES-12/13 and MERRA-2 model reanalysis parameters matched within ±7.5 minutes and 28×28 km^2 of a MESH95 cell maxima and MESH95 cell areas. Co-located MWR hail probability detections and ground spotter hail size reports are also included.

Figure 6 also highlights that certain satellite parameters, i.e., OT Probability, ECS Area, and ECS–Anvil BTD, are largely independent of the environment, and therefore should uniquely contribute to hailstorm discrimination. Meanwhile, some satellite/reanalysis correlations pairings, such as Anvil Height with T500 and TPW, are rather strong (r=0.71 and r=0.64, respectively). A warmer atmosphere results in higher equilibrium heights and higher cloud tops given sufficient instability and moisture.

In summary, multiple satellite and model parameters are shown to be potentially useful predictors of observed hailstorm characteristics. Overall, larger hail is linked to wider and colder updrafts occurring in environments with weaker low-level shear and increased convective inhibition and SHIP. Next, we highlight the outcomes of a DNN model used to establish quantitative links among these parameters to skillfully detect potentially severe hailstorms.

c. DNN validation

1) Association of DNN prediction to truth

Increasing likelihood of potentially severe hail is correlated with increasing MESH, with a greater range of hail size covered by MESH95 compared to that by MESH75 (Fig. 7). This positive correlation between likelihood and MESH lends confidence to the DNN performance. That is, the model is assigning the highest likelihoods to storms with the greatest potential severity, and therefore thresholds of hail likelihood can be tailored to focus on storms producing the largest hail.

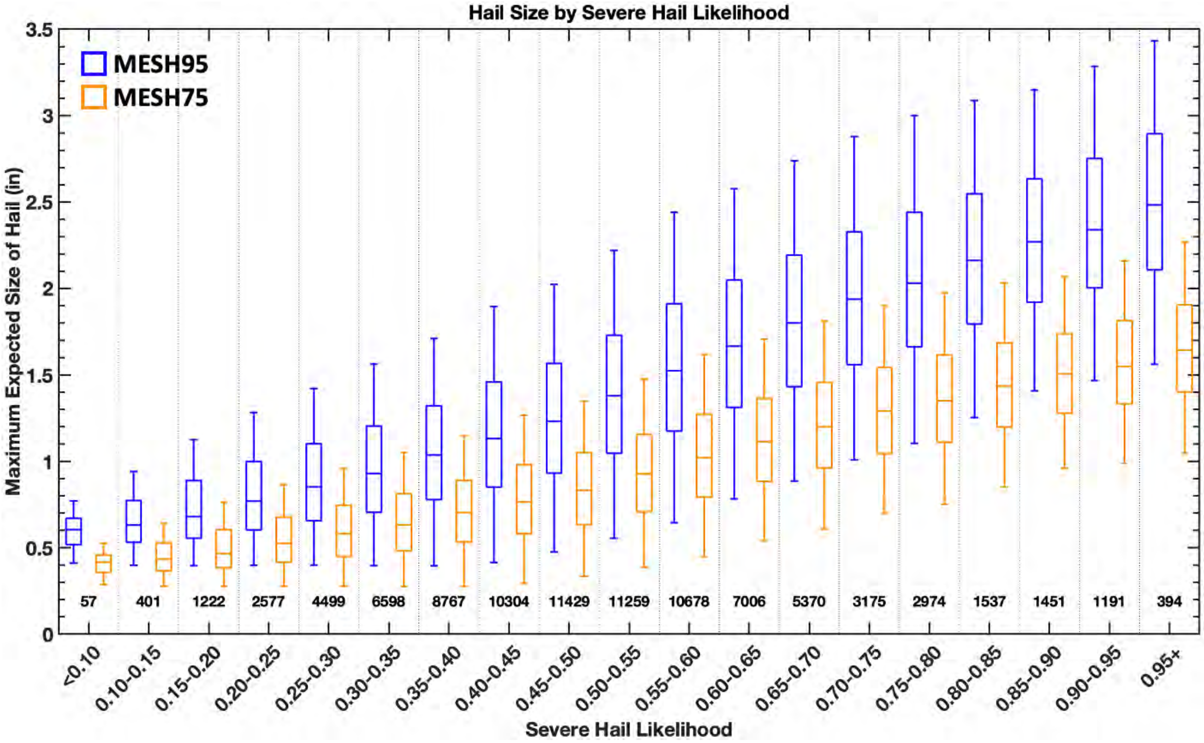

Fig. 7. Validation set distributions of MESH95 and MESH75 at 5% intervals of potentially severe hail likelihood, where potentially severe is defined as 1.5 inches for MESH95 and 1 inch for MESH75. Numbers indicate validation samples per bin interval. The severe hail likelihoods have been linearly scaled such that a pre-scaled likelihood of 0.8 would instead equal 1.

2) Input sensitivity experiment results

The DNN accuracy is assessed using receiver operating characteristic (ROC) curves based on models applied to the validation set (Fig. 8). A summary of the experiment skill scores for likelihoods determined by the *p* value at the inflection point of the ROC curve is provided in Table 9, which provides not only validation set skill metrics, but also those for the training and testing sets. While a ROC curve shows the relationship between POD and FA Rate, a performance diagram (PD) incorporates multiple validation metrics: POD, FA Ratio, and CSI (Gerapetritis 1995; Roebber 2009). Figure 9 shows a PD for the same ensemble of models shown in Fig. 8. Figure curves and table statistics represent the 6-fold-average results (standard deviations represented by parentheses in Table 9).

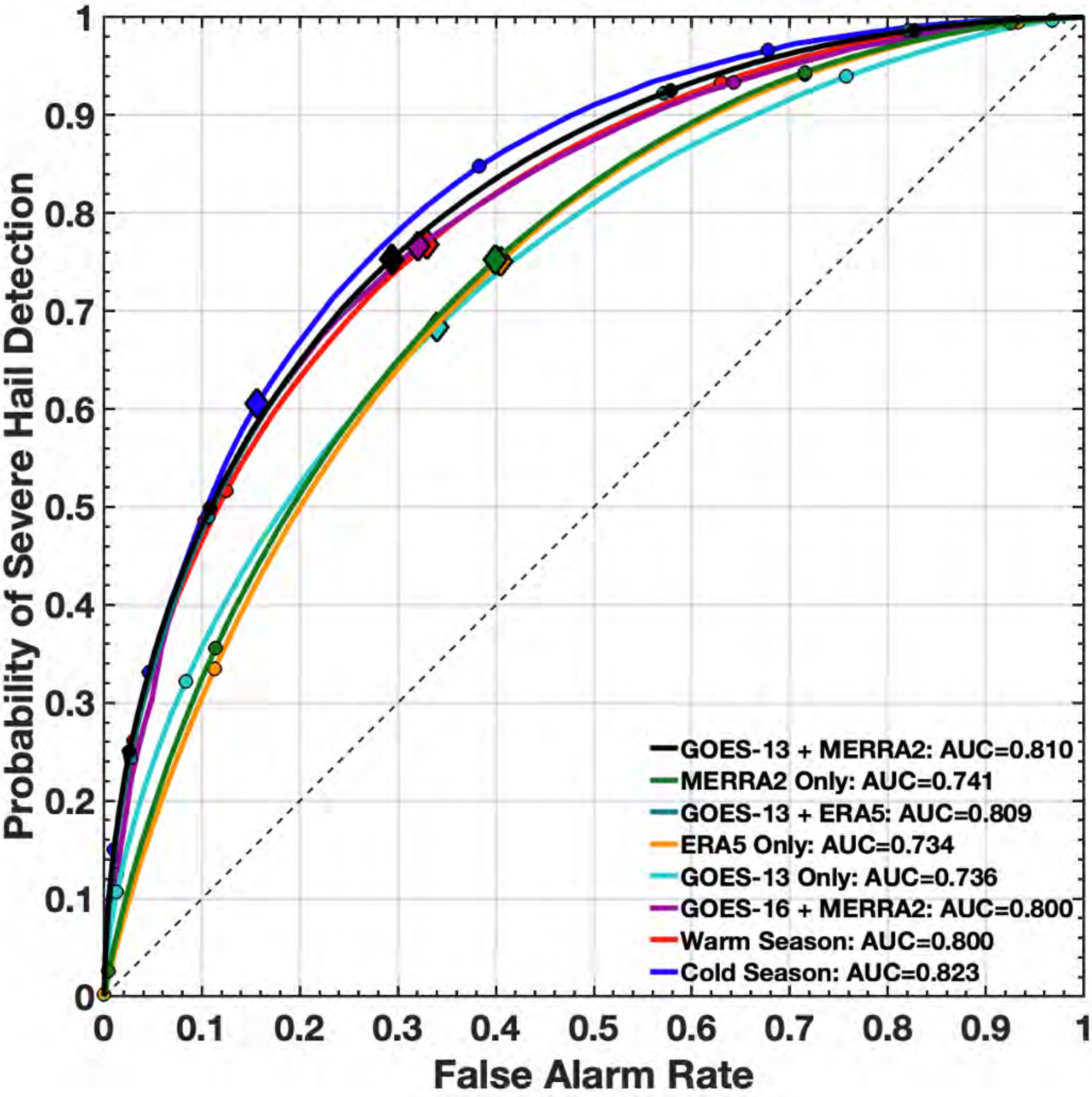

Fig. 8. Receiver operating characteristic (ROC) 6-fold-average curves illustrating the classification capability of the DNN as applied to the validation set for each experiment (see Table 6). Diamonds designate *p*=0.5 and circles mark *p* increments of 0.1 lower and higher than 0.5.

Dataset	Experiment	POD	FA Ratio	FA Rate	CSI	HSS	AUC	FBI	p
Training Set	**GOES-13 + MERRA-2**	**0.728 (0.015)**	**0.354 (0.007)**	**0.253 (0.013)**	**0.520 (0.004)**	**0.416 (0.004)**	**0.819 (0.003)**	**1.128 (0.033)**	**0.51 (0.01)**
	MERRA-2 Only	0.790 (0.027)	0.466 (0.010)	0.438 (0.031)	0.467 (0.003)	0.311 (0.006)	0.746 (0.003)	1.482 (0.079)	0.49 (0.01)
	GOES-13 + ERA5	**0.722 (0.016)**	**0.355 (0.006)**	**0.248 (0.013)**	**0.517 (0.007)**	**0.417 (0.004)**	**0.817 (0.004)**	**1.118 (0.032)**	**0.52 (0.01)**
	ERA5 Only	0.790 (0.011)	0.474 (0.004)	0.445 (0.012)	0.462 (0.005)	0.305 (0.004)	0.738 (0.004)	1.502 (0.024)	0.49 (0.01)
	GOES-13 Only	**0.749 (0.010)**	**0.462 (0.005)**	**0.406 (0.010)**	**0.456 (0.003)**	**0.306 (0.002)**	**0.741 (0.002)**	**1.391 (0.029)**	**0.48 (0.01)**
	GOES-16 + MERRA2	0.720 (0.029)	0.350 (0.015)	0.250 (0.027)	0.518 (0.008)	0.414 (0.005)	0.817 (0.002)	1.110 (0.070)	0.52 (0.01)
	Warm Season	**0.728 (0.015)**	**0.354 (0.007)**	**0.253 (0.013)**	**0.520 (0.004)**	**0.416 (0.004)**	**0.819 (0.003)**	**1.128 (0.033)**	**0.51 (0.01)**
	Cold Season	0.728 (0.015)	0.354 (0.007)	0.253 (0.013)	0.520 (0.004)	0.416 (0.004)	0.819 (0.003)	1.128 (0.033)	0.51 (0.01)
Validation Set	**GOES-13 + MERRA-2**	**0.721 (0.013)**	**0.363 (0.009)**	**0.259 (0.012)**	**0.511 (0.003)**	**0.407 (0.002)**	**0.810 (0.001)**	**1.131 (0.037)**	**0.51 (0.01)**
	MERRA-2 Only	0.787 (0.023)	0.469 (0.008)	0.439 (0.032)	0.464 (0.002)	0.308 0.009	0.741 (0.003)	1.483 (0.067)	0.49 (0.01)
	GOES-13 + ERA5	**0.711 (0.017)**	**0.361 (0.010)**	**0.251 (0.014)**	**0.507 (0.005)**	**0.408 (0.003)**	**0.809 (0.003)**	**1.114 (0.042)**	**0.52 (0.01)**
	ERA5 Only	0.788 (0.010)	0.475 (0.007)	0.445 (0.013)	0.460 (0.006)	0.304 (0.007)	0.734 (0.006)	1.501 (0.031)	0.49 (0.01)
	GOES-13 Only	**0.747 (0.005)**	**0.466 (0.005)**	**0.410 (0.010)**	**0.452 (0.003)**	**0.302 (0.004)**	**0.736 (0.002)**	**1.397 (0.021)**	**0.48 (0.01)**
	GOES-16 + MERRA2	0.712 (0.010)	0.369 (0.009)	0.257 (0.009)	0.503 (0.004)	0.403 (0.001)	0.800 (0.002)	1.128 (0.028)	0.52 (0.01)
	Warm Season	**0.660 (0.010)**	**0.317 (0.004)**	**0.220 (0.007)**	**0.505 (0.004)**	**0.393 (0.002)**	**0.800 (0.001)**	**0.972 (0.026)**	**0.54 (0.02)**
	Cold Season	0.685 (0.030)	0.513 (0.024)	0.208 (0.020)	0.397 (0.015)	0.419 (0.011)	0.823 (0.004)	1.312 (0.059)	0.54 (0.02)
Testing Set	**GOES-13 + MERRA-2**	**0.714 (0.018)**	**0.365 (0.012)**	**0.252 (0.024)**	**0.506 (0.006)**	**0.410 (0.013)**	**0.810 (0.007)**	**1.125 (0.046)**	**0.51 (0.01)**
	MERRA-2 Only	0.780 (0.025)	0.470 (0.017)	0.425 (0.046)	0.461 (0.006)	0.315 (0.020)	0.746 (0.011)	1.474 (0.095)	0.49 (0.01)
	GOES-13 + ERA5	**0.701 (0.015)**	**0.361 (0.015)**	**0.242 (0.017)**	**0.502 (0.010)**	**0.410 (0.008)**	**0.810 (0.006)**	**1.099 (0.041)**	**0.52 (0.01)**
	ERA5 Only	0.774 (0.012)	0.473 (0.016)	0.425 (0.026)	0.456 (0.012)	0.309 (0.017)	0.737 (0.012)	1.471 (0.054)	0.49 (0.01)
	GOES-13 Only	**0.744 (0.007)**	**0.474 (0.005)**	**0.410 (0.011)**	**0.446 (0.005)**	**0.301 (0.007)**	**0.734 (0.004)**	**1.414 (0.019)**	**0.48 (0.01)**
	GOES-16 + MERRA2	0.707 (0.023)	0.367 (0.022)	0.260 (0.018)	0.501 (0.019)	0.396 (0.007)	0.797 (0.006)	1.118 (0.050)	0.52 (0.01)
	Warm Season	**0.651 (0.015)**	**0.318 (0.009)**	**0.212 (0.004)**	**0.500 (0.011)**	**0.396 (0.010)**	**0.800 (0.009)**	**0.958 (0.040)**	**0.54 (0.02)**
	Cold Season	0.700 (0.052)	0.517 (0.047)	0.219 (0.047)	0.399 (0.033)	0.419 (0.027)	0.826 (0.013)	1.354 (0.113)	0.54 (0.02)

