Evolution of WRF-HAILCAST during the 2014–16 NOAA/Hazardous Weather Testbed Spring Forecasting Experiments

REBECCA D. ADAMS-SELIN,[a] ADAM J. CLARK,[b] CHRISTOPHER J. MELICK,[c] SCOTT R. DEMBEK,[b,d] ISRAEL L. JIRAK,[e] AND CONRAD L. ZIEGLER[b]

[a] *Atmospheric and Environmental Research, Inc., Lexington, Massachusetts*
[b] *NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma*
[c] *557th Weather Wing/16th Weather Squadron, Offutt AFB, Nebraska*
[d] *Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma*
[e] *NOAA/NWS/NCEP/Storm Prediction Center, Norman, Oklahoma*

(Manuscript received 10 February 2018, in final form 9 November 2018)

ABSTRACT

Four different versions of the HAILCAST hail model have been tested as part of the 2014–16 NOAA Hazardous Weather Testbed (HWT) Spring Forecasting Experiments. HAILCAST was run as part of the National Severe Storms Laboratory (NSSL) WRF Ensemble during 2014–16 and the Community Leveraged Unified Ensemble (CLUE) in 2016. Objective verification using the Multi-Radar Multi-Sensor maximum expected size of hail (MRMS MESH) product was conducted using both object-based and neighborhood grid-based verification. Subjective verification and feedback was provided by HWT participants. Hourly maximum storm surrogate fields at a variety of thresholds and Storm Prediction Center (SPC) convective outlooks were also evaluated for comparison. HAILCAST was found to improve with each version due to feedback from the 2014–16 HWTs. The 2016 version of HAILCAST was equivalent to or exceeded the skill of the tested storm surrogates across a variety of thresholds. The post-2016 version of HAILCAST was found to improve 50-mm hail forecasts through object-based verification, but 25-mm hail forecasting ability declined as measured through neighborhood grid-based verification. The skill of the storm surrogate fields varied widely as the threshold values used to determine hail size were varied. HAILCAST was found not to require such tuning, as it produced consistent results even when used across different model configurations and horizontal grid spacings. Additionally, different storm surrogate fields performed at varying levels of skill when forecasting 25- versus 50-mm hail, hinting at the different convective modes typically associated with small versus large sizes of hail. HAILCAST was able to match results relatively consistently with the best-performing storm surrogate field across multiple hail size thresholds.

1. Introduction

Hail is a significant severe weather hazard in the United States. For example, four of the seven $1 billion dollar severe weather disasters that occurred in the United States in 2017 were due largely or entirely to hail damage, including one hailstorm that caused over $1.5 billion of property damage in Colorado alone (NCEI 2017). Yet, successfully forecasting hail occurrence remains difficult, as the processes involved in producing hail are not fully resolved by current convection-allowing models (CAMs). Recent research has attempted to solve this issue from multiple directions, including through the use of machine-learning algorithms trained on model data (Gagne et al. 2017) and a physically based one-dimensional hail model called HAILCAST that can be embedded within a CAM (Adams-Selin and Ziegler 2016).

HAILCAST has been run and tested as part of the NOAA Hazardous Weather Testbed (HWT) Spring Forecasting Experiments (SFEs) since 2014 (Jirak et al. 2014). Each year, participants have subjectively evaluated its performance, and the data itself were archived for verification. Modifications were made to the HAILCAST model each year in response to the feedback. This work details each year's verification results and participant feedback, and the modifications made to HAILCAST in response to that feedback and objective evaluations. Multiple verification methods are used to evaluate HAILCAST hail size predictions.

Corresponding author: Rebecca D. Adams-Selin, rselin@aer.com

DOI: 10.1175/WAF-D-18-0024.1

TABLE 1. Modifications made to each version of HAILCAST.

Version	Brimelow et al. (2002); Jewell and Brimelow (2009)	v2014	v2015	v2016	vRerun
No. of embryos	1	5	5	5	5
Embryo type	Liquid	Frozen	Frozen	Frozen	Frozen
Embryo size	300 μm	From graupel size distribution	10–50 μm	5–10 mm	5–10 mm
Embryo insertion level	Cloud base	First model level with graupel cooler than 0°C	First model level cooler than 0°C	−13°C	−8° and −13°C
Hailstone density	Fixed at 900 kg m^{-3}	Variable dry growth	Variable dry growth	Rime layer soaking and variable dry growth	Rime layer soaking and variable dry growth
Motion across updraft	Fixed in center	Fixed in center	Fixed in center	Parameterized horizontal motion	Parameterized horizontal motion
Ice collection efficiency	Step function	Step function	Step function	Linear function	Linear function
Below-cloud melting	Elevated soundings only	All profiles	All profiles	All profiles	All profiles
Liquid water profile	From 1D cloud model	WRF cloud water	WRF cloud water	WRF cloud water	Adiabatic
Output		Mean size	Mean size	Mean size	Max size

Section 2 discusses HAILCAST's structure and its yearly modifications. Section 3 reviews the data sources used for verification, including observational datasets and the HWT SFE ensembles. The two methods used to verify the forecasts are provided in section 4, and sections 5 and 6 discuss the results obtained from these two methods.

2. HAILCAST structure and changes

HAILCAST consists of a one-dimensional, time-dependent hail growth model designed to be fed updraft information from a CAM. For this study, HAILCAST has been embedded within the Weather Research and Forecasting (WRF) Model with the Advanced Research solver (Skamarock et al. 2008). A full description of WRF-HAILCAST is provided in Adams-Selin and Ziegler (2016). This study will focus on the different versions of WRF-HAILCAST run as part of the 2014–16 HWT SFEs. An overview of the 2015 SFE is available in Gallo et al. (2017), and operational plans for these experiments can be found online (https://hwt.nssl.noaa.gov/Spring_2014, https://hwt.nssl.noaa.gov/Spring_2015, and https://hwt.nssl.noaa.gov/Spring_2016).

Adams-Selin and Ziegler (2016) explains all the modifications made to the original HAILCAST model developed by Poolman (1992), Brimelow et al. (2002), and Jewell and Brimelow (2009), including incorporating multiple embryo sizes, the insertion of embryos at multiple temperatures above 0°C, variable density, parameterization of hailstone motion across the updraft, and improved ice collection efficiency and liquid water shedding thresholds. These modifications were incorporated over the course of 2014–16, using constructive feedback from each year's SFE. The major modifications incorporated each year are detailed in Table 1.

The 2014 version of HAILCAST (v2014 hereafter) used five percentile points along the microphysical graupel size distribution to determine initial embryo sizes. The lowest 1st, 2nd, 5th, 10th, and 20th percentiles were used. HAILCAST 2015 (v2015) replaced these five variable embryo sizes with five constant, but small, embryos between 10 and 50 μm, and inserted them at the first model level cooler than 0°C. HAILCAST v2016 introduced quite a few more modifications. Hailstones were no longer required to effectively remain in the center of the one-dimensional updraft, only falling out once they grew large enough to overcome the updraft. Instead, a time-dependent multiplier was applied to the updraft speed to roughly parameterize the horizontal motion of the hailstone across the updraft. Many observational and modeling studies have found hailstone embryos are typically located on the updraft edge and are then advected horizontally across the updraft (Heymsfield et al. 1980; Heymsfield 1982; Heymsfield and Musil 1982; Nelson 1983; Heymsfield 1983a,b; Ziegler et al. 1983; Foote 1984; Miller et al. 1990). Embryo sizes were changed to a range between 5 and 10 mm, more in line with observations of hailstone embryos made by Magono and Nakamura (1965), Heymsfield (1982), Heymsfield and Musil (1982), Ziegler et al. (1983), and Nelson and Knight (1987).

Additional changes were made to the algorithm in response to comments from the 2016 SFE, in a version termed vRerun. Details about the modeling setup used for vRerun are provided in the next section. The main modification was an attempt to account for the typical CAM horizontal grid spacing of 1–4 km not fully resolving the bounded weak-echo region (BWER) typically found within an updraft core of a hail-producing storm (Heymsfield and Musil 1982). The weak-echo region is largely devoid of precipitation-sized particles, resulting in an almost adiabatic cloud water profile. HAILCAST vRerun calculates an adiabatic cloud liquid water profile from the WRF vertical temperature and pressure profiles, and the water vapor mixing ratio at cloud base. Full details of this calculation are provided in Adams-Selin and Ziegler (2016). This adiabatic liquid water profile is used in vRerun instead of the WRF cloud liquid water profile.

3. Data sources

a. NSSL-WRF and CLUE ensembles

WRF-HAILCAST was run as part of the National Severe Storms Laboratory (NSSL) WRF Ensemble during all of the 2014, 2015, and 2016 SFEs. The NSSL-WRF (Kain et al. 2010) Ensemble is a WRF-ARW ensemble with 4-km horizontal grid spacing that is run daily at 0000 UTC over the continental (CONUS) with forecasts out to 36 h. The ensemble is a single-physics design and uses the WSM6 microphysics scheme (Hong and Lim 2006). Spread is given by variations in the initial and boundary conditions by using either the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) analysis or 2100 UTC Short-Range Ensemble Forecast (SREF; Du et al. 2014) 3-h forecasts. In 2014, the ensemble had 9 members but expanded to 10 members in 2015 and 2016.

In 2016, an additional, 3-km horizontal grid-spacing WRF ensemble was added, called the Community Leveraged Unified Ensemble (CLUE). A full description of CLUE's design is available in Clark et al. (2018). Similar to the NSSL-WRF ensemble, all members of CLUE are initialized at 0000 UTC, and forecasts covering a CONUS domain run out to 36 h or more. The results shown here are from the 10 "single physics" members of CLUE, which uses the Thompson microphysics parameterization for all members. These specific members were run by the Center for Analysis and Prediction of Storms (CAPS) and used radar data assimilation at initialization. Spread is provided by variations in the initial and boundary conditions, using SREF perturbations [see Clark et al. (2018) for further details].

When verifying both sets of ensembles, only forecast hours 12–36 (or 1200–1200 UTC the following day) are retained to allow time for convective spinup and to match with the SPC forecast period. The hail size field from each member was accumulated over the 24-h period by saving the maximum hail size at any grid point within that 24-h period. This accumulation was done to lessen the impact of WRF convective timing on HAILCAST skill scores, as convective timing issues are unrelated to HAILCAST skill.

b. Rerun

To test additional modifications made to the HAILCAST algorithm after the 2016 SFE, two 3-week periods during the 2014 and 2015 SFEs were rerun using WRF with vRerun of HAILCAST. The exact dates were 7–21 May 2014 and 4–22 May 2015. The domain, initialization time, forecast duration, and model parameterizations were all the same as those used for the control members of the NSSL-WRF ensemble during the 2014 and 2015 SFEs, with one exception: WRF version v3.8.1 was used instead of v3.4.1. Again, only forecast hours 12–36, or 1200–1200 UTC the next day, were used for verification, and the maximum hail size over the 24-h period was retained.

c. Storm Prediction Center forecasts

In addition to the model data listed above, the Storm Prediction Center (SPC) day 1 convective outlooks for hail were also included in the verification process, to examine if HAILCAST could provide any information in addition to the convective outlooks. The 0600 UTC outlook, valid from 1200 to 1200 UTC the following day, was used. Convective outlook probabilities are issued for selected intervals only: specifically, 5%, 15%, 30%, 45%, and 60%. The outlook probabilities were reprojected onto the NSSL-WRF 4-km grid for verification using nearest-neighbor interpolation.

d. Maximum estimated size of hail

Verification of hail size was performed using the NOAA/NSSL Multi-Radar Multi-Sensor (MRMS) maximum estimated size of hail (MESH) product (Smith et al. 2016). The MRMS MESH (MESH hereafter) is produced by using the MRMS 1-km multiradar reflectivity mosaic, and a power-law relationship found by Witt et al. (1998) and Lakshmanan et al. (2006) between radar reflectivity values above the melting level and hail observed with storms in Oklahoma and Florida. MESH has been found to correlate well with the spatial coverage of hailfall observed by higher-resolution datasets (Wilson et al. 2009; Cintineo et al. 2012). It also demonstrates skill when delineating surface hailfall into general size categories (Wilson et al. 2009; Ortega 2018), but MESH does not have a one-to-one correlation

between the observed and MESH-estimated hail sizes (see Ortega 2018, Fig. 20). All statistical evaluations will bin the MESH hail size data into categories of 19–24, 25–49, 50–74, and >75 mm instead of using it directly. Although these thresholds do differ slightly (<5 mm) from the values found by Ortega (2018) to most closely correspond to nonsevere, severe, and significantly severe hail, they do align with thresholds used by Gagne et al. (2017) and Adams-Selin and Ziegler (2016).

