Just what is "good"? Musings on hail forecast verification through evaluation of FV3-HAILCAST hail forecasts

Rebecca D. Adams-Selin,[a] Christina Kalb,[b] Tara Jensen,[b] John Henderson,[c] Tim Supinie,[d] Lucas Harris,[e] Yunheng Wang,[f,g] Burkely T. Gallo,[f,h] Adam J. Clark[i,j]

[a] *Verisk Atmospheric and Environmental Research, Bellevue, Nebraska*

[b] *Research Applications Laboratory, National Center for Atmospheric Research, Boulder, Colorado*

[c] *Verisk Atmospheric and Environmental Research, Lexington, Massachusetts*

[d] *Center for Analysis and Prediction of Storms, Norman, Oklahoma*

[e] *NOAA/OAR Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey*

[f] *Cooperative Institute for Severe and High-Impact Weather Research and Operations, University of Oklahoma, Norman, Oklahoma*

[g] *NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma*

[h] *NOAA/NWS/NCEP Storm Prediction Center, Norman, Oklahoma*

[i] *NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma*

[j] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*

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

ABSTRACT: Hail forecasts produced by the CAM-HAILCAST pseudo-Lagrangian hail size forecasting model were evaluated during the 2019, 2020, and 2021 NOAA Hazardous Weather Testbed Spring Forecasting Experiments. As part of this evaluation, HWT SFE participants were polled about their definition of a “good” hail forecast. Participants were presented with two different verification methods conducted over three different spatiotemporal scales, and were then asked to subjectively evaluate the hail forecast as well as the different verificaiton methods themselves. Results recommended use of multiple verification methods tailored to the type of forecast expected by the end-user interpreting and applying the forecast.

The hail forecasts evaluated during this period included an implementation of CAM-HAILCAST in the Limited Area Model of the Unified Forecast System with the Finite Volume 3 (FV3) dynamical core. Evaluation of FV3-HAILCAST over both 1-h and 24-h periods found continued improvement from 2019 to 2021. The improvement was largely a result of wide intervariability among FV3 ensemble members with different microphysics parameterizations in 2019 lessening significantly during 2020 and 2021. Overprediction throughout the diurnal cycle also lessened by 2021. A combination of both upscaling neighborhood verification and an object-based technique that only retained matched convective objects was necessary to understand the improvement., agreeing with the HWT SFE participants’ recommendations for multiple verification methods.

SIGNIFICANCE STATEMENT: "Good" forecasts of hail can be determined in multiple ways and must depend on both the performance of the guidance and the perspective of the end-user. This work looks at different verification strategies to capture the performance of the CAM-HAILCAST hail forecasting model across three years of the Spring Forecasting Experiment (SFE) in different parent models. Verification strategies were informed by SFE participant input via a survey. Skill variability among models decreased in SFE 2021 relative to prior SFEs. The FV3 model in 2021, compared to 2019, provided improved forecasts of both convective distribution and 38-mm (1.5 in) hail size, as well as less overforecasting of convection from 1900–2300 UTC.

1. Introduction

Hail is the most consistently damaging hazard of severe thunderstorms, producing losses in the U.S. alone exceeding $10 billion per year over the past 13 years (Faust et al. 2021). With improved detection and prediction of severe hail along with understanding of hail characteristics and their impacts at the surface a good portion of this monetary loss could be avoided. Yet, much like the nature of weather forecasts in general (Murphy 1993), determination of what makes a hail forecast "good" is a surprisingly difficult concept. Public, private, and even academic interests in hail prediction vary, with location, timing, and size of the forecast hail all at various levels of importance depending on the forecast's end user. As such, identification of the most-desired"good" forecast characteristics from a cross-section of the severe hazard community is necessary.

The existence of multiple standards for a "good" forecast likely drives the proliferation of convective hazard verification methods in the literature. Convective hazards are highly spatially and temporally variable, making validation without undue penalization of missed forecasts difficult. Several verification configurations have been used that reward a convective hazard forecast if it successfully predicts occurrence of a hazard within some spatial and/or temporal interval surrounding the occurrence itself. Upscaling neighborhood approaches are one such option where forecast hazard occurrence is upscaled to a coarser grid (e.g., Marsh et al. 2012; Hitchens et al. 2013; Schwartz and Sobash 2017; Roberts et al. 2020; Gallo et al. 2021): a forecast is considered successful if the forecast and observed occurrences both occur within the same coarse grid box. Additional configurations of this option include smoothing the forecast to further account for spatial error.