Table 9. Average and variation (in parentheses) of skill metrics on training, validation, and testing sets for the different DNN experiments, evaluated with 6-fold cross-validation. The prediction likelihoods at which the statistics are determined are based on the inflection points of the ROC curves for each fold, the average and standard deviation of which are given in the column marked *p*.

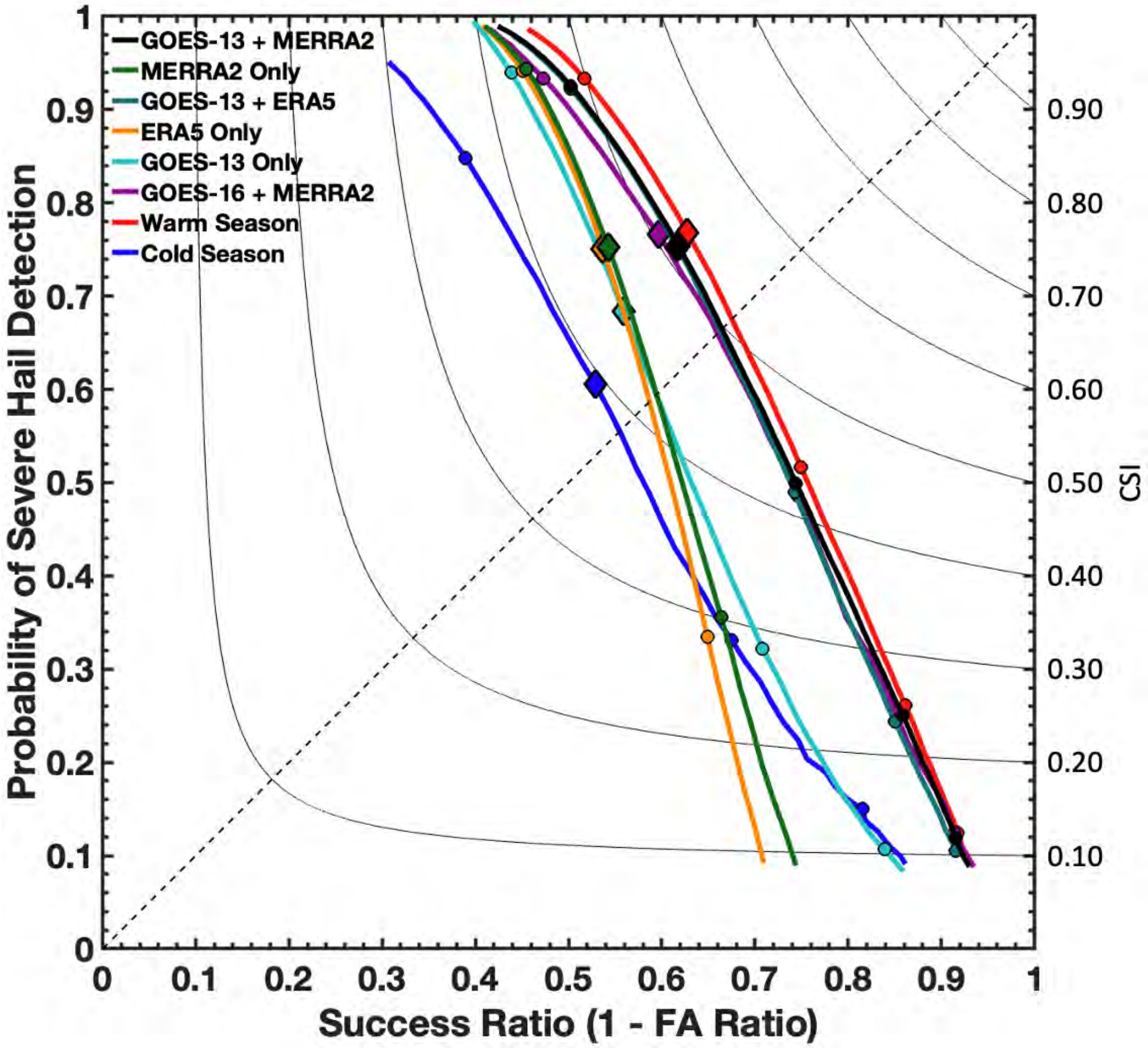

Fig. 9. Performance diagram 6-fold-averag curves illustrating the classification capability of the DNN as applied to the validation set for each experiment (see Table 6). Diamonds designate *p*=0.5 and circles mark *p* increments of 0.1 lower and higher than 0.5.

Before discussing the specific performance details, it is notable that the various skill metrics are rather consistent, per each experiment, across the training, validation, and testing sets. For example, in the GOES-13 + MERRA-2 experiment, both CSI and HSS on the training set outperform CSI and HSS on the validation set by only 0.009 (by 0.014 and 0.006 relative to the testing set, respectively). This consistency suggests that there is a low amount of variance in the model, meaning that it is comprehensive and not overfit to the training set. Furthermore, because the validation set and testing set skill scores are comparable, there is no concern that the model was inadvertently overfit to the validation set during the model tuning process. There is bias in the model (the DNN predictions are not perfect), but it is indiscriminate of the input dataset. Therefore, these results should help alleviate concerns that climatology results may not be well generalized.

3) Input sensitivity experiment discussion

Interpretation of results may be influenced by the skill metric being considered. The GOES-13 + MERRA-2 experiment offers what one might consider “well rounded” performance, as it scores relatively well in terms of CSI, HSS, and AUC (Fawcett 2005) at 0.511, 0.407, and 0.810, respectively, on the validation set, and 0.506, 0.410, and 0.810, respectively, on the testing set. The GOES-16 + MERRA-2 performance is comparable at 0.501-0.503 CSI, 0.396-0.403 HSS, and 0.797-0.800 AUC on the validation and testing sets. Other recent studies that have sought to detect severe hail events, using MESH, storm reports, or insurance reports as truth, and station data, radiosondes, radar reflectivity, lightning, and/or reanalysis as inputs, achieved CSI values ranging from 0.17 to 0.27, HSS ranging from 0.26 to 0.39 and AUC ranging from 0.77 to 0.78. The scope, goals, and success metrics of these previous studies, however, are unique and not necessarily aligned with those of our study, e.g., forecasting is not our intention (Kunz 2007; Gagne et al 2017; Czernecki et al. 2019; Gensini et al. 2021). One fold of the GOES-13 + MERRA-2 model, with scores of 0.512-0.514 CSI and 0.409-0.434 HSS on the validation and testing sets, is selected to compile the severe hail climatology and to demonstrate results from case studies shown below.

The GOES-13 + ERA5 experiment skill metrics are comparable to those of GOES-13 + MERRA-2. Despite the small CSI advantage of the latter and its overall better 6-fold stability, i.e., lower variance across the 6-fold evaluation, using either reanalysis source along with satellite observations will lead to effectively the same outcome. One reason for why the MERRA-2 DNN may marginally outperform the ERA5 DNN in this study, either alone or in combination with satellite inputs, might be owed to the fact that the tropopause height, tropopause temperature, and TPW inputs are sourced from MERRA-2 in all instances of their use. Whether an ERA5 DNN that sources ERA5 for those inputs would demonstrate notably higher skill is an area for future study.

In the case of the Cold Season analysis, the disparity in skill demonstrated by the ROC curve relative to the PD is notable. The Cold Season ROC curve is higher than those from the other satellite + reanalysis experiments, demonstrating the highest AUC out of all experiments at 0.823-0.826 and the highest HSS at 0.419. However, Cold Season is the poorest performer in terms of CSI at 0.397-0.399. This disparity in performance is owed to the high FA Ratio of 0.513-0.517 and relatively low FA Rate of 0.208-0.219. In the case of the Warm Season, CSI is closer to that for all seasons at 0.500-0.505, but with worse HSS than the Cold Season at 0.393-0.396 – outcomes that are owed to relatively low POD (0.651-0.660) but relatively favorable FA Ratio (0.317-0.318) and FA Rate (0.212-0.22).

A pattern similar to the Cold Season HSS and CSI discrepancy was reported in a large hail study performed by Czernecki et al. (2019), who, with a model trained to predict reported severe hail based on radar reflectivity, lightning detection, and ERA5 parameters, demonstrated HSS performance of 0.383 and CSI of 0.245. In that study, the discrepancy was driven by a significant difference between FA Rate (~1%) and FA Ratio (~67%), which may suggest that although there was adequate recognition of non-events, the model perhaps attempted relatively few positive predictions, or had few opportunities to do so. Relating the significance of such prediction opportunities to our Cold Season results, it is important to keep in mind that model sampling is dominated by warm season events, and therefore the DNN is biased toward such environments. Figure 10 demonstrates this seasonal dependence, revealing diminished MUCIN (Fig. 10a) in the cold season compared to the warm season for increasing MESH95, supporting previous findings on the distinct seasonal characteristics conducive to severe hail occurrence (Púčik et al. 2015). Aside from the higher CIN, a DNN dominated by warm season inputs learns to associate severe hail with overall reduced SHEAR01 (Fig. 10b) and higher cloud top heights (Fig. 10c). These differences between the cold and warm season distributions, with the exception of OT Probability (Fig. 10d), do not lend to a cold season DNN performance as favorable as that of the warm season. Improving cold season skill, perhaps with specialized training and increased sampling, is an area that warrants further study.