Witt et al. (1998) developed the MESH power-law relationship using only hail observations of 19 mm (0.75 in.) and larger (see their Fig. 8). Thus, in this study all MESH data used for verification were required to be at least 19 mm. Again, the dataset was accumulated over a 24-h period by taking the maximum hail size observed at any grid point during the period. The 0.01°-resolution MESH data were reprojected onto the NSSL-WRF domain by assigning the maximum hail size of the MESH product within a 2-km radius of each WRF grid point to that point. The reprojection process for the CLUE domain was similar except a 1.5-km radius was used. This ''maximum nearest neighbor'' interpolation method was used to ensure the largest hail size seen in the MESH dataset within each hail swath was preserved.

4. Verification methods

Verification of convective-based hazards, with their implicit high spatial and temporal variability, is a difficult proposition. Accumulating the datasets into 24-h periods helps eliminate some of the uncertainty due to temporal variability among observed and forecasted convection. However, spatial variability is still an issue. Hail forecasting relies not only on the successful prediction of hail size within convection, but also the successful prediction of the convection itself. However, only one of these factors can be controlled by an embedded hail model. To isolate these two factors, two different verification methods are used. The first method, object-based verification, uses the Method for Object-Based Diagnostic Evaluation (MODE; Davis et al. 2006a,b; available online at http://www.dtcenter.org/met/users) to match forecasted and observed swaths of hail size. By matching swaths across space, this verification method eliminates some of the dependency on correctly forecasted convection. The second method uses neighborhood grid-based verification.

a. Object-based verification

To utilize the object-based verification technique, the MESH dataset was accumulated over a 24-h period and reprojected onto the NSSL-WRF and CLUE model domains, as already described above, before being input into MODE. The MODE configuration chosen was designed to compare clusters of hail swath objects on the spatial scale of a swath produced by a single mesoscale convective system or a family of supercells [see Fig. 13 in Adams-Selin and Ziegler (2016) for an example of matched storm clusters]. Within each hail swath cluster, the maximum hail size was retained. Two-dimensional histogram plots showing the frequency of forecast versus observed maximum cluster hail sizes were created (e.g., Fig. 4). To construct these diagrams, bin counts for 25-mm hail size intervals were determined, and then the counts were normalized by the total number of objects for that year. A perfect forecast would cluster all bin counts along the 1-to-1 correlation line. Binned intervals were used instead of scatterplots because of the lack of skill MESH shows in discriminating among observed hail sizes at finer intervals (Cintineo et al. 2012; Ortega 2018). Thresholds of 25 and 50 mm (1 and 2 in.) were also used to compare observed and forecast cluster maximum hail sizes and construct typical contingency table statistics (Wilks 2006). For example, if matched forecast and observed clusters both had maximum sizes above 50 mm, that would be considered a hit; if only the forecast cluster had a maximum size above 50 mm, that would be considered a false alarm. These statistics can then be used to calculate the probability of detection (POD) and false alarm ratio (FAR) and plotted onto a performance diagram, as described by Roebber (2009).

In both instances, hail swath clusters were only compared if there were corresponding clusters in both the forecast and observed fields. By only evaluating matched clusters, the penalty for WRF failing to forecast convection is eliminated. In this respect the object-based verification is best characterized as an evaluation of the distribution of the forecasted hail size, not an evaluation of the occurrence of any hail.

b. Grid-based verification with Gaussian smoother

This verification technique was used to evaluate WRF and WRF-HAILCAST's hail forecasting skill as a whole, testing its abilities to both forecast convection and hail size within that convection. The technique used is the ''upscaling'' neighborhood approach described by Ben Bouallègue and Theis (2014) and in section 2b of Schwartz and Sobash (2017). The NSSL-WRF ensemble and vRerun 24-h period forecasts were remapped onto the approximately 80-km NCEP 211 grid by setting the value for each grid box equal to 1 if the forecast hail size at any point within that grid box was larger than a given threshold. An 80-km grid was selected to match SPC's severe weather outlooks, which predict the probability

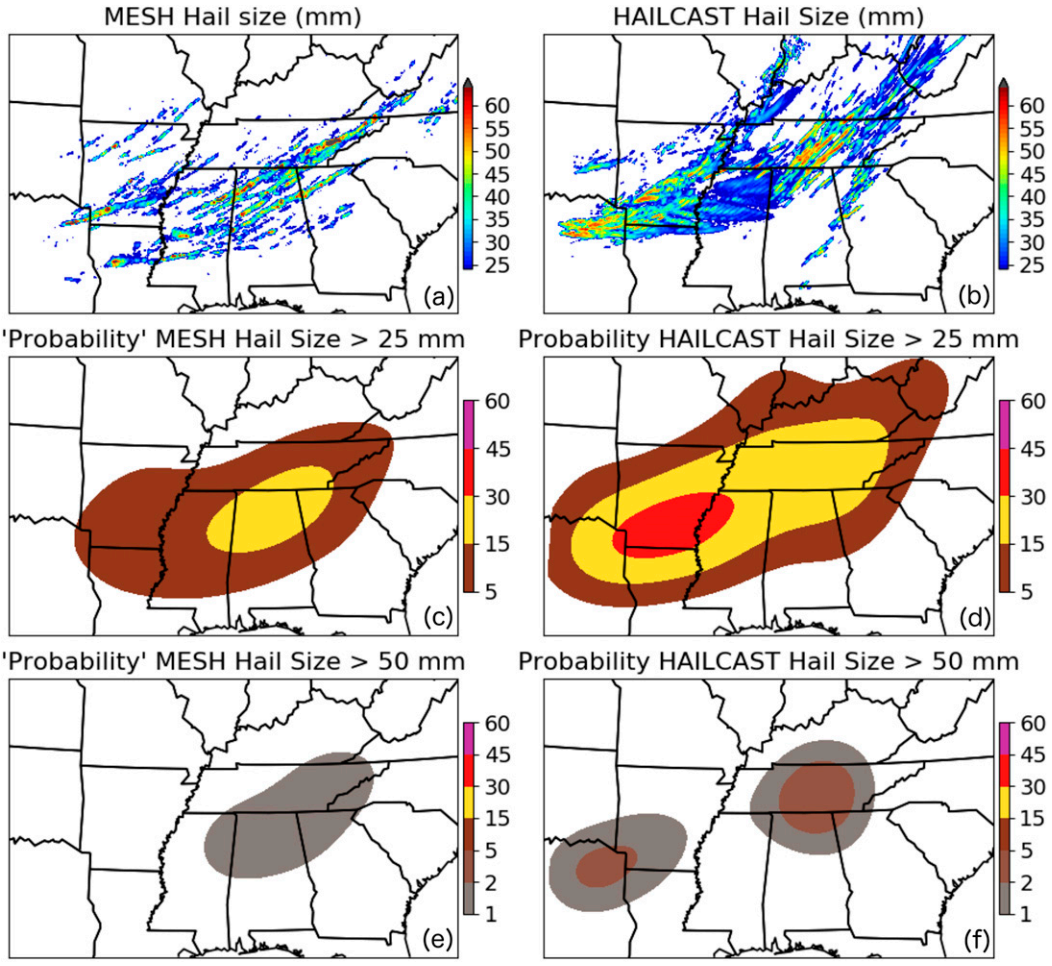

FIG. 1. Hail size (mm) from (a) MESH and (b) vRerun of HAILCAST from 0600 UTC 27 Apr to 0600 UTC 28 Apr 2011. Gaussian-smoothed probability of hail size larger than 25 mm from (c) MESH and (d) HAILCAST, and larger than 50 mm for (e) MESH and (f) HAILCAST.

of severe weather occurrence within a 40-km radius. The 80-km binary probabilities found for each ensemble member were then averaged, to create a forecast probability of the chosen threshold hail size for the ensemble. Finally, a two-dimensional Gaussian smoother with a standard deviation of 120 km was applied to the data [as in Sobash et al. (2011) and Hitchens et al. (2013)] to produce a forecasted probability field for a hail report of the given threshold size within 40 km. An observed ''probability'' field was created by feeding MESH data through the same process. Example 25- and 50-mm forecast and observed probability fields for a hail event on 27 April 2011, along with the HAILCAST vRerun forecast and MESH hail size, are shown in Fig. 1. The corresponding SPC convective outlook and storm reports for the same event are provided in Fig. 2.

When verifying the SPC probabilistic convective hail outlooks, the probabilities were remapped onto the same 80-km grid by using the maximum probability within that grid box. Those probabilities were not additionally treated by the Gaussian smoother. Because of the discrete nature of the forecast probabilities, a 5% forecast probability was considered verified if the observed Gaussian field was between 5% and 14%, a 15% probability between 15% and 29%, etc. For the observational dataset, the MESH hail size field was similarly remapped onto an 80-km grid by using the largest hail size within each grid box. The smoothed forecast probability field could then be directly compared to the resampled observed data.

Performance and attributes diagrams (Wilks 2006) were then constructed using 25- and 50-mm thresholds. In the attributes diagrams, climatology was calculated each year using the sample climatology. The Brier skill score (BSS) values (Murphy 1973) were calculated using climatology as the reference value.

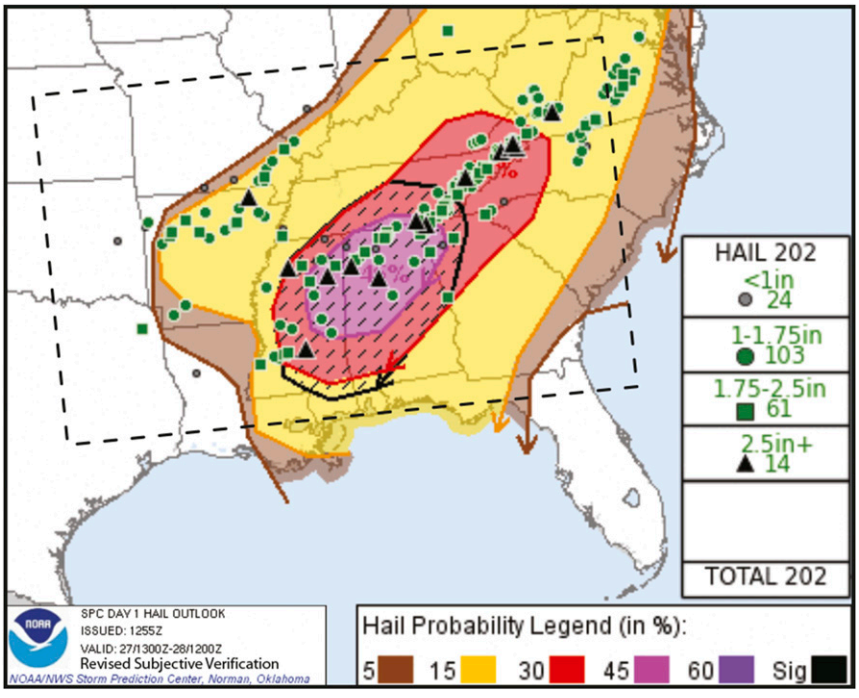

FIG. 2. SPC day 1 convective hail outlook and storm reports from 1300 UTC 27 Apr to 1200 UTC 28 Apr 2011.

c. Storm surrogate variables

Storm-surrogate fields such as updraft helicity have been shown to be successful proxies for predicting severe weather. For example, Sobash et al. (2016) showed ensemble surrogate severe probability forecasts (SSPFs) of updraft-helicity-produced forecasts with skill comparable to SPC outlooks. Gagne et al. (2017) found SSPFs of updraft helicity, column total graupel, and reflectivity at the −10°C level to produce positive levels of skill when specifically forecasting hail. Wendt et al. (2016) examined similar storm surrogate hail forecasts and found 2–5-km updraft helicity, using a tuned threshold, was the most successful during the 2012–15 HWTs. Thus, the question of whether or not HAILCAST can improve hail forecasts above the skill levels shown by these storm surrogate fields is valid.