Object-matching methods such as the Method for Object-based Diagnostic Evaluation (MODE hereafter, Davis et al. 2006a,b) or the technique developed by Skinner et al. (2018) for the NOAA Warn-on-Forecast System (WoFS; Wheatley et al. 2015) also allow for spatial errors in a convective hazard forecast by matching forecast and observed convective objects (e.g., hail swaths) and comparing their shape, size, separation distance, and magnitudes. These methods are designed to mimic subjective verification by forecasters. Object-based methods are also useful when both the forecasts and their verification need to remain on small spatial and temporal scales, such as for probabilistic convective forecasts produced in real-time by WoFS (e.g., Skinner et al. 2018; Potvin et al. 2020; Britt et al. 2020; Flora et al. 2021; Miller et al. 2021). Finally, both upscaling neighborhood and object-based verification methods, including the many variations therein, all still penalize a convective hazard forecast even if the underlying Numerical Weather Prediction (NWP) model failed to predict convection. Such an outcome is likely desired for forecasters interested in warning the population affected by the hazard. That outcome is not desired, however, by developers of the convective hazard forecasting method itself, who want to separate performance of the underlying NWP model from the performance of their hazard forecasting method. Such an outcome requires yet a different verification technique.

Given this variety of convective hazard verification methods, an evaluation of the verification methods themselves is needed, and must be informed by identified "good" forecast characteristics. In this study, the performance of the CAM-HAILCAST (Convection-Allowing Model-HAILCAST; Adams-Selin and Ziegler 2016; Adams-Selin et al. 2019) hail forecast model is used to explore both the idea of a "good" hail forecast and evaluate the effectiveness of several verification methods, including object-matching and upscaling neighborhood approaches. CAM-HAILCAST was deployed in the Limited Area Model (LAM; Black et al. 2021) versions of Finite-Volume Cubed-Sphere Dynamical Core (FV3; Putman and Lin 2007) model at the Center for Analysis and Prediction of Storms (CAPS) and the National Severe Storms Laboratory (NSSL) during the 2019, 2020, and 2021 NOAA Hazardous Weather Testbed (HWT) Spring Forecasting Experiments (SFEs; Clark et al. 2012a; Gallo et al. 2017a), and included in the High-Resolution Rapid Refresh - Ensemble (HRRR-E; Alexander et al. 2020) during the 2020 HWT SFE. The FV3 dynamical core is part of NOAA's effort to create a Unified Forecast System (UFS; `https://ufscommunity.org/`) across all modeled scales. The LAM FV3 will be the foundation of the new Rapid Refresh Forecasting System (RRFS), which is designed to subsume several of NOAA's current regional modeling systems including the HRRR. In addition, discussion of convective hazard forecasts from LAM FV3 configurations in the literature is growing (e.g., Snook et al. 2019; Zhang et al. 2019; Harris et al. 2019; Zhou et al. 2019; Gallo et al. 2021), but further study is needed.

It is our hypothesis that verification preferences will change based upon an individual's understanding of a hail forecast's purpose, which we expect will show significant variation. Section 2 details the implementation of FV3-HAILCAST, the configuration of the FV3 and HRRRE versions at each SFE, and describes the different verification methods, time, and space scales used. Section 3 discusses the SFE survey results about necessary elements of "good" hail forecasts and verification method effectiveness, and provides a case study verification method comparison. Section 4 uses these different methods to evaluate CAM-HAILCAST performance over 24-h periods across the three years. Section 5 examines the usefulness of temporally and spatially dependent verification, with a focus on forecasts over both 1-h and 24-h periods. Discussion and conclusions are presented in Section 6.

2. Methodology

a. FV3-HAILCAST

The HAILCAST of Adams-Selin and Ziegler (2016) and Adams-Selin et al. (2019), termed CAM-HAILCAST, is a one-dimensional psuedo-Lagrangian hail trajectory model designed to be embedded within any CAM. It is one-dimensional as it operates independently on each convective grid column in the CAM; each grid column serves as an input updraft profile for the hail trajectroy model. The "pseudo-Lagrangian" nature of CAM-HAILCAST is achieved by employing an updraft parameterization to simulate the updraft as experienced by a hailstone being advected across it. Previous verification studies have found CAM-HAILCAST deployed within the Weather Research and Forecasting model (WRF) to be most successful in the U.S. Great Plains and Midwest (e.g., Fig. 10 of Gagne et al. 2017) and for smaller hail (e.g., 25-mm; Adams-Selin et al. 2019). The reduced skill of WRF-HAILCAST in forecasting 50-mm hail or larger is not unexpected given the importance of increased updraft volume and hailstone residence time aloft in the production of larger hail (Kumjian and Lombardo 2020; Kumjian et al. 2021; Lin and Kumjian 2022), and hence, it must be assumed, two- or three-dimensional hail trajectory motions. Yet despite its issues, the CAM-HAILCAST hail forecasting method remains one of the most skillful yet operationally efficient model-based hail forecasting methods (Gagne et al. 2017; Adams-Selin and Ziegler 2016; Adams-Selin et al. 2019). CAM-HAILCAST was incorporated into the LAM configuration of FV3, termed FV3-HAILCAST. Understanding the performance of FV3-HAILCAST is important as the transition from HRRR to RRFS occurs.