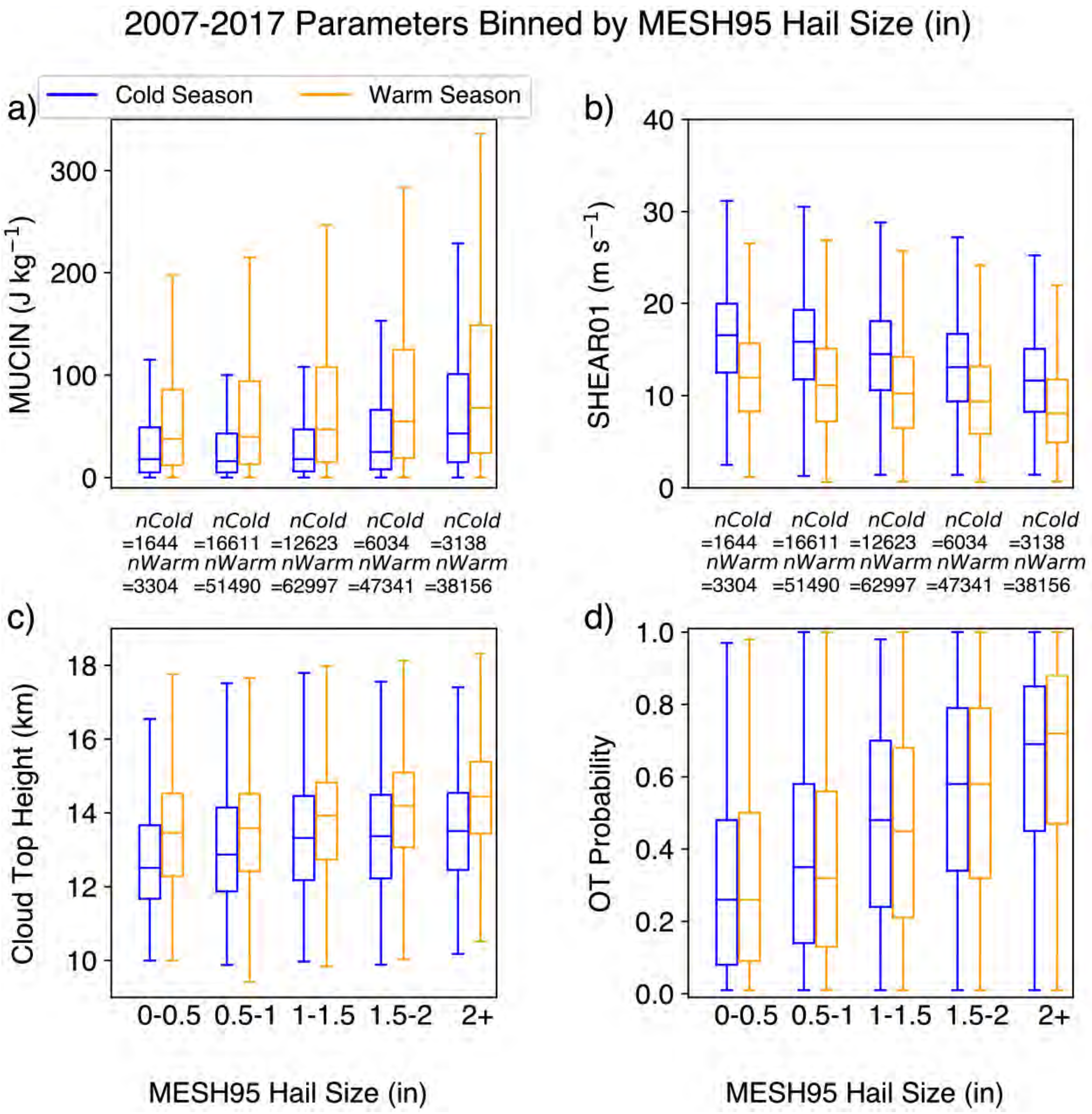

Fig. 10. Box and whisker distributions of select MERRA2 and GOES-12/13 parameters during the cold season (blue) and warm season (orange) stratified by MESH95 hail size bins (matched within ±7.5 minutes and 28×28 km^2) during 2007-2017. a) MUCIN, b) SHEAR01, c) Cloud Top Height, and d) OT Probability. The number of matches within each bin are listed below panels a) and b).

It is important to note that the performance of the DNNs with reanalysis-only inputs are biased toward better predictions because reanalysis parameters are only being extracted at ECS locations where convection is actively occurring. Situations not reflected in these statistics are cases where model parameters might have indicated a favorable severe storm environment, but no storms formed. Therefore, the importance of the satellite measurements alone in precisely identifying areas of convection within these favorable environments is perhaps more significant than what is apparent from these results. With this consideration in mind, the experiments that rely only on reanalysis input have comparable performance to that of GOES-13 Only. Model superiority depends on the position along the ROC and PD curves, but, considering only the optimal *p* values, the MERRA-2 Only and ERA5 Only models score marginally better than the GOES-13 Only model. Performance is notably better, however, when GOES-12/13 inputs are combined with reanalysis.

d. DNN severe hail climatology and case studies

The DNN model is applied to the 11-year database of merged MERRA-2 and GOES-12/13 parameters at ECS pixels to derive severe hail climatologies. ECS or OT pixels with varying hail likelihoods are aggregated at the 80×80-km^2 grid scale, matching the grid size of hail climatology truth datasets shown in Figure 3. The spatial variability and frequency of confident OT detections with p>50% likelihood of producing severe hail (Fig. 11b) highlights a region of maximum occurrence (~20-22 events year^{-1}) across the Central Plains, patterns that are supported by the MESH95 climatology (see contours of 3 and 6 MESH95 hail events year^{-1} in Fig. 11). Hail-producing ECSs predicted by the DNN reveal frequent severe hail days (~50 events year^{-1}, Fig. 11a) in these regions – far exceeding MESH event days but keeping the same general spatial pattern. The excessive ECS event days may be tied to the fact that ECS with low OT Probability originate from cold outflow, from a tropopause-relative perspective, very near to true updraft regions (Cooney et al. 2021). Therefore, a true updraft OT may have several ECSs nearby that can trigger a moderately high hail probability, and the greater frequency and coverage of these ECSs inflate the counts. When we constrain to only the ECSs likely to be true updrafts (Fig. 11b), the event days are in much better agreement with that of MESH. When a higher likelihood threshold (p>75%) corresponding to greater MESH hail size (Fig. 7) is used to compile a climatology, event day counts show even closer relative agreement.

The patterns are also consistent with observations from SPC and the findings of the Meteorological Phenomena Identification Near the Ground (MPING) project, which crowd-sourced volunteers to report precipitation events (Fig. 3b; Fig. 6 in Elmore et al. 2022). Other DNN hail maxima occur over the Sierra Madre, northeastern Mexico, and the Gulf Stream (Fig. 11), which are outside the NEXRAD range but are coincident with regions of MWR hail detections (Fig. 3d). Although radar confirmation is not possible in these areas of Mexico, previous intense convection climatology studies depict local maxima in these regions (Edwards 2006; Zisper et al. 2006; Farfán et al. 2020).

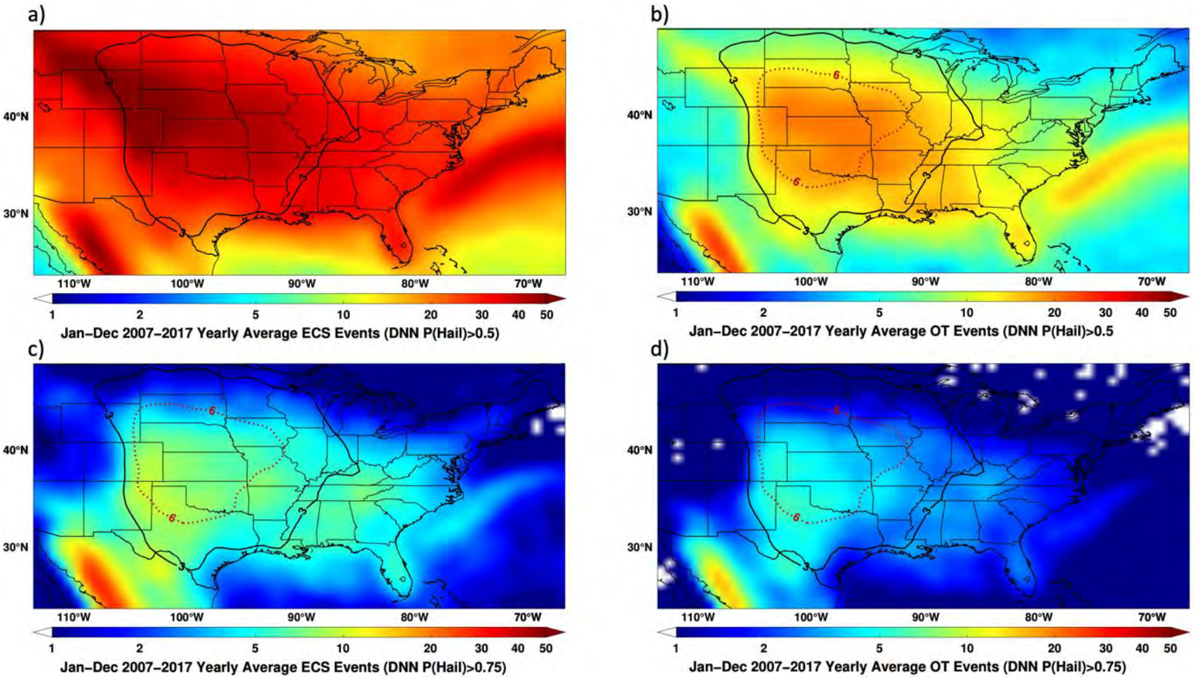
a)
b)
c)
d)
40°N
30°N
110°W
100°W
90°W
80°W
70°W
6
1 2 5 10 20 30 40 50
Jan–Dec 2007–2017 Yearly Average ECS Events (DNN P(Hail)>0.5)
Jan–Dec 2007–2017 Yearly Average OT Events (DNN P(Hail)>0.5
Jan–Dec 2007–2017 Yearly Average ECS Events (DNN P(Hail)>0.75)
Jan–Dec 2007–2017 Yearly Average OT Events (DNN P(Hail)>0.75)

Fig. 11. Yearly average GOES-12/13 ECS (left) and OT (right) event days detected during 2007-2017. a) ECSs filtered by >50% likelihood of severe hail, b) OTs filtered by >50% likelihood of severe hail, c) ECSs filtered by >75% likelihood of severe hail, and d) OTs filtered by >75% likelihood of severe hail. Black and dotted red contours show regions of 3 and 6 hail events year^{-1}, respectively, for MESH95 ≥ 1.5 inches.

Case studies reveal interesting relationships between satellite parameters, large-scale environmental conditions, DNN-predicted hail likelihood, ground spotter hail reports, and 5-minute potentially severe hail swaths from MESH95 (1.5+ inches), which allows for qualitative assessment of DNN performance. The first case, 16-17 May 2017, was initially highlighted in Fig. 1, although without showing consideration of the broader environmental conditions, where it is evident that the most prominent cells (Figs. 1c-1e) are generally consistent with areas of ~60%+ severe hail likelihood (Fig. 1f). Grey arrows in Fig. 1f highlight a range of likelihood estimates for comparably strong satellite parameters. This likelihood range, which shows as decreasing in the south-to-north direction, is driven by tendencies in the environment toward decreasing SHIP and MUCIN but increasing SHEAR01 and TPW at the time of the image (not shown) – patterns that are anti-correlated with MESH95 (Fig. 6).

In Fig. 12, compilations of select reanalysis and satellite variable maximum intensities are shown, summarizing the strongest convection during the event across the region and at each GOES gridded pixel. A large area of overshooting convection occurred on this day, evidenced by widespread swaths of IR–Tropopause < 0 K (Fig. 12b), OT Probabilities near 1.0 (Fig. 12c), and tropopause-relative cloud top heights up to 2-3 km above the tropopause

(Fig. 12f). MERRA-2 reveals coincident high SHIP and environmental stability along a front (Figs. 12a and 12e). It appears that the DNN is properly combining the GOES and MERRA-2 inputs to derive high hail likelihood (Fig. 12g) in regions where high MESH95 was observed and hail/tornadoes were reported (Figs. 12d and 12h), such as the Texas Panhandle and Minnesota/Wisconsin. Despite similarly intense cloud heights, temperatures, and overshooting tops elsewhere, such as in Iowa, the DNN hail likelihood was not as high because of weaker SHIP, thereby demonstrating the significant influence of this 3rd-most contributing model predictor (see Table 4).