Fields of maximum 2–5-km updraft helicity (UHMAX), column-maximum updraft speed (WMAX), maximum column-integrated graupel (COLGMAX), and 1-km reflectivity (DBZ1KMMAX) were all verified, along with the HAILCAST data, using the same two methods described above. "Hourly maximum fields" of each of these variables were output in each model output file, as in Kain et al. (2010). Maximum fields over the full 24-h period from 1200 to 1200 UTC were then calculated similarly to the hail field as described in section 3 and run through the object-based and grid-based verification methods as described above. A range of thresholds were examined for each surrogate field and are given later (see Table 3). Verification figures were then constructed using two sets of thresholds selected from the range. The first set, or "a priori," thresholds, were estimated from which thresholds showed the best results through a review of the literature (Sobash et al. 2011; Wendt et al. 2016; Gagne et al. 2017) and previous experience with storm surrogate parameters. This set of thresholds is meant to mimic thresholds an operator would choose upon first designing a model configuration (the italicized rows in Table 3). The second set of thresholds, or "best" thresholds, were chosen after the model runs as the thresholds that produced the highest BSS values over the entire period (boldface rows in Table 3). These thresholds are meant to represent the best possible predictive skill for storm surrogate fields.

For the MODE configuration file within the object-based verification method, both a "convolution radius" and a "convolution threshold" are required. The convolution radius is used to average over that number of grid points when creating a smoothed field; contiguous areas in the smoothed field that contain values larger than the convolution threshold are considered hail swath objects. A convolution radius of four grid points was selected, as recommended by Davis et al. (2006a,b). A convolution threshold of 12.5 mm (0.5 in.) was selected in agreement

TABLE 2. MODE configuration.

Configuration option	Setting
Convolution radius	Four grid points
Convolution threshold	0.5 in.
Area threshold	Four grid points
Max distance allowed between centroids	400 km

with Adams-Selin and Ziegler (2016) for HAILCAST; for the storm surrogate fields, convolution thresholds of $37.5\,m^2\,s^{-2}$ for UHMAX for a 3-km grid and $20\,m^2\,s^{-2}$ for a 4-km grid, $17.5\,m\,s^{-2}$ for WMAX, $12.5\,kg\,m^{-2}$ for COLGMAX, and 45 dBZ for DBZ1KMMAX were used. See Table 2 for additional details on the MODE configuration.

5. Results from object-based methods

The results from object-based verification of HAILCAST hail forecasts from all 2014–16 NSSL-WRF and CLUE members, as well as the rerun, are shown in Figs. 3 and 4. Figure 3a displays the skill of HAILCAST in forecasting hail swath objects with a maximum hail size of at least 50 mm. Determining HAILCAST's skill in forecasting objects with a maximum size of 25 mm is more difficult. As mentioned in section 3d, only MESH values of 19 mm and larger were used. Objects with maximum sizes between 19 and 25 mm occurred less frequently than objects with maximum sizes larger than 25 mm. Conversely, performance diagrams are designed to display the skill of forecast events that occur only relatively infrequently as they do not incorporate correct forecasts of null events (Roebber 2009). To account for this discrepancy, Fig. 3b displays HAILCAST's skill in forecasting objects with a maximum size of 25 mm or larger using a plot comparing the probability of detection versus the percent correct rejection (PCR), as opposed to the success ratio as in Fig. 3a. PCR measures the number of observed "no" objects (objects with a maximum hail size less than 25 mm) that were correctly forecast [Eq. (7) in Roebber (2009)]. Contours of the Peirce skill score (PSS; also known as the Hanssen and Kuipers skill score), which is a better discriminant of skill than the critical success index for more frequent events in a population (Peirce 1884; Woodcock 1976; Manzato 2007), are provided for reference. A perfect forecast in Fig. 3b is still in the top right-hand corner, just as in Fig. 3a. As a symbol shifts to lower values of POD in Fig. 3b, fewer objects with observed >25-mm hail size were correctly forecast; as it shifts to lower PCR, fewer objects with observed <25-mm hail size were correctly forecast.

While Figs. 3 and 4 generally indicate the same results, the combination of the two can more fully explain the

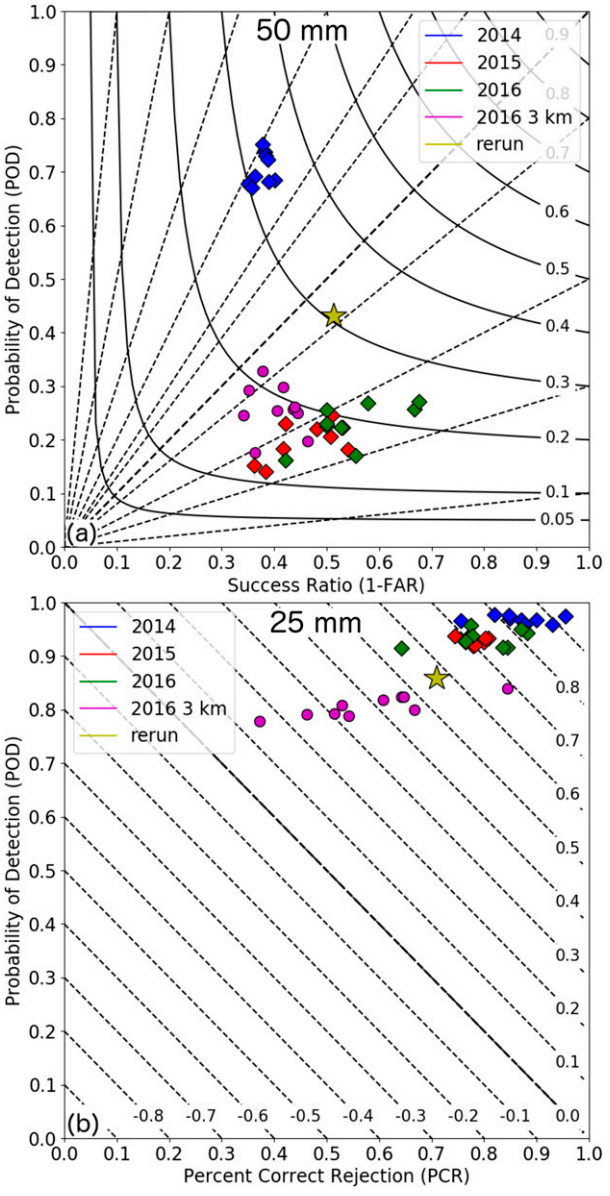

FIG. 3. Object-based verification for the 2014 (blue), 2015 (red), and 2016 (green) NSSL-WRF ensemble members; the 2016 CLUE single-physics members (purple); and vRerun (gold). Symbols for each year, ensemble, and rerun are as labeled. (a) Performance diagram for prediction of objects with a maximum hail size of 50 mm or larger. Solid lines of CSI are labeled; dashed bias lines are 1 along the diagonal, and correspond to overforecast values of 1.3, 1.5, 2, 3, 5, and 10 above the diagonal, and underforecast values of 0.8, 0.5, 0.3, and 0.1 below the diagonal. (b) Plot of probability of detection of objects with a maximum hail size of 25 mm or larger vs the percent of correct rejections of objects with maximum hail size less than 25 mm. Dashed lines are lines of constant PSS as per Peirce (1884). Perfect forecasts in both (a) and (b) would be in the top-right corner.

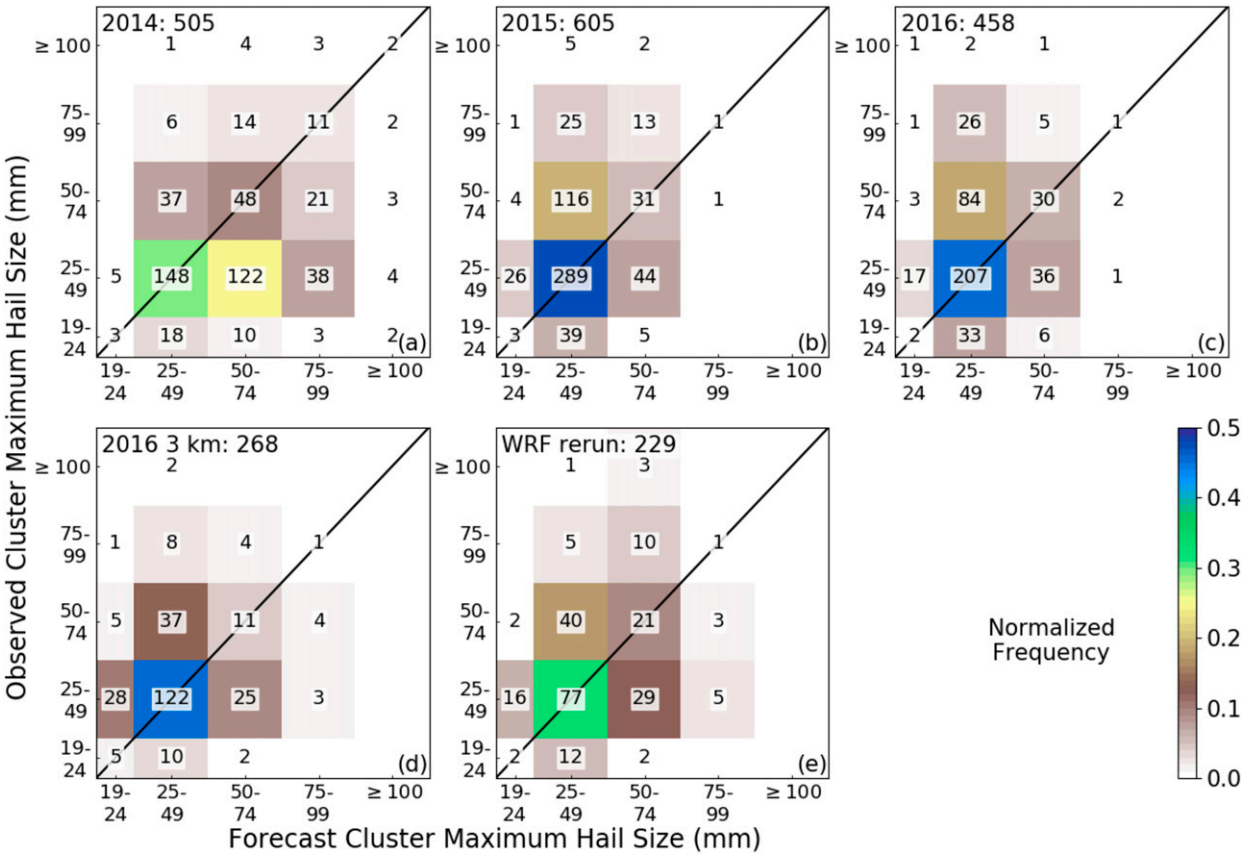

FIG. 4. Two-dimensional histogram plot showing the frequency of maximum observed and forecasted hail size values for each matched cluster in the object-based verification: (a) 2014, (b) 2015, and (c) 2016 NSSL-WRF control members, (d) 2016 CLUE control member, and (e) vRerun. Number within each bin is the number of matched object clusters; color shading is normalized by the total number of matched clusters each year, displayed next to the year label in each panel.

different models' levels of performance. The results from 2014 (Figs. 3 and 4a) indicate a large overforecasting bias, particularly for hail sizes larger than 50 mm. While ensemble members were able to make somewhat skillful forecasts of 25-mm hail swath objects (Figs. 3b and 4a), members frequently predicted many more >50-mm hail swaths than actually occurred (Fig. 3a). The bin counts are also clustered below the 1-to-1 correlation line in Fig. 4a. Subjective feedback from the SFE agreed with this assessment. Per the 2014 SFE Preliminary Findings and Results report (available online at http://hwt.nssl.noaa.gov/Spring_2014/HWT_SFE_2014_Prelim_Findings.pdf), "it became very apparent that HAILCAST substantially over-predicted hail sizes."

The only change made to HAILCAST between v2014 and v2015 was the embryo size determination method (Table 1). Using percentile points from the graupel size distribution within the microphysical parameterization greatly increased the dependency of the HAILCAST hail size forecast on which microphysics scheme was used. This variability was particularly evident with the Thompson microphysics scheme, which produced embryo sizes over twice the size of other schemes such as Morrison or WRF double-moment 6-class [WDM6; see Adams-Selin and Ziegler (2016, Fig. 2)]. Even with the WSM6 microphysics scheme, however, the NSSL-WRF ensemble still saw a high forecasting bias.

To correct this problem before the 2015 SFE, embryo sizes in v2015 HAILCAST were set to constant values within a range of 10–50 μm. These sizes are significantly smaller than are observed in nature (Magono and Nakamura 1965; Heymsfield 1982; Heymsfield and Musil 1982; Ziegler et al. 1983; Nelson and Knight 1987), but larger embryo sizes seemed to result in continued overforecasting. Figure 4b and the red symbols in Fig. 3, results from v2015, show an underforecasting of 50-mm hail (Fig. 3a, as well as a slight decrease in the ability to forecast <25-mm hail (Fig. 3b). When asked at the 2015 SFE if hail size forecasts provided "additional useful information relative to traditionally used hourly maximum

fields (HMFs) like UH [updraft helicity]'', respondents unanimously answered ''yes'' (results available at http://hwt.nssl.noaa.gov/Spring_2015/HWT_SFE_2015_Prelim_Findings_Final.pdf). The SFE questionnaire did not ask participants to elaborate on their reasons. Such answers could indicate an impression that HAILCAST is more skillful, or simply a preference to viewing forecasts of hail size in units of hail size, as opposed to requiring thresholds.

To address the large hail underforecasting problem in v2015, it was desired to increase the embryo sizes in v2016. However, if they were increased to typical observed sizes (1–10 mm), overforecasting became an issue. Review of one-dimensional hailstone trajectories within the v2015 model revealed that hailstones were frequently being advected above the supercooled water layer into the top of the storm by strong updrafts, reducing their ability to grow (see Adams-Selin and Ziegler 2016, Fig. 8). Because of the time-invariant nature of the updraft in v2015 (Table 1), growing hailstones were basically ''stuck'' in the center of the updraft until either they grew big enough to fall out or the updraft shut off. As already mentioned, observational and modeling studies indicate hailstone embryos form around the edge of the updraft before being advected horizontally across the core [see Heymsfield and Musil (1982, Fig. 13)]. HAILCAST v2016 incorporated a time-dependent multiplier to the updraft speed to simulate the hailstone's horizontal motion relative to the updraft [see Eq. (1) in Adams-Selin and Ziegler (2016)]. With this new multiplier, embryo sizes and insertion temperatures closer to observed values could be used (Table 1).

HAILCAST v2016 results are shown by green (NSSL-WRF) and purple (CLUE) symbols in Fig. 3 and Figs. 4c and 4d. The modifications made between v2015 and v2016, while improving the underlying physics of the algorithm, slightly improved the verification skill shown for 50-mm hail in Fig. 3a, led to some decrease in forecasting ability for 25-mm hail but still proved skillful (Fig. 3b), and resulted in minimal differences in Figs. 4c and 4d. Figure 3a still indicates an underforecasting of 50-mm hail, also similar to v2015. Subjective feedback from the 2016 SFE was more positive, with HAILCAST being rated the most highly among the three hail forecasting methods being evaluated [the Gagne machine-learning method (Gagne et al. 2017) and the Thompson method, which directly uses graupel size information from the microphysical parameterizations]. SFE feedback indicated larger areal coverage of 25-mm hail compared to the other two methods (see p. 28 of http://hwt.nssl.noaa.gov/Spring_2016/HWT_SFE_2016_preliminary_findings_final.pdf).

Interestingly, the NSSL-WRF 4-km ensemble (green) and CLUE 3-km ensemble (purple) did not show drastic differences in their verification results at 50-mm thresholds (Fig. 3a). In Fig. 3b, the CLUE 3-km ensemble members do show a reduction in skill in forecasting 25-mm hail, which is largely a result of slight underforecasting of objects with 25-mm hail (not shown). Figures 4c and 4d show only a small improvement between the distribution of maximum hail sizes from matched clusters between the NSSL-WRF (Fig. 4c) and the CLUE (Fig. 4d) results. The consistency of results across different model horizontal grid spacings indicates that the inclusion of parameterized horizontal motion reduced HAILCAST's sensitivity to the absolute values of the updraft strength. It also indicates HAILCAST can perform consistently across different microphysical parameterizations. Thus, as of v2016, HAILCAST can be run within models for a range of horizontal grid spacings or microphysical parameterization schemes without needing to be tuned or its forecasting thresholds changed.

The combination of verification results and SFE feedback indicated that v2016 of HAILCAST was still underforecasting large hail (50 mm and larger), but potentially overforecasting smaller (25 mm) hail. To address these biases, the moisture profile was examined. As already mentioned in section 2, hail-producing convection typically includes a BWER that contains an adiabatic cloud water profile due to a reduced number of precipitation scavengers (Heymsfield and Musil 1982). Thus, HAILCAST vRerun was modified to use an adiabatic cloud water profile based on the model cloud-base water vapor mixing ratio and the vertical profile of temperature and pressure. This was employed instead of directly using the model cloud water profile, which would have been artificially scavenged by precipitation due to CAMs not fully resolving the BWER region [see Adams-Selin and Ziegler (2016), section 3a]. With larger amounts of liquid cloud water available, vRerun was able to produce larger hailstones, reducing the bias and significantly improving the skill in forecasting 50-mm hail (yellow star in Fig. 3a), while improving the skill when forecasting 25-mm hail objects (Fig. 3b) compared to the CLUE members and one 2016 NSSL-WRF member. In Fig. 4e, the number of forecast and observed matched cluster maximum hail sizes both falling between 50 and 75 mm is increased compared to previous years, indicating an increased ability of vRerun to correctly forecast this larger hail.

Validation of storm surrogate fields from the control member of each ensemble for 50-mm hail swath objects over the same four time periods is presented in Fig. 5. The strong dependence of skill upon the surrogate threshold chosen is immediately apparent. The largest symbol in each panel in Fig. 5 corresponds to the ''best''

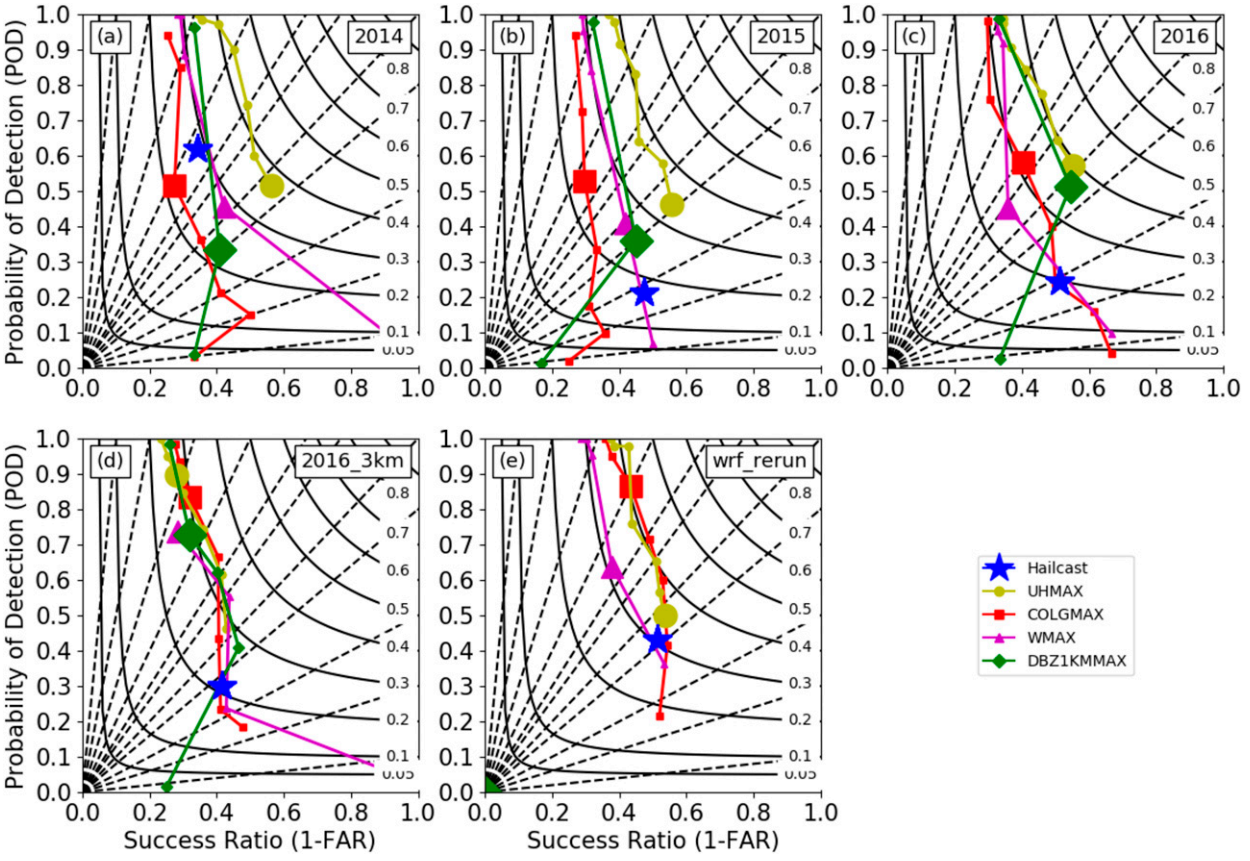

FIG. 5. Performance diagrams for object-based verification for 50-mm hail for HAILCAST and storm surrogate fields using a range of thresholds for (a) 2014, (b) 2015, and (c) 2016 NSSL-WRF control members, (d) 2016 CLUE control member, and (e) vRerun. Storm surrogate thresholds for the 4-km runs are the values given in Table 3. Thresholds in (d) are the same as the 4-km runs for COLGMAX and DBZ1KMMAX, and range from 60 to 160 $m^2 s^{-2}$ for UHMAX and from 20 to 50 $m s^{-1}$ for WMAX. The largest symbols in all panels correspond to the best thresholds. CSI and bias lines are as in Fig. 3a.

threshold as determined by the Brier skill score (see Table 3). Generally, WMAX (purple, triangles) provides a neutral to slight overforecast of 50-mm hail objects, while COLGMAX (red, squares) more strongly overforecasts. The tendency of DBZ1KMMAX (green, diamonds) is toward underforecasting. The best-performing surrogate is UHMAX (yellow, circles). HAILCAST is only able to show roughly equivalent skill to the updraft helicity as a predictor of 50-mm hail with vRerun, after all modifications have been incorporated.

For the object-based validation of the 3-km CLUE, the range of storm surrogate thresholds used for the calculation of skill was shifted to account for more fully resolved updrafts and to ensure the range covered a wide range of skill levels. COLGMAX was evaluated from 20 to 50 $kg m^{-2}$, DBZ1KMMAX from 36 to 60 dBZ, UHMAX from 60 to 160 $m^2 s^{-2}$, and WMAX from 20 to 50 $m s^{-1}$. Because the 3-km CLUE was not verified using the grid-based verification method owing to its different resolution, a set of best thresholds as determined by BSS calculations was not available for the 3-km configuration. Instead in Fig. 5d, the largest symbols correspond to the threshold producing the highest critical success index (CSI) score: 35 $kg m^{-2}$ for COLGMAX, 60 dBZ for DBZ1KMMAX, 140 $m^2 s^{-2}$ for UHMAX, and 35 $m s^{-1}$ for WMAX. Despite the threshold changes, however, all of the fields tend to overforecast hail size. Both UHMAX and DBZ1KMMAX show the largest change, shifting from almost no bias to a bias of 1.5. The significant change in skill between the 3- and 4-km 2016 ensembles highlights the importance of reevaluating the selected threshold used to determine hail size after a model horizontal grid-spacing, or potentially even configuration, change.