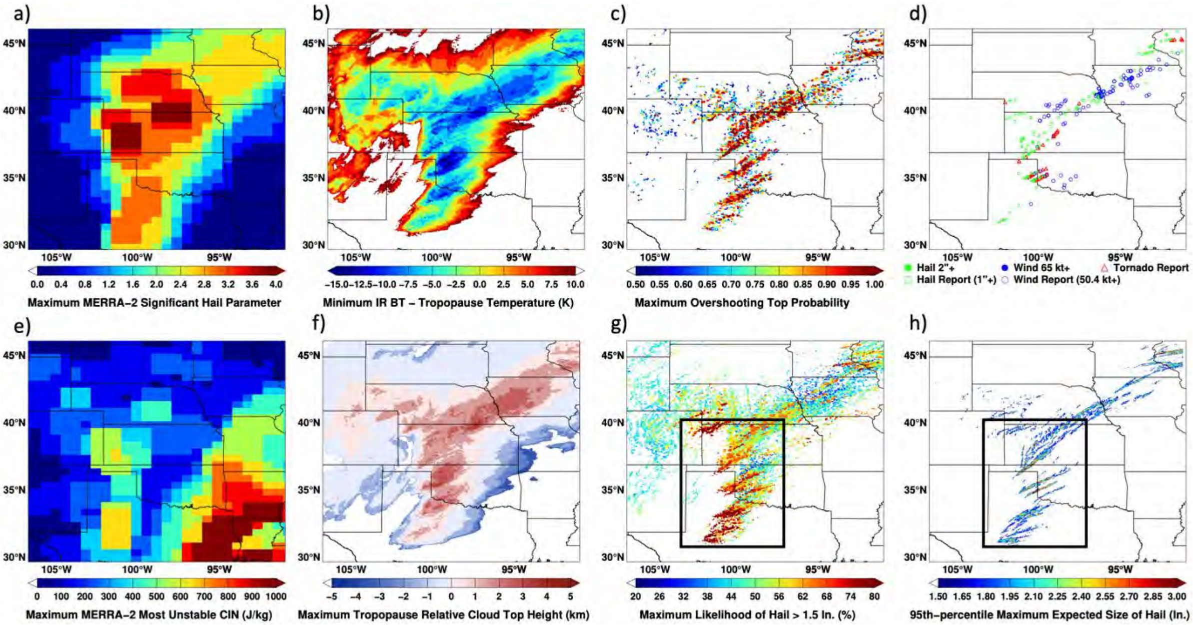

Fig. 12. Maximum intensity of parameters aggregated from 18 UTC 16 May 2017 through 6 UTC 17 May 2017. a) MERRA-2 SHIP, b) GOES-13 IR–Tropopause, c) GOES-13 OT Probability, d) SPC severe wind, hail, and tornado reports, e) MERRA-2 MUCIN, f) GOES-13 tropopause-relative cloud top height, g) DNN-predicted severe hail likelihood exceeding 20%, and h) 5-minute MESH95 exceeding 1.5 inches.

Zoomed views of the areas outlined in Figs. 12g and 12h are provided in the top panels of Fig. 13. Results from the DNN applied to scaled GOES-16 parameters (see Section 3.d.2), aggregated at GOES-13-like 15-min intervals, are also shown for this same case (Fig. 13b). GOES-16 hail likelihoods are comparable to those of Rapid Scan GOES-13 (Fig. 13a) in terms of spatial distribution of low and high likelihoods although with overall higher intensity, which may be owed to inherit imperfections in transferring parameters trained on scaled GOES-13 data to scaled GOES-16 data. On the other hand, the 15-min GOES-16 results appear to better represent actual MESH data, with fewer alarms in the western Texas Panhandle, central to eastern Kansas, and eastern Colorado, and more precise areal alignment with 1.5+ inch MESH95, thereby highlighting the importance of GOES data in the DNN

(Fig. 13d). The amplification of hail likelihood as a result of the increased GOES-16 spatial and temporal resolution warrants further study (Fig. 13c). That is, rather than applying imperfect scaling, a DNN trained using an extensive record of GOES-16 scans at the native 5-min-interval resolution should be developed for applications to GOES-16 measurements.

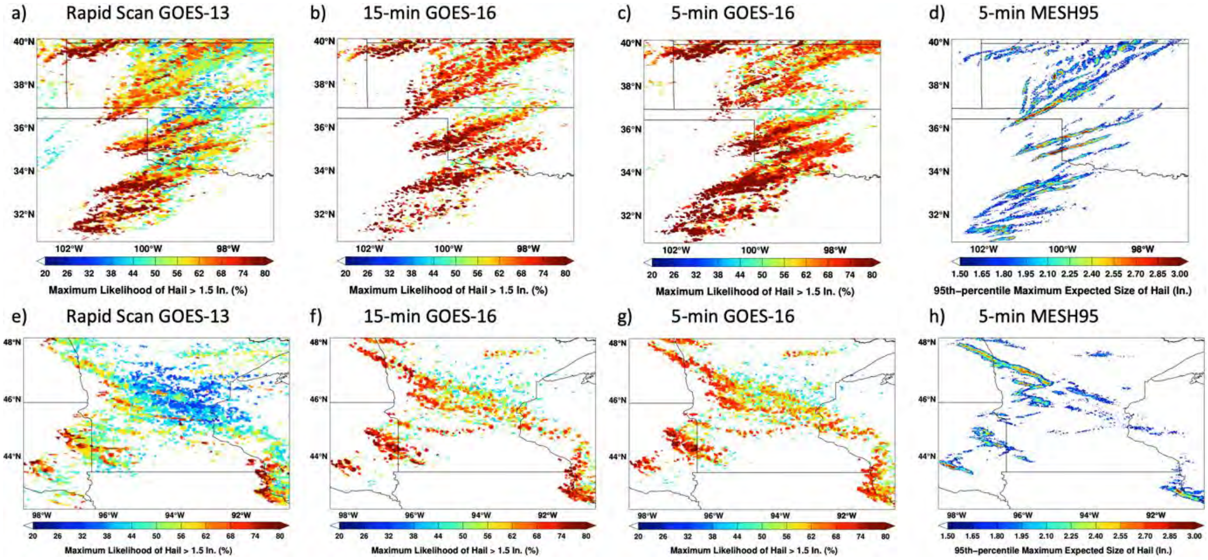
a) Rapid Scan GOES-13
b) 15-min GOES-16
c) 5-min GOES-16
d) 5-min MESH95
Maximum Likelihood of Hail > 1.5 In. (%)
95th-percentile Maximum Expected Size of Hail (In.)
e) Rapid Scan GOES-13
f) 15-min GOES-16
g) 5-min GOES-16
h) 5-min MESH95
Maximum Likelihood of Hail > 1.5 In. (%)
95th-percentile Maximum Expected Size of Hail (In.)

Fig. 13. Enhanced views of a-c) DNN-predicted severe hail likelihoods exceeding 20% and d) 5-minute MESH95 exceeding 1.5 inches for areas outlined in black in Figs. 12g and 12h, and, similarly, e-g) DNN-predicted severe hail likelihoods and h) 5-minute MESH95 for areas outlined in black in Figs. 14g and 14h. Likelihoods are determined using Rapid Scan GOES-13, 15-minute GOES-16, and 5-min GOES-16.

Another case study on 11 July 2017 highlights a summer-time hail, wind, and tornado event across Minnesota, Illinois, and the Dakotas (Fig. 14). Much of the reported severe hail for this case occurred in northern Minnesota, eastern South Dakota, eastern Iowa, and northern Illinois (Fig. 14d). The 5-minute MESH95, however, shows additional regions (isolated cells in Kansas) of potentially severe hail (1.5+ inches, Fig. 14h). SHIP and 0-1 km wind shear conditions supportive of severe convection (Figs. 14a and 14e) are aligned with the most intense observed overshooting cloud tops (Figs. 14b, 14c, and 14f) and severe weather reports (Fig. 14d). The GOES-13 DNN indicates severe hail likelihoods of ~60% and higher along the primary swaths of MESH95 potentially severe hail, in addition to isolated convective cells in the Central Plains (Figs. 14g and 14h). Figures 13e-13h show zoomed views of the areas outlined in Figs. 14g and 14h. As with the May case, the 15-min GOES-16 DNN application (Fig. 13 f) highlights overall higher intensity than that for the Rapid Scan GOES-13 case (Fig. 13e), again with better precision of the 15-min GOES-16 results in aligning with 1.5+ inch MESH95. That is, the GOES-16-based DNN application does well in showing well-aligned and enhanced likelihoods along true hail tracks, with fewer false alarms in Fig. 13f than in Fig. 13e, especially in central/northeast Minnesota and northeast Nebraska. The 5-min GOES-16 results (Fig. 13g) are comparable to those of the 15-min GOES-16 results in this case. DNN performance is also demonstrated for an early-season hail event (5 April 2017) in the Southeast US in Appendix B.

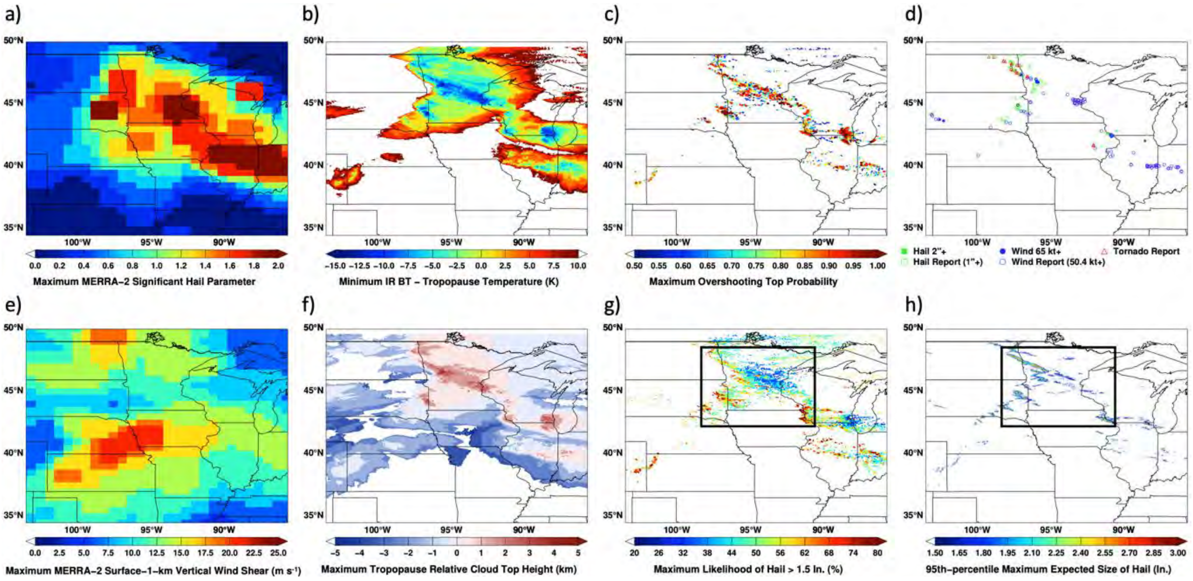

Fig. 14. Maximum intensity of parameters aggregated from 12 UTC 11 July 2017 through 8 UTC 12 July 2017. a) MERRA-2 SHIP, b) GOES-13 IR–Tropopause, c) GOES-13 OT Probability, d) SPC severe wind, hail, and tornado reports, d) MERRA-2 SHEAR01, f) GOES-13 tropopause-relative cloud top height, g) DNN-predicted severe hail likelihood exceeding 20%, and h) 5-minute MESH95 exceeding 1.5 inches.

5. Conclusions

Prediction of storms that produce potentially severe hail was accomplished through the training and application of a DNN that ties historical and present-day GOES satellite-based detections and intensity metrics at the satellite-detected updraft scale, together with reanalysis environmental characteristics, to NEXRAD maximum expected hail size (MESH) over the CONUS. The DNN generates severe hail likelihood at each satellite-detected embedded cold spot (ECS) detection, which serve as proxies for storm updraft regions. To the authors' knowledge, this is a first-of-its-kind study applying deep learning to merged reanalysis and individual pixel-scale satellite-derived parameters to estimate hail likelihood. This study emphasizes hailstorm detection using historical GOES-12/13 data for the purposes of assessing climatological hail risk, while also highlighting statistical performance, as well as case studies with GOES-13 data during a period in 2017 with overlapping GOES-16 data. Our goal is to create a model that can produce a hail likelihood for any satellite-identified ECS and coincident reanalysis inputs as a first step toward global applicability.

A set of parameters were identified that correlate with MESH95 and exhibit a relatively low amount of regional dependency, and recursive feature analysis was performed to rank their contribution to potentially severe hailstorm detection. These parameters encapsulate several GOES IR BT variables that define storm-top height and updraft intensity, 0-1 km wind shear, SHIP (which includes most unstable CAPE, mixing ratio, mid-level lapse rate, 500-hPa temperature, and 0-6-km wind shear), CIN, and TPW. It was found that these parameters are much better correlated with MESH95 maxima and MWR hail likelihood than observed hail size, an outcome attributed to uncertainties with hail size reporting. It was also shown that that DNN severe hail likelihood is strongly correlated with MESH. Various metrics, computed using *k*-fold cross-validation, were used to evaluate the DNN predictive capability. The performance of the experiments was shown to be dependent on what skill metrics are evaluated. The GOES-13 + MERRA-2 model demonstrates an exceptionally high-level and balanced performance compared to other recent similar efforts, with CSI of 0.511 and HSS of 0.407. The GOES-16 + MERRA-2 model performs comparably.