6. Results from grid-based verification

The grid-based verification methods described above in section 4b show generally similar results to the object-based methods, which is encouraging. Performance diagrams for 25- and 50-mm hail forecasts are provided in

TABLE 3. BSS values for grid-based verification of 25- and 50-mm hail probability forecasts from HAILCAST, storm surrogates for a range of thresholds, and SPC day 1 convective outlooks. Skill scores in boldface are the "best" thresholds over the entire period and were used to construct the attributes diagrams shown in Figs. 9 and 11. Skill scores in italics are the a priori thresholds and were used to construct Figs. 8 and 10. BSS is recorded as "null" if fewer than 1000 grid points contained nonzero forecast probabilities.

	25 mm				50 mm			
Field	v2014	v2015	v2016	vRerun	v2014	v2015	v2016	vRerun
HAILCAST	−0.211	−0.060	−0.017	−0.124	−0.457	0.010	0.005	−0.218
SPC	0.023	0.025	0.029	0.046				
COLGMAX ($kg\,m^{-2}$)								
20	−0.049	−0.164	−0.215	−0.028	−0.984	−1.481	−1.285	−0.726
25	*0.007*	*−0.021*	*−0.060*	*0.009*	−0.396	−0.735	−0.537	−0.398
30	**−0.011**	**−0.001**	**−0.035**	**0.015**	−0.129	−0.246	0.179	−0.186
35	−0.038	−0.021	−0.042	−0.000	**−0.035**	**−0.051**	**−0.051**	**−0.066**
40	Null	Null	−0.051	−0.012	Null	Null	−0.017	−0.018
50	Null	Null	Null	Null	*Null*	*Null*	*Null*	*Null*
DBZ1KMMAX (dBZ)								
36	−3.213	−4.411	−4.435	Null	−5.285	−6.664	−6.786	Null
48	−1.327	−1.811	−2.042	−0.237	−3.175	−3.733	−3.942	−0.747
55	**−0.051**	**−0.121**	**−0.177**	**Null**	**−0.814**	**−1.229**	**−1.079**	**Null**
60	*Null*	*Null*	*Null*	*Null*	*Null*	*Null*	*Null*	*Null*
UHMAX ($m^2\,s^{-2}$)								
30	0.032	0.026	−0.035	−0.010	−0.693	−0.872	−0.589	−0.411
40	*0.039*	*0.044*	*−0.011*	*−0.002*	−0.402	−0.448	−0.298	−0.220
50	**0.030**	**0.037**	**−0.008**	**−0.007**	−0.223	−0.213	−0.146	−0.132
60	0.018	0.022	−0.011	−0.014	−0.117	−0.085	−0.059	−0.071
70	0.003	0.008	−0.017	−0.018	−0.057	−0.019	−0.022	−0.036
80	−0.010	−0.004	−0.024	−0.023	*−0.026*	*0.011*	*−0.004*	*−0.019*
90	−0.019	−0.015	−0.029	−0.028	**−0.009**	**0.027**	**0.005**	**−0.009**
WMAX ($m\,s^{-1}$)								
12	−0.840	−1.246	−1.025	−0.336	−2.545	−3.131	−2.613	−1.681
16	−0.408	−0.686	−0.512	−0.149	−2.082	−2.557	−2.077	−1.228
20	*−0.105*	*−0.283*	*−0.184*	*−0.035*	−1.526	−2.022	−1.396	−0.768
22.5	0.004	−0.102	−0.089	0.002	−1.059	−1.561	−0.962	−0.474
25	**0.032**	**0.000**	**−0.036**	**0.013**	*−0.546*	*−0.892*	*−0.516*	*−0.244*
30	−0.030	−0.019	−0.035	−0.008	**−0.019**	**−0.031**	**−0.061**	**−0.028**

Fig. 6. The skill lines only show forecast probabilities from 5% to 60%, the probability thresholds the SPC uses to forecast hail. Specific probability thresholds most frequently used by the SPC (15%, 30%, and 45%) are additionally highlighted. These thresholds were chosen to evaluate HAILCAST's direct use to the forecaster.

Figure 6 shows that v2014 of HAILCAST overforecast both 25- and 50-mm hail, as was also noticed in the previous section. The interval of skill between the 15% and 45% probabilities (indicated by upward- and downward-facing triangles) is entirely in a region of positive bias for both 25- and 50-mm hail. The 30% probability of 25-mm hail has a bias of over 2 and that of 50-mm hail is approximately 3. As also noticed in the previous section, HAILCAST v2015 and v2016 improve the 25-mm hail forecast by reducing this overforecasting bias; the 15%–45% probability interval is now centered about a bias line of 1.5. However, 50-mm hail forecasts are reduced in skill, with the 30% forecast probability of v2015 or v2016 having a POD of either 0 or almost 0. (Grid-based validation of storm surrogate variables discussed later in the section will show a good portion of the low skill is due to the 2016 NSSL-WRF struggling with predicting convection, at least compared to the 2014 and 2015 results.) HAILCAST vRerun improved 50-mm hail forecasts; its 30% probability showed a bias of 1 and positive skill. Again, these results are similar to those noted in the previous section. However, the 25-mm grid-based results (Fig. 6a) show vRerun is overforecasting 25-mm hail, although not as badly as in v2014.

The different trends seen in the 25- and 50-mm hail forecasts seem to indicate that the hail size distribution produced by HAILCAST versions might not have a long enough right-hand tail or, in other words, may have trouble with extra-large hail size values. Similar results were presented in Adams-Selin and Ziegler (2016), although it should be noted that MESH also struggles with

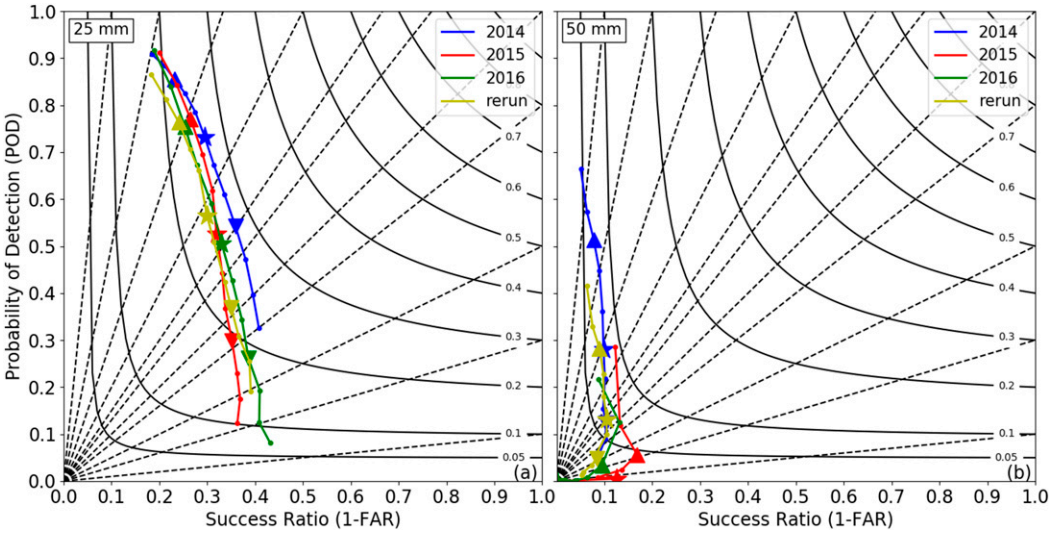

FIG. 6. Performance diagrams for grid-based verification for all 2014 (blue), 2015 (red), and 2016 (green) NSSL-WRF ensemble members, and the rerun (yellow) for hail thresholds of (a) 25 and (b) 50 mm. Curves extend from forecasted probabilities of 5%–60%. The upward triangle, star, and downward triangle are forecasted probabilities of 15%, 30%, and 45%, respectively.

larger hail sizes (Cintineo et al. 2012; Ortega 2018). The climatological rarity of extra-large hail sizes also makes drawing conclusions about forecasting methods difficult. However, to examine this possibility, Fig. 7 was constructed to examine the ratios between the number of forecast and observed grid points over a range of hail size intervals. The ratio was calculated by taking the mean number of forecasted grid points across the ensemble, and dividing by the number of observed grid points, of each hail size bin. Again, fairly large interval size bins are chosen in accordance with MESH's abilities. If the hail size distribution was perfectly predicted, the ratio would be equal to 1 across all size bins. Once again, v2014's overforecasting of all hail sizes is confirmed. HAILCAST v2015 and v2016 improved 19–25- and 25–50-mm hail forecasts at the expense of >50-mm hail forecasts. HAILCAST vRerun shows improvement for >50-mm hail, but at the expense of forecasting too much 25–50-mm hail. These results suggest vRerun, while an improvement over the previous three versions for 50-mm hail, still produces a hail size distribution that is too sharp. The amount of forecasted <50-mm hail needs to be reduced, while the amount of forecasted large hail (>50 mm) needs to be increased. Again, however, these results are difficult to confirm due to the low frequency and difficulty associated with observing very large hail sizes.

More interesting to a forecaster, however, is how HAILCAST compares to other forecasting tools at their disposal, such as the storm surrogate fields. Figures 8 and 9 show attributes diagrams for HAILCAST compared to the storm surrogate fields and the SPC day 1 hail outlooks for all three years and vRerun for the a priori and best thresholds. (Note that SPC forecasts are issued only at the discrete intervals of probability described in section 3c.) In both Figs. 8 and 9, it is evident that vRerun does overforecast 25-mm hail as already noted, although not as badly as v2014. HAILCAST v2015 and v2016 were particularly successful for lower forecast probabilities. Using the a priori thresholds, WMAX indicates nearly equivalent reliability to the HAILCAST forecasts in v2014, although HAILCAST was able to improve upon WMAX in v2015 and v2016. With the exception of some underforecasting of lower probabilities in 2014 and the rerun, COLGMAX tends to overforecast the occurrence of 25-mm hail. In addition, DBZ1KMMAX performed poorly as it appeared the model only rarely produced 60-dBZ reflectivities, while UHMAX performed the most reliably, producing the largest BSS values of all the storm surrogate fields using the a priori thresholds. HAILCAST v2016s BSS score was very close to that of UHMAX using the a priori threshold.

Comparison of results from the a priori thresholds (Fig. 8) to the best thresholds (Fig. 9) underscores the need for calibration of storm surrogate thresholds before using them to predict hail. In particular, DBZ1KMMAX shifts from showing no skill at any probability level, to showing positive skill at many lower forecast probabilities in 2014 and 2015. Likely the model configuration was able to produce a larger population of 55-dBZ storms. While still reliable, UHMAX is no longer the most reliable storm surrogate, with WMAX in 2014 and COLGMAX and WMAX in the rerun both having larger BSS values.