Various experiments with different DNN inputs and evaluations during the warm vs. cold seasons were conducted to further understand model capability. The results demonstrate choice of either MERRA-2 or ERA5 for reanalysis training has rather minimal impact on the final skill, although MERRA-2 does perform marginally better. Focusing on only warm season application improves the DNN performance based FA Ratio and FA Rate, but at the cost of reduced POD. The cold season success ratio is much lower than that of the warm season, likely due to the more dominant sampling of warm season environments and observations. This discrepancy could likely be mitigated with increased sampling and training specific to the seasons, an area that warrants further study.

Furthermore, it was shown that although a satellite-only DNN performs slightly worse than a reanalysis-only DNN at optimal *p* values. There is notable improvement when the satellite and reanalysis are used together. An important consideration here is the fact that models relying solely on reanalysis are biased toward better predictions because parameters are extracted where convection has already been detected by GOES. Therefore, the marginally improved performance of MERRA-2 or ERA5 Only models compared to a GOES-13 Only model is misleading. The DNN reproduced the general features of the 11-year CONUS severe hail climatology defined by MESH95, SPC, and satellite MWR, with distinct maxima in hail frequency over the Central Plains, the Midwest, and northwestern

Mexico. Good agreement between GOES-13/GOES-16 hail likelihoods and 5-minute resolution MESH95 severe hail is also demonstrated for several case studies. Fewer false alarms were apparent in these cases when the DNN is applied to GOES-16 data.

Many open questions remain regarding optimal experimental design for achieving maximum, generalized predictive skill, which can be better answered as sampling is improved. Future efforts to enhance DNN hail predictions may include: 1) using MWR hail probabilities to better understand how severe hail environments differ regionally and allow for model application/validation on a global scale, 2) incorporating additional satellite data like Geostationary Lightning Mapper flash characteristics/rate or visible channel parameters such as cloud top reflectance or spatial texture, 3) incorporating a mesoscale weather prediction model or reanalysis that can better resolve the environment ingested into storms, 4) modifying spatio-temporal interpolation of reanalysis data to the satellite grid (e.g., use the model timestep before a satellite ECS detection), and 5) training with a longer data record and incorporating 5-min data from GOES-16, coupled with the more recently available 5-min resolution GridRad Severe MESH dataset. Pursuing these advances should help further generalize the model and improve likelihood estimates, which is beneficial to those interested in better understanding hail risk across the globe. The release of the co-located GOES, reanalysis, and MESH95 dataset at the storm cell scale enables additional experimentation and improvements to a DNN model by the community.

Acknowledgments.

An open-source Python package, Tracking and Object-Based Analysis of Clouds (tobac; https://tobac.readthedocs.io/en/latest/index.html) v1.2, was modified to optimize the feature detection and watershed segmentation functionality to the hourly MESH95 data to determine locations of extrema and their areas. The University of Wisconsin Madison Space Science and Engineering Center McIDAS-V visualization and analysis software (https://www.ssec.wisc.edu/mcidas/software/v/) was used to quality-check the alignment of radar, model, and satellite parameters, and for figure creation. McIDAS-X was used to acquire and subset the GOES input data to derive convective cloud top parameters. The authors acknowledge the NASA Applied Sciences Disasters Program award (18-DISASTER18-0008), which provided the funding for this work. C. Homeyer acknowledges support from NASA grant 80NSSC20K1410. J. Allen acknowledges support from the National Science Foundation (AGS-1855054).

Data Availability Statement.

The matched ECS and MESH cell dataset used in this paper can be found at https://science-data.larc.nasa.gov/LaRC-SD-Publications/2023-01-25-001-BRS/. Severe storm reports are available through NOAA's Storm Prediction Center (https://www.spc.noaa.gov/climo/reports/). ECMWF Reanalysis v5 reanalysis data were downloaded from the Copernicus Climate Change Service Climate Data Store (https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels?tab=overview). MERRA-2 data were provided by the NASA GMAO (https://disc.gsfc.nasa.gov/datasets/M2T1NXSLV_5.12.4/summary). Environmental parameters were derived using Python package, xcape (https://github.com/xgcm/xcape) Hourly GridRad volumes used to derive MESH are available online (http://weather.ou.edu/~chomeyer/MESH/). Five-minute GridRad Severe volumes used to derive Severe MESH are available online (http://weather.ou.edu/~chomeyer/GridRad_Severe_MESH/).

APPENDIX

Appendix A: Parameter sensitivity to imager resolution, hail report type, and season

GOES-13 and GOES-16 parameters are stratified by MESH95 bins to assess the sensitivity to imager resolution (Fig. A1). Convective intensity increases with MESH95 for all four parameters shown, although there are differences in the GOES-13 and GOES-16 distributions. The median IR–Tropopause ranges from ~0 (3) K in the 0.5-1.0-inch bin to -6 (-3) K in the 2+ inch bin for GOES-16 (GOES-13), revealing a comparable relative sensitivity of GOES-16 IR–Tropopause to MESH95 but starting at a lower temperature, likely owing to the ability of a finer resolution imager to resolve colder IR pixels from intense convection (Khlopenkov et al. 2021; Cooney et al. 2021). ECS area is greater for GOES-13 than GOES-16 because of the greater pixel size of the former.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-22-0042.1.

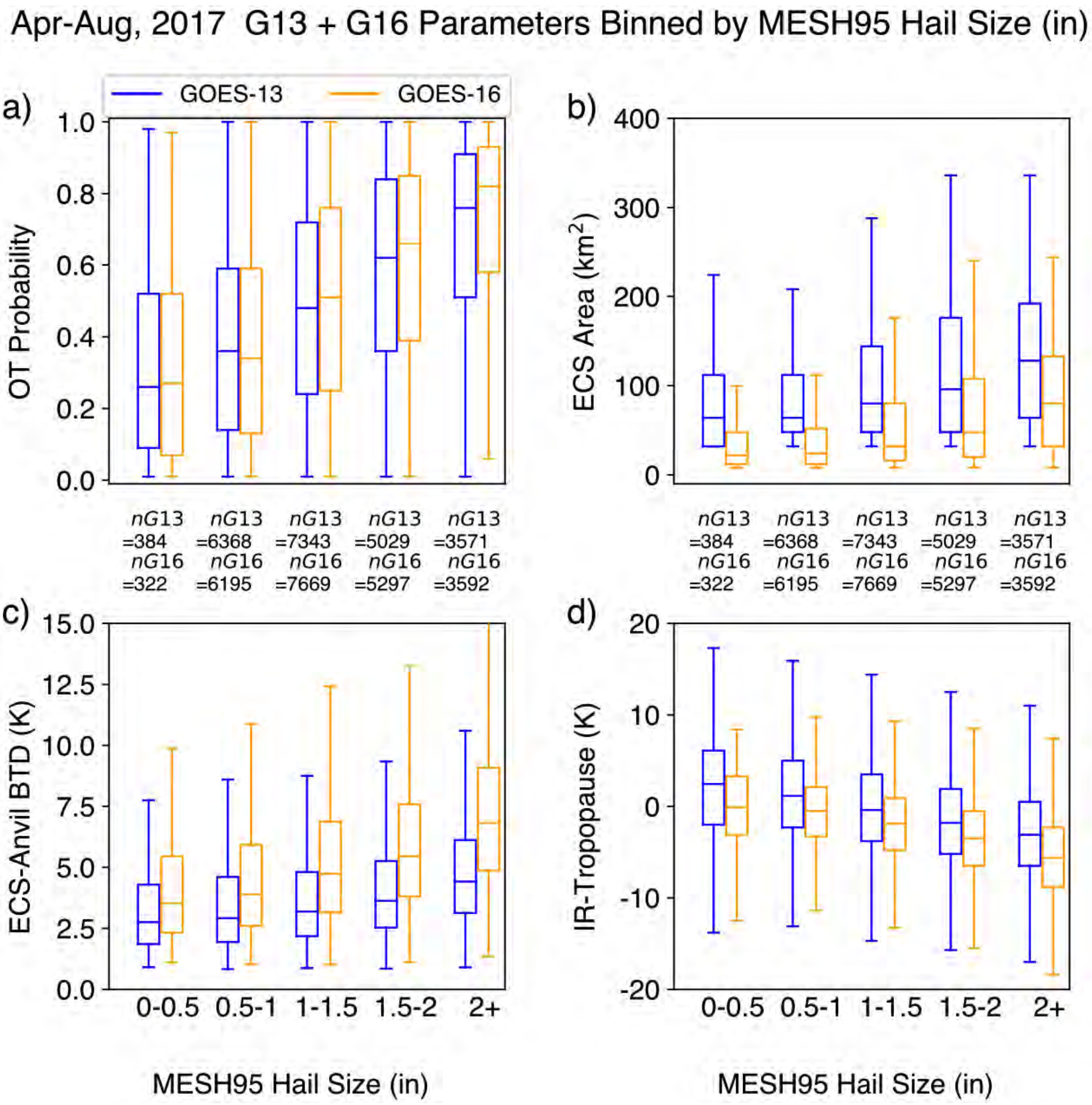

Fig. A1. Box and whisker distributions of select GOES-13 (blue) and GOES-16 (orange) IR cloud top parameters stratified by MESH95 hail size bins (matched within ±7.5 minutes and 30×30 km^2) during the 2017 warm season. a) OT Probability, b) ECS area, c) ECS–Anvil BTD, and d) IR–Tropopause. Number of matches within each bin are displayed below panels a) and b).

To compare the sensitivity of these GOES-12/13 parameters to other hail detection datasets, Fig. A2 stratifies GOES-12/13 parameters as a function of MWR hail probability. Although there is a significantly reduced sample size relative to MESH95 analyses (Fig. 4), the GOES-12/13 parameters exhibit similar increases in intensity as the MWR hail probability increases. The IR–Tropopause IQR in the lowest (highest) hail probability bin ranges roughly between -2.5 K and 2.5 K (-5.5 K and -2.5 K), revealing comparable sensitivity as that found with MESH95, but with greater (more negative) median intensity for

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0042.1.

matches across all bins. This implies that MWR hail detections are associated with more intense overshooting convection than MESH cell detections.

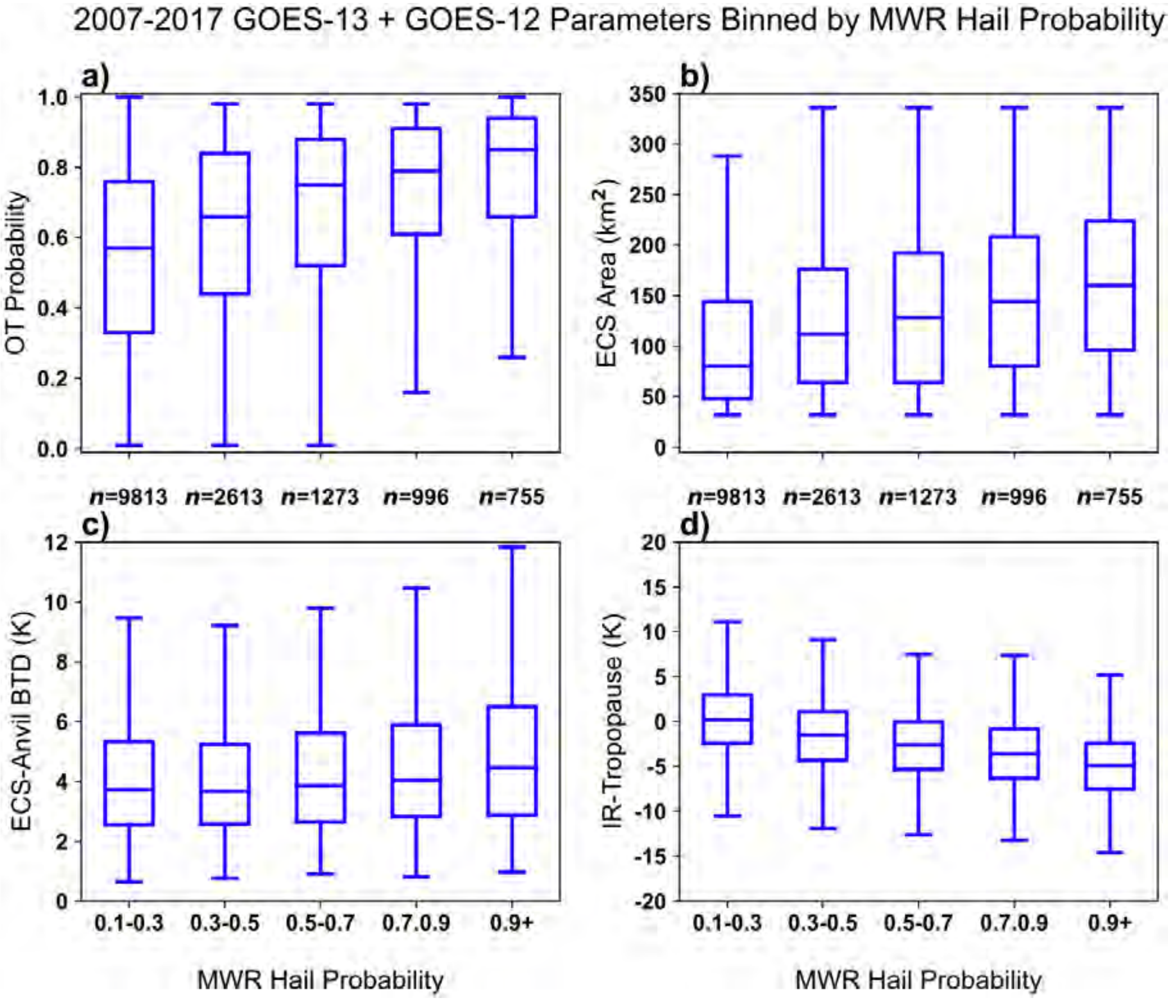

Fig. A2. Box and whisker distributions of select GOES-12/13 IR cloud top parameters stratified by MWR hail probability bins (matched within ±7.5 minutes and 28×28 km^2) during 2007-2017. a) OT Probability, b) ECS area, c) ECS–Anvil BTD, and d) IR–Tropopause.