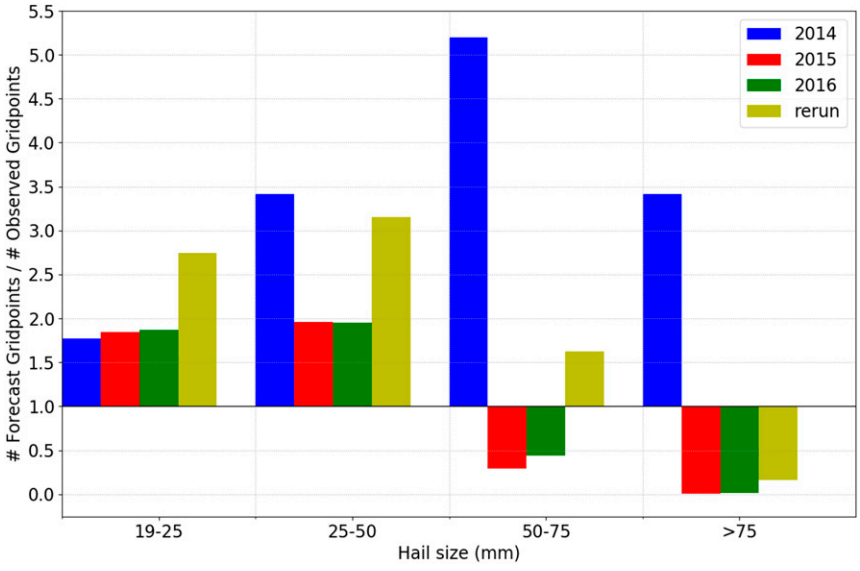

FIG. 7. Ratio of the number of forecast to observed grid points over a range of hail size bins. Shown are the 2014 (blue), 2015 (red), and 2016 (green) NSSL-WRF ensemble members, and the rerun (yellow). The number of forecast gridpoint values for the NSSL-WRF ensemble is calculated by taking the mean of all members. A perfect size forecast corresponds to the horizontal solid black line at $y = 1.0$.

The SPC forecasts are the only forecasts that show positive BSS values across all four years; neither HAILCAST nor a storm surrogate at any threshold was able to show similar performance. SPC forecasts show an underforecast of 25-mm hail, but this issue, also seen in Melick et al. (2014) and Herman et al. (2018), is likely due to the stairstep nature of SPC probabilities and to the SPC forecasters being evaluated on and self-calibrated to severe hail reports as opposed to MESH. If the probabilities were instead interpolated as a continuous field, the skill of the SPC outlooks would improve as in Herman et al. (2018).

The results for the grid-based verification for storm surrogates for the 50-mm forecasts are shown in attributes diagrams for a priori (Fig. 10) and best (Fig. 11) thresholds as well as the right columns in Table 3. Immediately evident from these figures is the fact that all forecasting methods struggle with forecasting 50-mm hail. Both COLGMAX and DBZ1KMMAX fail to predict any 50-mm hail using a priori thresholds, and their skill improves little when using the best threshold. Using the a priori threshold, WMAX overforecasts across all time periods; the results improve when using the best threshold but the simulation still performs poorly at higher forecast probabilities. The UHMAX results show perhaps the most consistent reliability across years and the two sets of thresholds, as is borne out by its higher BSS values (Table 3) as compared to the other storm surrogates. As in the 25-mm hail, v2016 of HAILCAST performs equivalently with UHMAX, showing equivalent BSS values for the best threshold and improves upon UHMAX when evaluated with the a priori threshold.

In contrast to the object-based verification, HAILCAST vRerun does not improve upon v2016 when forecasting 50-mm hail. One possibility for this discrepancy is that while vRerun improves upon the number of storm objects with 50-mm hail, it produces a too-large areal coverage of 50-mm hail.

7. Discussion and conclusions

The HAILCAST one-dimensional hail model, embedded within WRF, was tested for three years' worth of NOAA Hazardous Weather Testbed (HWT) Spring Forecasting Experiments (SFEs) from 2014 to 2016. WRF-HAILCAST was run as part of the NSSL-WRF 4-km ensemble with 9 or 10 members for all three years, as well as within the single-physics members of the CLUE 3-km ensemble during 2016. Verification results and subjective participant feedback were obtained from the SFEs each year, and resulting improvements to the hail model were incorporated yearly. Modifications made to HAILCAST after the 2016 SFE were tested via rerunning WRF over a 6-week period during 2014 and 2015, using a model domain and configuration similar to the NSSL-WRF.

Two different types of verification methods were used. The first, object-based verification, used MODE software to match the observed and forecast swaths of hail across space. This method allowed for evaluation of

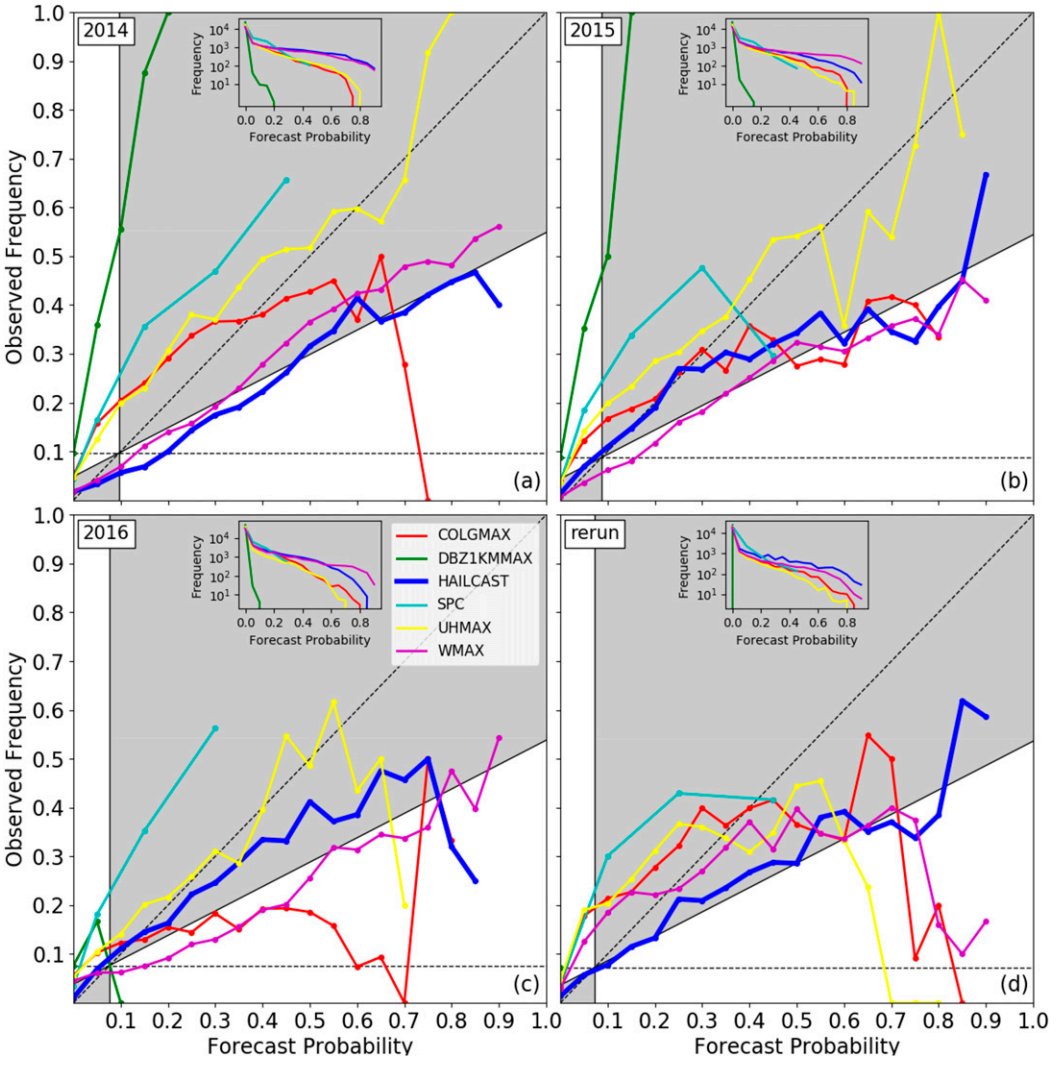

FIG. 8. Attributes diagram for grid-based verification for 25-mm hail for HAILCAST, storm surrogate fields, and SPC day 1 convective outlook forecasts from (a) 2014, (b) 2015, (c) 2016, and (d) rerun. The shaded gray area indicates skillful forecasts; the dashed diagonal line is a forecast of perfect reliability. The horizontal dashed line is a climatological forecast. The inset shows the frequency of forecasts in each probability bin. Storm surrogate thresholds for 25-mm hail were the "a priori" thresholds noted by the italicized rows in Table 3.

HAILCAST hail forecasting skill without penalizing it for WRF not successfully forecasting convection. The second method used neighborhood grid-based verification methods to examine WRF-HAILCAST's skill in hail forecasting as a whole.

Both methods, in addition to participant feedback, found that the 2014 version of HAILCAST significantly overforecasted all hail sizes. This result was due to a combination of using embryo sizes retrieved from percentile points along the graupel distribution within the microphysical parameterization, and requiring the hailstone be locked in the center of the updraft until it could grow big enough to fall out. The graupel distributions within different WRF microphysics schemes vary widely, so the assigned embryo sizes and eventual forecast hail size did as well. To address this inconsistency, before the 2015 HWT, HAILCAST's embryo sizes were set to constant, albeit small, values. In the 2015 version of HAILCAST, the object-based verification found a small underforecasting of large (≥50 mm) hail sizes. The grid-based verification agreed, but also noted a significant improvement in forecasting smaller, 25-mm hail, particularly compared to the 2014 version. Participant feedback from the 2015 SFE agreed.

In 2016, HAILCAST was modified to include larger embryo sizes and insertion points that better matched

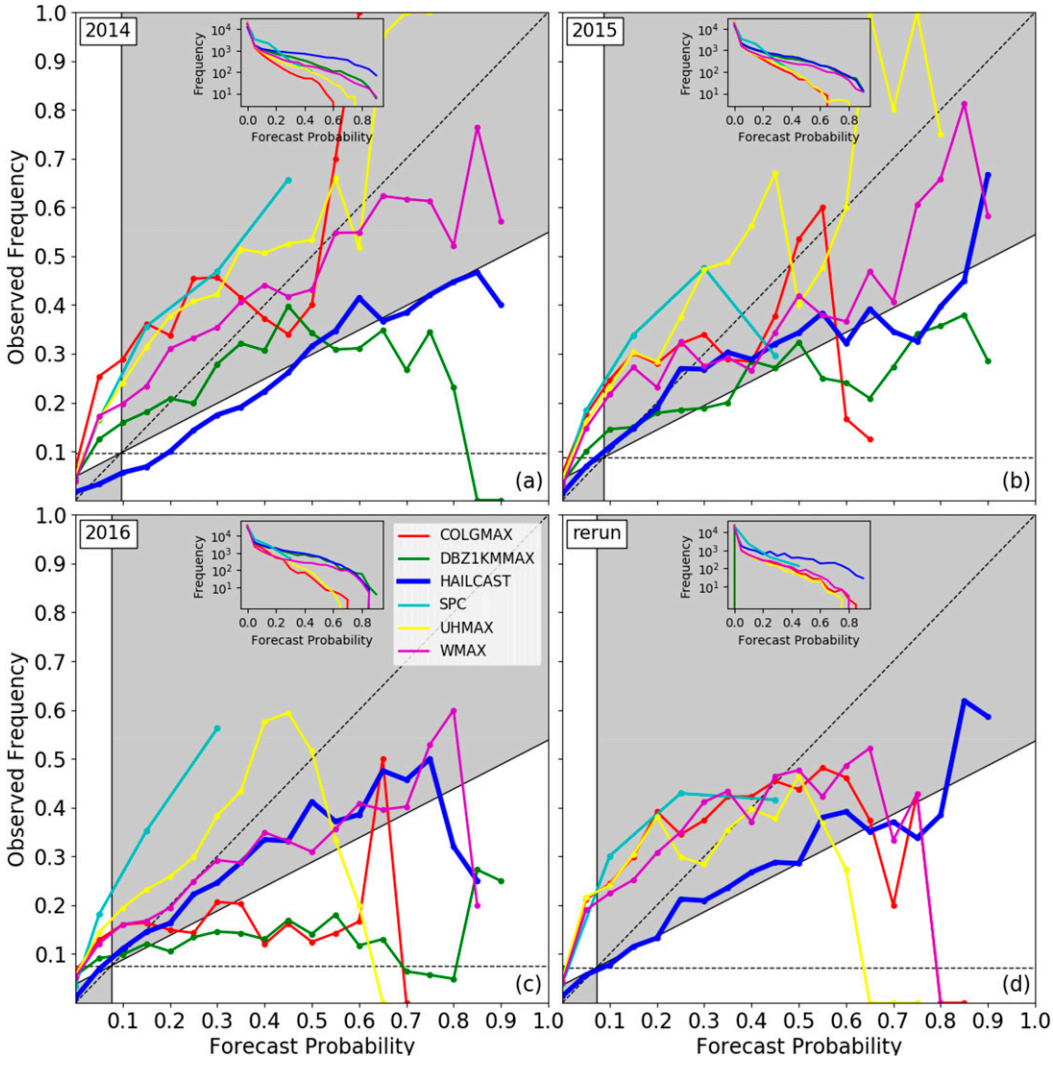

FIG. 9. As in Fig. 8, but storm surrogate thresholds for 25-mm hail were determined by the highest BSS value over the 3-yr period, noted by the boldface rows in Table 3.

with values observed in previous studies. It also included a time-dependent multiplier to the updraft speed that parameterized the embryo's horizontal motion across the updraft. The combination of these two improved 25- and 50-mm forecasts compared to the 2015 version as determined by both the object-based and grid-based verification methods. Subjective feedback from the 2016 SFE was positive as well. Verification results were largely consistent across the two 2016 ensembles of different horizontal grid spacings, indicating WRF-HAILCAST's lessened sensitivity to absolute updraft speed as well as microphysical parameterization. This is an important consideration and indicates that HAILCAST could be run across multiple model configurations and horizontal grid spacings without needing to retune its thresholds.