GOES-12/13 parameters are stratified by SPC hail size in Fig. A3. The median GOES-12/13 match with SPC reports for all hail size bins is colder than the tropopause (Fig. A3d) with greater than 0.6 median OT Probability (Fig. A3a). However, these high intensities come at the price of lower sensitivity to hail size and significant IQR overlap among bins. The median ECS Area reveals only ~20 km^2 sensitivity between the lowest (1.0-1.5 inches) and highest (3+ inches) reported hail size bins (Fig. A3b), the median ECS–Anvil BTD reveals only ~0.75-K sensitivity (Fig. A3c), and the median IR–Tropopause ranges between ~-1.5 K and -3.5 K (Fig. A3d). These results are consistent with the work of Murillo and Homeyer (2019) and Sandmael et al. (2019), who did not find notable correlation between GOES-13/14 parameters and reported hail size.

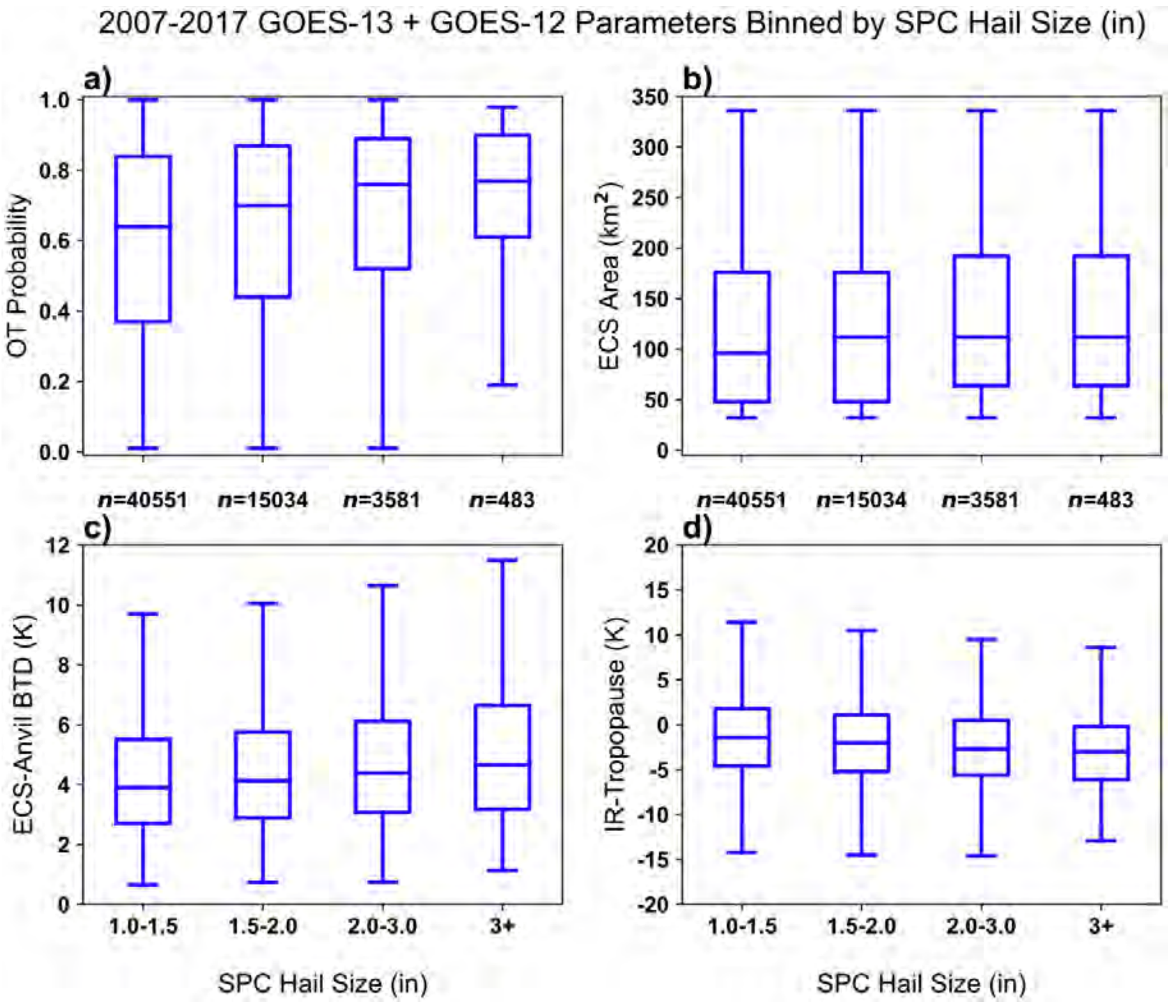

Fig. A3. Box and whisker distributions of select GOES-12/13 IR cloud top parameters stratified by ground spotter hail size report bins (matched within ±7.5 minutes and 28×28 km^2) during 2013-2017. a) OT Probability, b) ECS area, c) ECS–Anvil BTD, and d) IR–Tropopause.

Appendix B: Additional case study – 5 April 2017

A case study of an early season severe hail-, tornado-, and wind-producing event across the Southeastern US is shown in Fig. A4. Extremely large MERRA-2 SHIP (Fig. A4a) aligned with environments of elevated SHEAR01 (Fig. A4d). Combined, these environmental forcings contributed to formation of severe hailstorms across Alabama, Georgia, and, to a lesser extent, across areas of the Ohio River valley (Figs. A4d and A4h). GOES-13 reveals local regions of cloud tops colder and higher than the tropopause (Figs. A4b and A4f) aligned with OT Probabilities at or near 1 (Fig. A4c), contributing to good agreement between the DNN hail likelihood with swaths of severe reports and potentially severe MESH95 (Figs. A4d, A4g, and A4h). Isolated occurrences of potentially severe MESH95 in other areas are also correctly predicted as likely severe by the DNN. Enhanced views of Figs. A4g and A4h are shown in Fig. A5, combined with data from 15-min (Fig. A5b) and 5-min (Fig. A5c)

GOES-16. Figure A5 provides a better view of the agreement between Rapid Scan GOES-13 (Fig. A5a) and MESH95 swaths (Fig. A5d), the former of which is reasonably comparable to 5-min GOES-16 (Fig. A5b) along the strongest cell tracks, although with many false alarms elsewhere. The 15-min GOES-16 (Fig. A5c) application also shows more relative false alarms than we have seen from this result in the previous case studies. These minor shortcomings might be owed to the DNN's over-reliance on Great Plains hail MESH events, and therefore the environmental conditions for strong Southeast hail events might not be as well understood by the model.

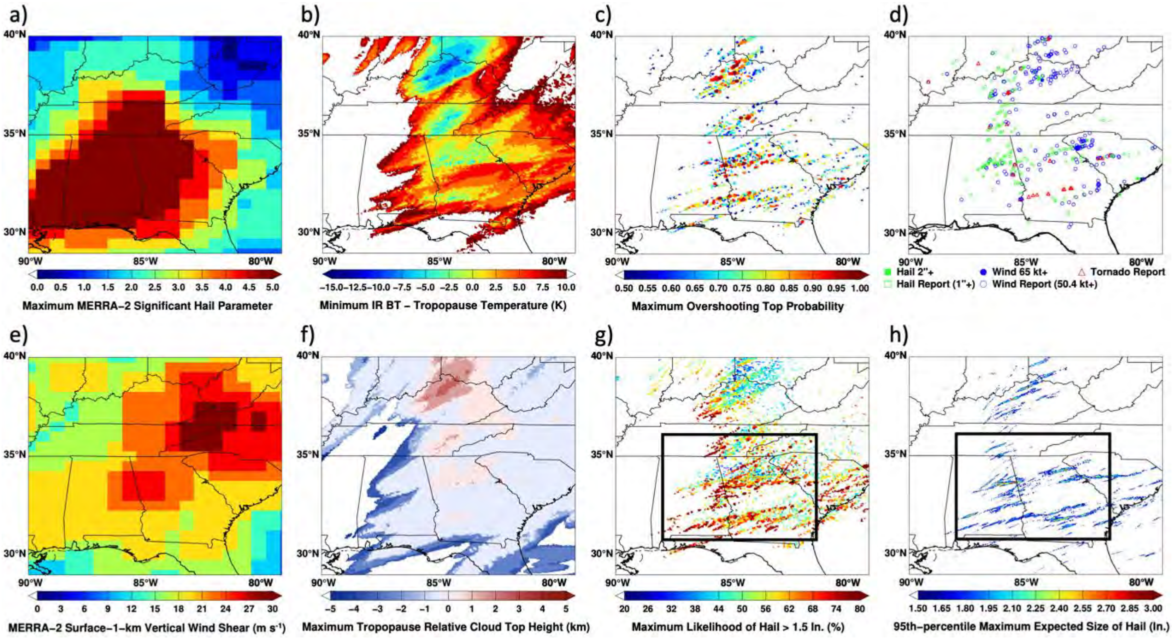

Fig. A4. Maximum intensity of parameters aggregated from 13 UTC 5 April 2017 to 6 UTC 6 April 2017. a) MERRA-2 SHIP, b) GOES-13 IR–Tropopause, c) GOES-13 OT Probability, d) SPC severe wind, hail, and tornado reports, e) MERRA-2 SHEAR01, f) GOES-13 tropopause-relative cloud top height, g) DNN-predicted severe hail likelihood exceeding 20%, and h) 5-minute MESH95 exceeding 1.5 inches.