After the 2016 SFE, in order to address the continued underforecasting of hail sizes $\geq$ 50 mm, the use of an adiabatic cloud liquid water profile was introduced. Because HAILCAST can be implemented within a CAM running at 3- or 4-km horizontal grid spacing, the precipitation-free bounded weak-echo region where maximum hail growth typically occurs is not fully resolved, and the cloud liquid water was being artificially scavenged within the CAM. Thus, an adiabatic cloud liquid water profile was recreated using model temperature, pressure, and water vapor fields. With these modifications, WRF-HAILCAST was rerun. Its skill in forecasting 50-mm hail improved as determined by the object-based verification method, but the grid-based verification method showed a decrease in skill. The rerun version showed an improvement in forecasting <25-mm hail per the object-based

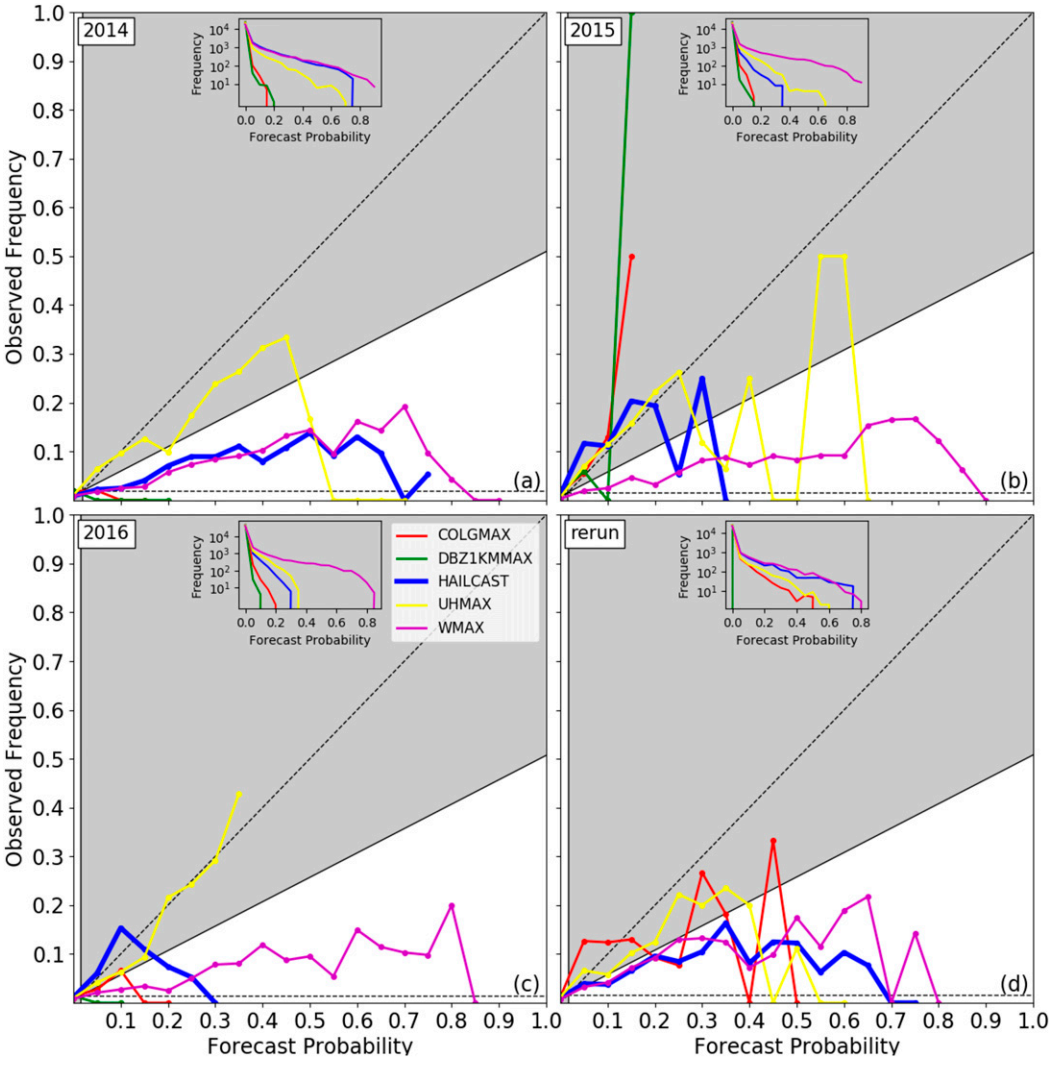

FIG. 10. As in Fig. 8, but for 50-mm hail.

verification method, but a decrease in skill of 25–50-mm hail per both verification methods.

Hourly maximum storm surrogate fields, including updraft helicity, column-integrated graupel, maximum column updraft speed, and 1-km radar reflectivity were also examined as predictors for hail over the same period, using both the object- and grid-based verification methods. Thresholds of the storm surrogate fields were selected using two methods: one designed to mimic the decision process of an operator designing a model configuration (a priori), and the second designed to show the best possible performance of the threshold fields (best). SPC hail convective outlooks were also evaluated using the grid-based methods.

Comparison of verification results from the two threshold selection methods emphasizes the need for calibration of the threshold for the specific model configuration, a consideration not required by HAILCAST. When using a priori thresholds, updraft helicity was the most skilled surrogate at detecting 25-mm hail, but when using the best thresholds, updraft speed or column-integrated graupel was the most skilled surrogate in some years. The performance of the radar reflectivity storm surrogate varied widely across the different thresholds. WRF-HAILCAST v2016 produced results roughly equivalent in skill to the best-performing surrogate across both types of thresholds; vRerun did overforecast the occurrence of 25-mm hail. SPC outlooks performed the best of any method. For 50-mm hail, updraft helicity was the best performer across both sets of thresholds, although all model-based forecasting methods performed poorly. However, WRF-HAILCAST vRerun performed comparatively to updraft helicity with the best threshold when evaluated using the objective verification method,

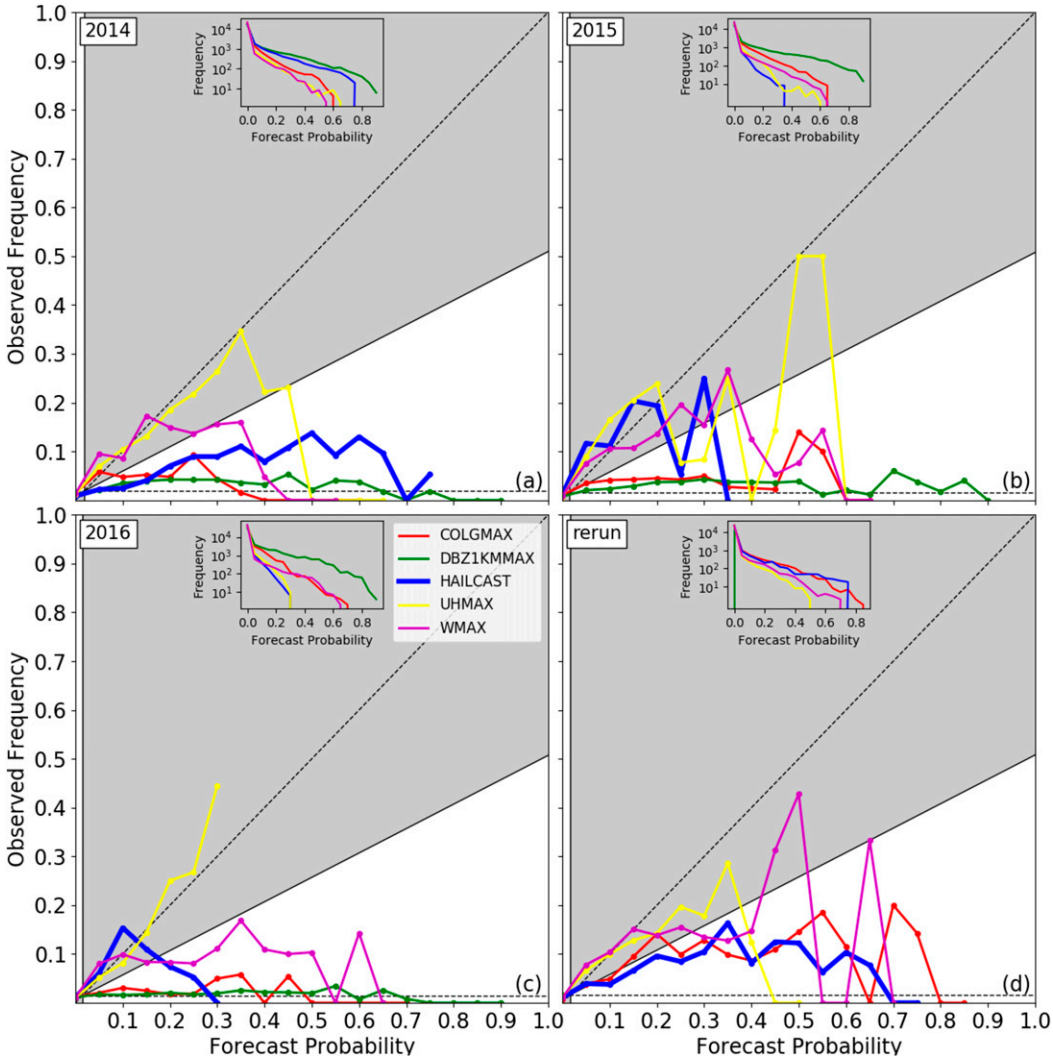

FIG. 11. As in Fig. 9, but for 50-mm hail.

and v2016 performed equivalently using the regridded verification method.

The different levels of performance of the storm surrogate fields in forecasting 25- versus 50-mm hail highlights potentially contrasting methods of development for the two hailstone sizes. It is generally hypothesized that supercell storm structure is required to produce the necessary strength, size, and volume of updraft to generate large, 50-mm hail (Foote and Frank 1983; Foote 1984; Miller et al. 1988; Dennis and Kumjian 2017). Since the supercell updraft structure would be best captured in the model by the updraft helicity parameter, it is unsurprising that updraft helicity is the best-performing storm surrogate for 50-mm hail. Conversely, 25-mm hail is more likely to be generated by multicellular, less-organized, and nonsupercellular convection. Storm surrogate fields of updraft speed or column-integrated graupel would better capture development in the model of these storm types. The ability of vRerun or v2016 of HAILCAST to produce forecasts of both 25- and 50-mm hail similar in skill to the best-performing storm surrogates indicates that it can skillfully forecast hail across a variety of convective regimes and is worthy of note.