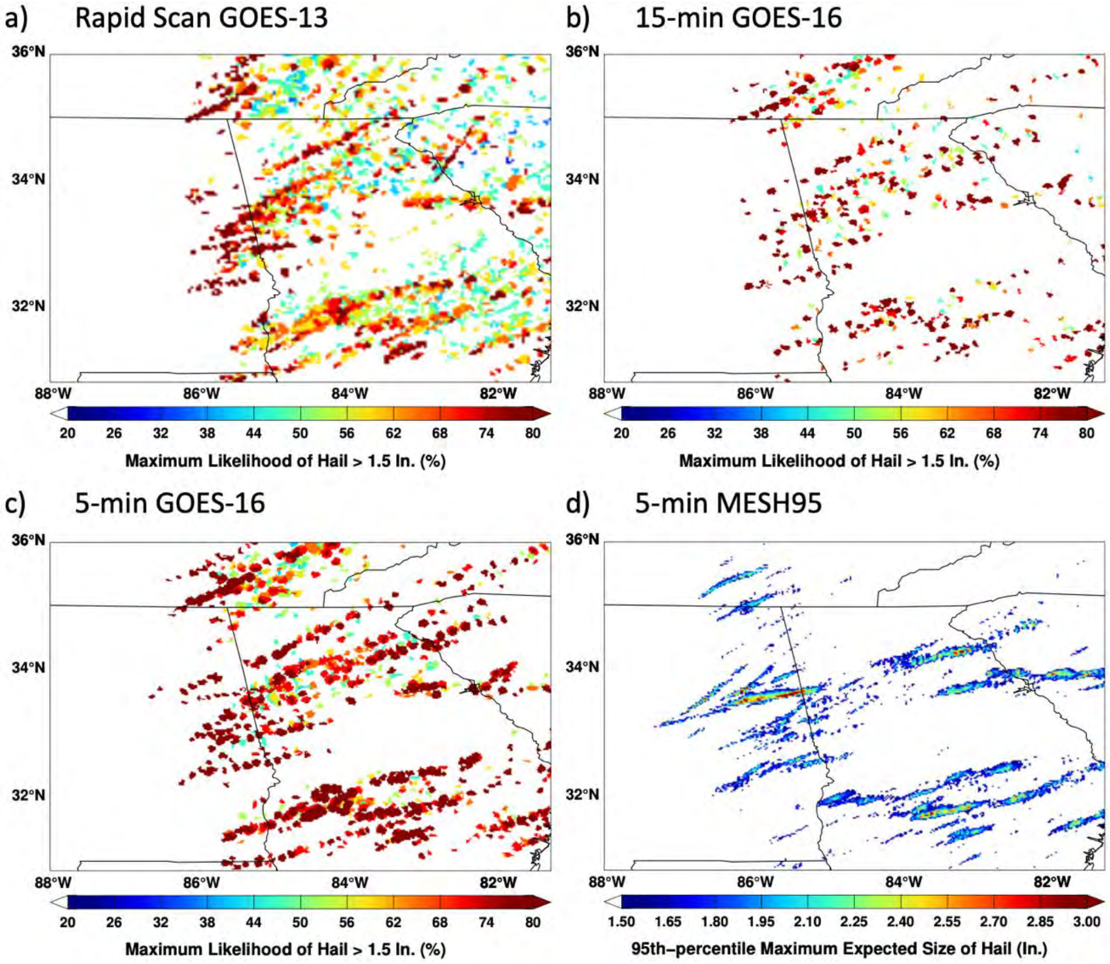

Fig. A5. Enhanced views of Figs. A4g and A4h black outlined areas, starting from 17 UTC 5 April 2017 to 6 UTC 6 April 2017, showing a-c) DNN-predicted severe hail likelihoods exceeding 20% and d) 5-minute MESH95 exceeding 1.50 inches Likelihoods are determined using Rapid Scan GOES-13, 15-minute GOES-16, and 5-min GOES-16.

REFERENCES

Allen, J. T. and M. K. Tippett, 2015: The characteristics of United States hail reports: 1955–2014. *Electron. J. Severe Storms Meteor.*, **10**, 575–595, http://www.ejssm.org/ojs/index.php/ejssm/article/viewArticle/149.

Allen, J.T., M. K. Tippett, and A. H. Sobel, 2015: An empirical model relating U.S. monthly hail occurrence to large-scale meteorological environment. *J. Adv. Model. Earth Syst.*, **7**, 226–243, https://doi.org/10.1002/2014MS000397.

Bang, S. D. and D. J. Cecil, 2021: Testing passive microwave-based hail retrievals using GPM DPR Ku-band radar. *J. Appl. Meteor. Climatol.*, **60**, 255–271, https://doi.org/10.1175/JAMC-D-20-0129.1.

Bang, S. D. and D. J. Cecil, 2019: Constructing a multifrequency passive microwave hail retrieval and climatology in the GPM domain. *J. Appl. Meteor. Climatol*., **58,** 1889–1904, https://doi.org/10.1175/JAMC-D-19-0042.1.

Barnston, A. G., 1992: Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score. *Notes and Correspondence*, **7**, 699–709, https://doi.org/10.1175/1520-0434(1992)007%3C0699:CATCRA%3E2.0.CO;2.

Barnes, L. R., D. M. Schultz, E. C. Gruntfest, M. H. Hayden, and C. C. Benight, 2009: CORRIGENDUM: False alarm rate or false alarm ratio? *Wea. Forecasting*, **24**, 1452–1454, https://doi.org/10.1175/2009WAF2222300.1.

Bedka K. M., J. T. Allen, H. J. Punge, M. Kunz, and D. Simanovic, 2018: A long-term overshooting convective cloud-top detection database over Australia derived from MTSAT Japanese advanced meteorological imager observations. *J. Appl. Meteor. Climatol.*, **57**, 937–951, https://doi.org/10.1175/JAMC-D-17-0056.1.

Bedka K. M. and K. Khlopenkov, 2016: A probabilistic multispectral pattern recognition method for detection of overshooting cloud tops using passive satellite imager observations. *J. Appl. Meteor. Climatol.*, **55**, 1983–2005, https://doi.org/10.1175/JAMC-D-15-0249.1.

Bengio, Y., 2012: Practical recommendations for gradient-based training of deep architectures. *Neural Networks: Tricks of the Trade, Lecture Notes in Computer Science*, *vol 7700*, G. Montavon, G. B. Orr, and K. R. Müller, Eds., Springer, Berlin, Heidelberg, 437–374, https://doi.org/10.1007/978-3-642-35289-8_26.

Bowman, K. P. and C. R. Homeyer, 2017: GridRad - three-dimensional dridded NEXRAD WSR-88D radar data. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, Boulder, CO., https://doi.org/10.5065/D6NK3CR7.

Brooks, H. E., 2009: Proximity soundings for severe convection for Europe and the United States from reanalysis data, *Atmos. Res.*, **93**, 546–553, https://doi.org/10.1016/j.atmosres.2008.10.005.

Brooks, H. E., C. A. Doswell III, and J. Cooper, 1994: On the environments of tornadic and nontornadic mesocyclones, *Wea. and Forecasting*, **9**, 606–618, https://doi.org/10.1175/1520-0434(1994)009%3C0606:OTEOTA%3E2.0.CO;2.

Brooks, H. E., J. W. Lee, and J. P. Craven, 2003: The spatial distribution of severe thunderstorm and tornado environments from global reanalysis data, *Atmos. Res.*, **67**, 73–94, https://doi.org/10.1016/S0169-8095(03)00045-0.

Bruick, Z. S., K. L. Rasmussen, and D. J. Cecil, 2019: Subtropical South American hailstorm characteristics and environments. *Mon. Wea. Rev.*, **147**, 4289–4304, https://doi.org/10.1175/MWR-D-19-0011.1.

Burke, A., N. Snook, D. J. Gagne II, S. McCorkle, and A. McGovern, 2020: Calibration of machine learning–based probabilistic hail predictions for operational forecasting, *Wea. and Forecasting*, **35**, 149–168, https://doi.org/10.1175/WAF-D-19-0105.1.

Cecil, D. J., 2009: Passive microwave brightness temperatures as proxies for hailstorms. *J. Appl. Meteor. Climatol.*, **48**, 1281–1286, https://doi.org/10.1175/2009JAMC2125.1.

Cecil, D. J. and C. B. Blankenship, 2012: Toward a global climatology of severe hailstorms as estimated by satellite passive microwave imagers. *J. Climate*, **25**, 687–703, https://doi.org/ 10.1175/JCLI-D-11-00130.1.

Cecil D. J., S. J. Goodman, D. J. Boccippio, E. J. Zipser, and S. W. Nesbitt, 2005: Three years of TRMM precipitation features. Part I: Radar, radiometric, and lightning characteristics. *Mon. Wea. Rev.*, **133**, 543–566, https://doi.org/10.1175/MWR-2876.1.

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, L. Cronce, and J. Brunner, 2020: NOAA ProbSevere v2.0—ProbHail, ProbWind, and ProbTor. *Wea. Forecasting*, **35**, 1523–1543, https://doi.org/10.1175/WAF-D-19-0242.1.

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, D. T. Lindsey, L. Cronce, J. Gerth, B. Rodenkirch, J. Brunner, and C. Gravelle, 2018: The NOAA/CIMSS ProbSevere model: Incorporation of total lightning and validation. *Wea. Forecasting*, **33**, 331–345, https://doi.org/10.1175/WAF-D-17-0099.1.

Cintineo, J. L., T. M. Smith, V. Lakshmanan, H. E. Brooks, and K. L. Ortega, 2012: An objective high-resolution hail climatology of the contiguous United States. *Wea. Forecasting*, **27**, 1235–1248, https://doi.org/10.1175/WAF-D-11-00151.1.

Coniglio, M. C. and R. E. Jewell, 2022: SPC mesoscale analysis compared to field-project soundings: Implications for supercell environment studies. *Mon. Wea. Rev.*, **150**, 567–588, https://doi.org/10.1175/MWR-D-21-0222.1.

Coniglio, M. C. and M. D. Parker, 2020: Insights into supercells and their environments from three decades of targeted radiosonde observations. *Mon. Wea. Rev.*, **148**, 4893–4915, https://doi.org/10.1175/MWR-D-20-0105.1.

Cooney J. W., K. M Bedka, K. P. Bowman, K. V. Khlopenkov, and K. Itterly, 2021: Comparing tropopause-penetrating convection identifications derived from NEXRAD and GOES over the contiguous United States. *J. of Geophys. Res.: Atmos.*, **126**, https://doi.org/10.1029/2020JD034319.

Czernecki, B., Taszarek, M., Marosz, M., Półrolniczak, M., Kolendowicz, L., Wyszogrodzki, A., and Szturc, J., 2019: Application of machine learning to large hail prediction-The importance of radar reflectivity, lightning occurrence and convective parameters derived from ERA5. *Atmospheric Research*, **227**, 249–262, https://doi.org/10.1016/j.atmosres.2019.05.010.

Dennis, E. J. and M. R. Kumjian, 2017: The Impact of Vertical Wind Shear on Hail Growth in Simulated Supercells. *J. Atmos. Sci.*, **74**, 641–663, https://doi.org/10.1175/JAS-D-16-0066.1.

Doswell, C. A. III, H. E. Brook, and M. P. Kay, 2005: Climatological estimates of daily local nontornadic severe thunderstorm probability for the United States. *Wea. Forecasting*, **20**, 577–595, https://doi.org/10.1175/WAF866.1.

Dworak, R., K. Bedka, J. Brunner, and W. Feltz, 2012: Comparison between GOES-12 overshooting-top detections, WSR-88D radar reflectivity, and severe storm reports. *Wea. Forecasting*, **27**, 684–699, https://doi.org/10.1175/WAF-D-11-00070.1.

Edwards, R., 2006: Supercells of the Serranias Del Burro (Mexico). *Preprints, 23rd Conf. on Severe Local Storms*, St. Louis, MO, Amer. Meteor. Soc., P6.2, https://www.spc.noaa.gov/publications/edwards/delburro.pdf.

Elmore, K. L., J. T. Allen, and A. E. Gerard, 2022: Sub-severe and severe hail. *Wea. and Forecasting*, Early Online Release, https://doi.org/10.1175/WAF-D-21-0156.1.

Farfán, L. M., B. S. Barrett, G. B. Raga, and J. J. Delgado, 2020: Characteristics of mesoscale convection over northwestern Mexico, the Gulf of California, and Baja California Peninsula. *Inter. J. Climatol.*, **41**, E1062–E1084, https://doi.org/10.1002/joc.6752.

Fawcett, T., 2006: An introduction to ROC analysis. *Pattern Recognition Letters*, **27**, 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

Gagne, D. J. II, A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles, *Wea. and Forecasting*, **32**, 1819–1840, https://doi.org/10.1175/WAF-D-17-0010.1.

Gagne, D. J. II, S. E. Haupt, D. W. Nychka, and G. Thompson, 2019: Interpretable deep learning for spatial analysis of severe hailstorms, *Mon. Wea. Rev.*, **147**, 2827–2845, https://doi.org/10.1175/MWR-D-18-0316.1.

Gelaro, R., W. McCarty, M. J. Suárez, R. and Todling, 2017: The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2), *J. of Climate*, **30**. https://doi.org/10.1175/JCLI-D-16-0758.1.

Gensini, V. A. and M. K. Tippett, 2019: Global Ensemble Forecast System (GEFS) predictions of days 1–15 U.S. tornado and hail frequencies. *Geophys. Res. Lett.*, **46**, 2922–2930. https://doi.org/10.1029/2018GL081724.

Gensini, V. A., C. Converse, W. S. Ashley, and M. Taszarek, 2021: Machine learning classification of significant tornadoes and hail in the United States using ERA5 proximity soundings. *Wea. and Forecasting*, **36**, 2143–2160, https://doi.org/10.1175/WAF-D-21-0056.1.

Gerapetritis, H., J. M. Pelissier, and S. C. Greer, 1995: The critical success index and warning strategy. 17th Conference on Probability and Statistics in the Atmospheric Sciences, https://ams.confex.com/ams/84Annual/techprogram/paper_70691.htm.