Overall, both object- and grid-based verification methods indicated that HAILCAST improved steadily in response to feedback received from three years' worth of NOAA HWTs. HAILCAST v2016 was better or at least comparable in skill to the storm surrogate fields and the SPC forecasts when predicting both 50-mm hail forecasts; only updraft helicity was able to produce 50-mm hail forecasts comparable in skill to HAILCAST's. HAILCAST vRerun further improved its 50-mm hail forecasts, but at the expense of overforecasting 25-mm hail. Future research will work to improve the forecast

hail size distribution, focusing on the importance of embryo source regions and in-storm hail trajectories, as already noted by Dennis and Kumjian (2017).

Acknowledgments. This work was performed as part of the Cooperative Research Data Agreement between the 557th Weather Wing and Atmospheric and Environmental Research, Inc. (AER). Additional support for this effort was provided through AER internal research funds. Computing resources were provided by the Navy Department of Defense (DoD) Supercomputing Resource Center (Navy DSRC), which is sponsored by the DoD High Performance Computing Modernization Program. CJM was provided support by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce. Rob Hepper provided the SPC convective outlook data. David John Gagne provided access to the MESH data as well as a constructive review; the comments of two additional anonymous reviewers were also helpful.

REFERENCES

Adams-Selin, R., and C. Ziegler, 2016: Forecasting hail using a one-dimensional hail growth model within WRF. *Mon. Wea. Rev.*, **144**, 4919–4939, https://doi.org/10.1175/MWR-D-16-0027.1.

Ben Bouallègue, Z., and S. E. Theis, 2014: Spatial techniques applied to precipitation ensemble forecasts: From verification results to probabilistic products. *Meteor. Appl.*, **21**, 922–929, https://doi.org/10.1002/met.1435.

Brimelow, J. C., G. W. Reuter, and E. R. Poolman, 2002: Modeling maximum hail size in Alberta thunderstorms. *Wea. Forecasting*, **17**, 1048–1062, https://doi.org/10.1175/1520-0434(2002)017<1048:MMHSIA>2.0.CO;2.

Cintineo, J. L., T. M. Smith, and V. Lakshmanan, 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.

Clark, A. J., and Coauthors, 2018: The Community Leveraged Unified Ensemble (CLUE) in the 2016 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Bull. Amer. Meteor. Soc.*, **99**, 1433–1448, https://doi.org/10.1175/BAMS-D-16-0309.1.

Davis, C., B. Brown, and R. Bullock, 2006a: Object-based verification of precipitation forecasts. Part I: Methods and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772–1784, https://doi.org/10.1175/MWR3145.1.

——, ——, and ——, 2006b: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. *Mon. Wea. Rev.*, **134**, 1785–1795, https://doi.org/10.1175/MWR3146.1.

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.

Du, J., and Coauthors, 2014: NCEP regional ensemble update: Current systems and planned storm-scale ensembles. *26th Conf. on Weather Analysis and Forecasting/22nd Conf. on Numerical Weather Prediction*, Atlanta, GA, Amer. Meteor. Soc., J1.4, https://ams.confex.com/ams/94Annual/webprogram/Paper239030.html.

Foote, G. B., 1984: A study of hail growth utilizing observed storm conditions. *J. Climate Appl. Meteor.*, **23**, 84–101, https://doi.org/10.1175/1520-0450(1984)023<0084:ASOHGU>2.0.CO;2.

——, and H. W. Frank, 1983: Case study of a hailstorm in Colorado. Part III: Airflow from triple-Doppler measurements. *J. Atmos. Sci.*, **40**, 686–707, https://doi.org/10.1175/1520-0469(1983)040<0686:CSOAHI>2.0.CO;2.

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

Gallo, B. T., and Coauthors, 2017: Breaking new ground in severe weather prediction: The 2015 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Wea. Forecasting*, **32**, 1541–1568, https://doi.org/10.1175/WAF-D-16-0178.1.

Herman, G. R., E. R. Nielsen, and R. S. Schumacher, 2018: Probabilistic verification of Storm Prediction Center convective outlooks. *Wea. Forecasting*, **33**, 161–184, https://doi.org/10.1175/WAF-D-17-0104.1.

Heymsfield, A. J., 1982: A comparative study of the rates of development of potential graupel and hail embryos in high plains storms. *J. Atmos. Sci.*, **39**, 2867–2897, https://doi.org/10.1175/1520-0469(1982)039<2867:ACSOTR>2.0.CO;2.

——, 1983a: A technique for investigating graupel and hail development. *J. Climate Appl. Meteor.*, **22**, 1143–1160, https://doi.org/10.1175/1520-0450(1983)022<1143:ATFIGA>2.0.CO;2.

——, 1983b: Case study of a hailstorm in Colorado. Part IV: Graupel and hail growth mechanisms deduced through particle trajectory calculations. *J. Atmos. Sci.*, **40**, 1482–1509, https://doi.org/10.1175/1520-0469(1983)040<1482:CSOAHI>2.0.CO;2.

——, and D. J. Musil, 1982: Case study of a hailstorm in Colorado. Part II: Particle growth processes at mid-levels deduced form *in-situ* measurements. *J. Atmos. Sci.*, **39**, 2847–2866, https://doi.org/10.1175/1520-0469(1982)039<2847:CSOAHI>2.0.CO;2.

——, A. R. Jameson, and H. W. Frank, 1980: Hail growth mechanisms in a Colorado storm. Part II: Hail formation processes. *J. Atmos. Sci.*, **37**, 1779–1813, https://doi.org/10.1175/1520-0469(1980)037<1779:HGMIAC>2.0.CO;2.

Hitchens, N. M., H. E. Brooks, and M. P. Kay, 2013: Objective limits on forecasting skill of rare events. *Wea. Forecasting*, **28**, 525–534, https://doi.org/10.1175/WAF-D-12-00113.1.

Hong, S.-Y., and J.-O. J. Lim, 2006: The WRF single-moment 6-class microphysics scheme (WSM6). *J. Korean Meteor. Soc.*, **42**, 129–151.

Jewell, R., and J. Brimelow, 2009: Evaluation of Alberta hail growth model using severe hail proximity soundings from the United States. *Wea. Forecasting*, **24**, 1592–1609, https://doi.org/10.1175/2009WAF2222230.1.

Jirak, I., and Coauthors, 2014: An overview of the 2014 NOAA Hazardous Weather Testbed Spring Forecasting Experiment. *27th Conf. on Severe Local Storms*, Madison, WI, Amer. Meteor. Soc., 46, https://ams.confex.com/ams/27SLS/webprogram/Paper254650.html.

Kain, J. S., S. R. Dembek, S. J. Weiss, J. L. Case, J. J. Levit, and R. A. Sobash, 2010: Extracting unique information from high-resolution forecast models: Monitoring selected fields and phenomena every time step. *Wea. Forecasting*, **25**, 1536–1542, https://doi.org/10.1175/2010WAF2222430.1.

Lakshmanan, V., T. Smith, K. Hondl, G. J. Stumpf, and A. Witt, 2006: A real-time, three-dimensional, rapidly updating, heterogeneous radar merger technique for reflectivity, velocity, and derived products. *Wea. Forecasting*, **21**, 802–823, https://doi.org/10.1175/WAF942.1.

Magono, C., and T. Nakamura, 1965: Aerodynamic studies of falling snowflakes. *J. Meteor. Soc. Japan*, **43**, 139–147, https://doi.org/10.2151/jmsj1965.43.3_139.

Manzato, A., 2007: A note on the maximum Peirce skill score. *Wea. Forecasting*, **22**, 1148–1154, https://doi.org/10.1175/WAF1041.1.

Melick, C. J., I. L. Jirak, and J. Correia Jr., A. R. Dean, and S. J. Weiss, 2014: Exploration of the NSSL maximum expected size of hail (MESH) product for verifying experimental hail forecasts in the 2014 Spring Forecast Experiment. *27th Conf. on Severe Local Storms*, Madison, WI, Amer. Meteor. Soc., 76, https://ams.confex.com/ams/27SLS/webprogram/Paper254292.html.

Miller, L. J., J. D. Tuttle, and C. A. Knight, 1988: Airflow and hail growth in a severe northern high plains supercell. *J. Atmos. Sci.*, **45**, 736–762, https://doi.org/10.1175/1520-0469(1988)045<0736:AAHGIA>2.0.CO;2.

——, ——, and G. B. Foote, 1990: Precipitation production in a large Montana hailstorm: Airflow and particle growth trajectories. *J. Atmos. Sci.*, **47**, 1619–1646, https://doi.org/10.1175/1520-0469(1990)047<1619:PPIALM>2.0.CO;2.

Murphy, A., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600, https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

NCEI, 2017: U.S. billion-dollar weather and climate disasters: Overview. National Centers for Environmental Information, https://www.ncdc.noaa.gov/billions/.

Nelson, S. P., 1983: The influence of storm flow structure on hail growth. *J. Atmos. Sci.*, **40**, 1965–1983, https://doi.org/10.1175/1520-0469(1983)040<1965:TIOSFS>2.0.CO;2.

——, and N. C. Knight, 1987: The hybrid multicellular–supercellular storm—An efficient hail producer. Part I: An archetypal example. *J. Atmos. Sci.*, **44**, 2042–2059, doi.org/10.1175/1520-0469(1987)044%3c2042:THMSEH>2.0.CO;2.

Ortega, K., 2018: Evaluating multi-radar, multi-sensor products for surface hailfall diagnosis. *Electron. J. Severe Storms Meteor.*, **13** (1), http://www.ejssm.org/ojs/index.php/ejssm/article/viewArticle/163.

Peirce, C. S., 1884: The numerical measure of the success of predictions. *Science*, **4**, 453–454, https://doi.org/10.1126/science.ns-4.93.453-a.

Poolman, E. R., 1992: Die voorspelling van haelkorrelgroei in Suid-Afrika (The forecasting of hail growth in South Africa). M.S. thesis, Faculty of Engineering, University of Pretoria, 113 pp.

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

Schwartz, C. S., and R. A. Sobash, 2017: Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: A review and recommendations. *Mon. Wea. Rev.*, **145**, 3397–3418, https://doi.org/10.1175/MWR-D-16-0400.1.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., http://dx.doi.org/10.5065/D68S4MVH.

Smith, T. M., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) severe weather and aviation products: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 1617–1630, https://doi.org/10.1175/BAMS-D-14-00173.1.

Sobash, R. A., J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011: Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme phenomena in convection-allowing model forecasts. *Wea. Forecasting*, **26**, 714–728, https://doi.org/10.1175/WAF-D-10-05046.1.

——, C. S. Schwartz, G. S. Romine, K. R. Fossell, and M. L. Weisman, 2016: Severe weather prediction using storm surrogates from an ensemble forecasting system. *Wea. Forecasting*, **31**, 255–271, https://doi.org/10.1175/WAF-D-15-0138.1.

Wendt, N., I. Jirak, and C. Melick, 2016: Verification of severe weather proxies from the NSSL-WRF for hail forecast. *28th Conf. on Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., 110, https://ams.confex.com/ams/28SLS/webprogram/Paper300913.html.

Wilks, D., 2006: Forecast verification. *Statistical Methods in the Atmospheric Sciences*, D. S. Wilks, Ed., 2nd ed., Academic Press, 260–268.

Wilson, C. J., K. L. Ortega, and V. Lakshmanan, 2009: Evaluating multi-radar, multi-sensor hail diagnosis with high resolution hail reports. *25th Conf. on Interactive Information Processing Systems*, Phoenix, AZ, Amer. Meteor. Soc., P2.9, https://ams.confex.com/ams/89annual/techprogram/paper_146206.htm.

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

Woodcock, F., 1976: The evaluation of yes/no forecasts for scientific and administrative purposes. *Mon. Wea. Rev.*, **104**, 1209–1214, https://doi.org/10.1175/1520-0493(1976)104<1209:TEOYFF>2.0.CO;2.

Ziegler, C. L., P. S. Ray, and N. C. Knight, 1983: Hail growth in an Oklahoma multicell storm. *J. Atmos. Sci.*, **40**, 1768–1791, https://doi.org/10.1175/1520-0469(1983)040<1768:HGIAOM>2.0.CO;2.