Glorot, X., A. Bordes, and Y. Bengio, 2010: Deep sparse rectifier neural networks. *J. of Mach. Learn. Re.*, **15**, 315–323, https://proceedings.mlr.press/v15/glorot11a/glorot11a.pdf.

Goodfellow, I., Y. Bengio, and A. Courville, 2020: Deep learning, *MIT Press*, http://www.deeplearningbook.org.

Griffin, S., K. M. Bedka, and C. Velden, 2016: A method for calculating the height of overshooting convective cloud tops using satellite-based IR imager and CloudSat Cloud Profiling Radar observations. *J. Appl. Meteor. Climatol.*, **55**, 479–491, https://doi.org/10.1175/JAMC-D-15-0170.1.

Gunturi, P., and M. K. Tippett, 2017: Managing severe thunderstorm risk: Impact of ENSO on U.S. tornado and hail frequencies. *Tech. Rep.*, WillisRe.

He, H., 2009: Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.*, **21**, 1263–1284, https://doi.org/10.1109/TKDE.2008.239.

Heikenfeld, M., P. J. Marinescu, M. Christensen, D. Watson-Parris, F. Senf, S. C. van den Heever, and P. Stier, 2019: tobac 1.2: towards a flexible framework for tracking and analysis of clouds in diverse datasets. *Geosci. Model Dev.*, **12**, 4551–4570, https://doi.org/10.5194/gmd-12-4551-2019.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quar. J. of the Roy. Met. Soc.*, **146**, 1999–2049. https://doi.org/10.1002/qj.3803.

Homeyer, C. R. and M. R. Kumjian, 2015: Microphysical characteristics of overshooting convection from polarimetric radar observations. *J. Atmos. Sci.*, **72**, 870–891, https://doi.org/10.1175/JAS- D-13-0388.1.

Homeyer, C. R. and K. P. Bowman, 2017: Algorithm description document for version 3.1 of the three-dimensional gridded NEXRAD WSR-88D radar (GridRad) dataset. *Tech. Rep.*, 23 pp., http:// gridrad.org/pdf/GridRad-v3.1-Algorithm-Description.pdf.

Hyvärinen, O., 2014: A probabilistic derivation of Heidke skill score. *Wea. Forecasting*, **29**, 177–181, https://doi.org/10.1175/WAF-D-13-00103.1.

Johnson, A. 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**, 1–42.

Khlopenkov, K.V., K. M. Bedka, J. W. Cooney, and K. Itterly, 2021: Recent advances in detection of overshooting cloud tops from longwave infrared satellite imagery. *J. of Geophys. Res.: Atmos.*, **126**, https://doi.org/10.1029/2020jd034359.

King, A. T. and A. D. Kennedy, 2019: North American supercell environments in atmospheric reanalyses and RUC-2. *J. Appl. Meteor. Climatol.*, **58**, 71–92, https://doi.org/10.1175/JAMC-D-18-0015.1.

Kingma, D. P. and J. L. Ba, 2015: Adam: A method for stochastic optimization. *Proc. ICLR,* https://doi.org/10.48550/arXiv.1412.6980.

Kunz, M., 2007: The skill of convective parameters and indices to predict isolated and severe thunderstorms. *Nat. Hazards Earth Syst. Sci.*, **7**, 327–342, https://doi.org/10.5194/nhess-7-327-2007.

Lazzara, M.A., J.M. Benson, R.J. Fox, D.J. Laitsch, J.P. Rueden, D.A. Santek, D.M. Wade, T.M. Whittaker, and J.T. Young, 1999: The man computer interactive data access system: 25 years of interactive processing. *Bull. Amer. Meteor. Soc.,* **80**, 271–284, https://doi.org/10.1175/1520-0477(1999)080<0271:TMCIDA>2.0.CO;2.

LeCun Y., L. Bottou, G. B. Orr, and K. R. Müller, 1998: Efficient BackProp. *Neural Networks: Tricks of the Trade, Lecture Notes in Computer Science, vol 1524*, G. B. Orr, and K. R. Müller, Eds., Springer, Berlin, Heidelberg, 9–50, https://doi.org/10.1007/3-540-49430-8_2.

Lepora, C., R. Abernathey, N. Henderson, J. T. Allen, and M. K. Tippett, 2021: Future global convective environments in CMIP6 models. *Earth's Future*, **9**, https://doi.org/10.1029/2021EF002277.

Lin., T.-Y., P. Goyal, R. Girshick, K. He, and P. Dollár, 2020: Focal loss for dense object detection. *IEEE Trans. on Patt. Analy. and Mach. Intel.*, **42**, 318–327, https://doi.org/10.1109/TPAMI.2018.2858826.

Liu, C., E. J. Zipser, D. J. Cecil, S. W. Nesbitt, and S. Sherwood, 2008: A cloud and precipitation feature database from nine years of TRMM observations. *J. Appl. Meteor. Climatol.*, **47**, 2712–2728, https://doi.org/10.1175/2008JAMC1890.1.

Marion, G. R., R. J. Trapp, and W. Nesbitt, 2019: Using overshooting top area to discriminate potential for large, intense tornadoes. *Geo. Res. Lett.*, **46**, 12520–12526, https://doi.org/10.1029/2019GL084099.

Masters, D. and C. Luschi, 2018: Revisiting small batch training for deep neural networks. *Graphcore Research*, 18 pp., https://doi.org/10.48550/arXiv.1804.07612.

Mecikalski, J. R., T. N. Sandmæl, E. M. Murillo, C. R. Homeyer, K. M. Bedka, J. M. Apke, and C. P. Jewett, 2021: A random-forest model to assess predictor importance and nowcast severe storms using high-resolution radar–GOES satellite–lightning observations. *Mon. Wea. Rev.*, **149**, 1725–1746. https://doi.org/10.1175/MWR-D-19-0274.1.

Murillo, E., C. Homeyer, and J. T. Allen, 2021: A 23-year severe hail climatology using GridRad MESH observations. *Mon. Wea. Rev.*, **149**, 945–958, https://doi.org/10.1175/MWR-D-20-0178.1.

Murillo, E. M. and C. R. Homeyer, 2019: Severe hail fall and hailstorm detection using remote sensing observations. *J. Appl. Meteor. Climatol.*, **58**, 947–970. https://doi.org/10.1175/JAMC-D-18-0247.1.

NASA Earth Science Applied Sciences Disasters, 2021: 2021 annual summary. Accessed 1 Feb 2022, https://appliedsciences.nasa.gov/sites/default/files/2022-03/NASA%20Disasters%202021%20Annual%20Summary.pdf.

NOAA, 2022: Significant Hail Parameter (SHiP). Accessed 1 Feb 2022, https://www.spc.noaa.gov/exper/soundings/help/ship.html.

Nesbitt, S. W., E. J. Zipser, and D. J. Cecil, 2000: A census of precipitation features in the tropics using TRMM: Radar, ice scattering, and lightning observations. Mon. Wea. Rev., 13, 4087–4106, https://doi.org/10.1175/1520-0442(2000)013,4087:ACOPFI.2.0.CO;2.

Ortega, K. L., 2018: Evaluating multi-radar, multi-sensor products for surface hail-fall diagnosis. *Electronic J. Severe Storms Meteor.*, **13**, 1–36, https://ejssm.org/archives/wp-content/uploads/2021/09/vol13-1.pdf.

North American Hail Workshop Panel Discussion: A new view: Hail science through the lens of early career scientists. 2022 North American Workshop on Hail & Hailstorms, Boulder, CO, Accessed 5 January 2023, https://www.youtube.com/watch?v=OjVq4qyi5tQ&list=PLHgkMmlD5xYULMkfYRa5yMHqY1fBchpYo&index=2&t=126s.

Prechelt, L., 1998: Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks*, **11**, 761–767, https://doi.org/10.1016/S0893-6080(98)00010-0.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-22-0042.1.

Prein, A. F. and G. J. Holland, 2018: Global estimates of damaging hail hazard. *Weather and Climate Extremes*, **22**, 10–23. https://doi.org/10.1016/j.wace.2018.10.004.

Púčik, T., P. Groenemeijer, D. Rýva, and M. Kolář, 2015: Proximity soundings of severe and nonsevere thunderstorms in Central Europe. *Mon. Wea. Rev.*, **143**, 4805–4821, https://doi.org/10.1175/MWR-D-15-0104.1.

Punge, H. J., Bedka, K. M., Kunz, M., Bang, S. D., and Itterly, K. F., 2023: Characteristics of hail hazard in South Africa based on satellite detection of convective storms, Nat. Hazards Earth Syst. Sci., 23, 1549–1576, https://doi.org/10.5194/nhess-23-1549-2023, 2023.

Punge H. J., K. M. Bedka, M. Kunz, and A. Reinbold, 2017: Hail frequency estimation across Europe based on a combination of overshooting top detections and the ERA-INTERIM reanalysis. *Atmos. Res.*, **198**, 34–43, https://doi.org/10.1016/j.atmosres.2017.07.025.

Punge H. J., K. M. Bedka, M. Kunz, A. Werner, 2014: A new physically based stochastic event catalog for hail in Europe. *Nat. Hazards*, **73**, 1625–1645, https://doi.org/10.1007/s11069-014-1161-0.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Notes and Correspondence*, **24**, 601–608, https://doi.org/10.1175/2008WAF2222159.1.

Sandmæl T. N., C. R. Homeyer, K. M. Bedka, J. M. Apke, J. R. Mecikalski, and K. Khlopenkov, 2019: Evaluating the ability of remote sensing observations to identify significantly severe and potentially tornadic storms. *J. Appl. Meteor. Climatol.,* **58**, 2569–2590, https://doi.org/10.1175/JAMC-D-18-0241.1.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570–575, https://doi.org/10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2.

Taszarek, M., J. T. Allen, M. Marchio, and H. E. Brooks, 2021: Global climatology and trends in convective environments from ERA5 and rawinsonde data. *Climate and Atmos. Sci.*, **4**.

Taszarek, M., J. T. Allen, T. Púčik, K. A. Hoogewind, and H. E. Brooks, 2020: Severe convective storms across Europe and the United States. Part 2: ERA5 environments associated with lightning, large hail, severe wind and tornadoes. *J. Climate*, **33**, 10263–10286, https://doi.org/10.1175/JCLI-D-20-0346.1.

Taszarek, M., N. Pilguj, J. T. Allen, V. Gensini, H. E. Brooks, and P. Szuster, 2021: Comparison of convective parameters derived from ERA5 and MERRA-2 with rawinsonde data over Europe and North America. *J. Climate*, **34**, 3211–3237, https://doi.org/10.1175/JCLI-D-20-0484.1.

University of Oklahoma School of Meteorology, 2021: GridRad-Severe - Three-dimensional gridded NEXRAD WSR-88D radar data for severe events. *Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, Boulder, CO*., Accessed 1 Feb 2022, https://doi.org/10.5065/2B46-1A97.

Wendt, N. A. and I. L. Jirak, 2021: An hourly climatology of operational MRMS MESH-diagnosed severe and significant hail with comparisons to *Storm Data* hail reports. *Wea. Forecasting*, **36**, 645–659, https://doi.org/10.1175/WAF-D-20-0158.1.

Wilks, D. S., 2006: Statistical Methods in the Atmospheric Sciences. 2nd ed. International Geophysics Series, **100**, Academic Press, 648 pp.

Witt, A., M. D. Eilts, G. J. Stumpf, J. T. Johnson, E. D. W. Mitchell, and K. W. Thomas, 1998: An enhanced hail detection algorithm for the WSR-88D. *Wea. Forecasting*, **13**, 286–303, https://doi.org/10.1175/1520-0434(1998)013,0286:AEHDAF. 2.0.CO;2.

Yost, C. R., Bedka, K. M., Minnis, P., Nguyen, L., Strapp, J. W., Palikonda, R., Khlopenkov, K., Spangenberg, D., Smith Jr., W. L., Protat, A., and Delanoe, J., 2018: A prototype method for diagnosing high ice water content probability using satellite imager data, *Atmos. Meas. Tech.*, **11**, 1615–1637, https://doi.org/10.5194/amt-11-1615-2018.

Zhou, Z., Q. Zhang, J. T. Allen, X. Ni, and C. Ng, 2021: How many types of severe hailstorm environments are there globally? *Geophys. Res. Lett.*, **48**, https://doi.org/10.1029/2021GL095485.

Zisper, E. J., D. J. Cecil, C. Liu, S. W. Nesbitt, and D. P. Yorty, 2006: Where are the most intense thunderstorms on Earth? *Bull. Amer. Meteor. Soc.,* **87**, 1057–1072, https://doi.org/10.1175/BAMS-87-8-1057.