The overall design of both WRF-HAILCAST and FV3-HAILCAST are quite similar. In both cases, CAM-HAILCAST is coupled in one direction only to its underlying CAM: no microphysical information is passed back to the CAM. Additional details of the physics are provided in Adams-Selin and Ziegler (2016) and Adams-Selin et al. (2019). All microphysics packages are supported. The workflow for the RRFS was updated to support FV3-HAILCAST in early 2022.

b. Model data

During the 2019 SFE, CAPS ran an LAM FV3 ensemble consisting of 14 members with both mixed physics and perturbations in initial conditions. Seven of the members (*core*) were initialized with the North American Mesoscale Model (NAM) with a variety of boundary layer, microphysics, land surface, and surface layer parameterizations. One member was initialized using GFS analyses and forecasts. The remaining six members (*pert*) used the same physics options, but were initialized with initial condition perturbations from the 2100 UTC version of the Short Range Ensemble Forecast System (SREF) added to the NAM analyses. The full configuration of all members is provided in Tables 2 and 3 of the 2019 SFE operations plan (`https://hwt.nssl.noaa.gov/sfe/2019/docs/HWT_SFE2019_operations_plan.pdf`). Results from a representative subset of members will be discussed; their configurations are listed in Table 1.

During the 2020 SFE, FV3-HAILCAST was run by NSSL within the *sarfv3-ICs02* CLUE member. It used the LAM FV3 configuration with initial and boundary conditions from the Unified Model (UM) as part of an experiment testing UM ICs (Roberts et al. 2022). In the 2021 SFE, FV3-HAILCAST was run within NSSL's FV3-LAM with initial and boundary conditions from the GFS version 16 (GFSv16). Physics options for both years' configurations are listed in Table 1.

WRF-HAILCAST was also run as part of the experimental HRRR Ensemble (HRRRE; Kalina et al. 2021), with the physics configuration for the 2020 SFE summarized in Dowell (2020).

The HRRRE uses the WRF-ARW dynamical core and initial/boundary conditions are generated by the 36-member HRRR Data Assimilation ensemble analysis System (HRRRDAS). Additional configuration details are provided in Table 2. In addition to WRF-HAILCAST, two other hail forecasts were produced using HRRRE data and evaluted during the 2020 SFE: the Thompson method, generated using the hail size distribution within the microphysical parameterization (see discussion of method in Milbrandt and Yau 2006; Gagne et al. 2019), and calibrated machine learning methods (ML; Gagne et al. 2017; Burke et al. 2020). Subjective verification discussion during the 2020 SFE evaluated all three hail forecasting methods.

TABLE 1. SFE FV3 configurations. All used RRTMG radiation (Iacono et al. 2008); Thompson (Thompson and Eidhammer 2014), NSSL (Mansell et al. 2010), or Morrison (Morrison et al. 1997) microphysics, scale-aware MYNN (Olson et al. 2019) or GFS EDMF (Han et al. 2016) boundary layer parameterizations, NOAH (Chen and Dudhia 2001) or RUC (Smirnova et al. 2016) land surface models, and GFS (Long 1986, 1989) or MYNN (Olson et al. 2021) surface layer parameterizations.

Year	Name	ICs/LBCs	Microphysics	PBL	LSM	SFC Layer
2019	*core_cntl*	NAM	Thompson	MYNN-SA	NOAH	GFS
2019	*core_mp1*	NAM	NSSL	MYNN-SA	NOAH	GFS
2019	*core_mp2*	NAM	Morrison	MYNN-SA	NOAH	GFS
2019	*core_pbl2*	NAM	Thompson	EDMF	NOAH	GFS
2019	*pert_sfcl1*	NAM+SREF	Thompson	MYNN-SA	RUC	MYNN
2020	*sarfv3-ICs02*	UM	Thompson	MYNN-SA	NOAH	GFS
2021	*NSSL FV3-LAM*	GFSv16	NSSL	MYNN-SA	NOAH	GFS

TABLE 2. 2020 HRRRE configuration, using Thompson microphysical (Thompson and Eidhammer 2014), MYNN planetary boundary and surface layer (Nakanishi and Niino 2009; Benjamin et al. 2016), and RUC land surface (Smirnova et al. 2016) parameterizations.

Year	Name	ICs/LBCs	Microphysics	PBL	LSM	SFC Layer
2020	HRRRE	HRRRDAS	Thompson	MYNN	RUC	MYNN

The domain and initialization timing of all SFE models follow the design of the Community Leveraged Unified Ensemble (CLUE; Clark et al. 2018), which during 2019-2021 consisted of a CONUS domain with 3-km horizontal grid-spacing and initialization daily at 0000 UTC. The verification results shown here will be limited to the portion of CONUS defined daily at each SFE as the “domain of the day” to ensure the objective and subjective verification results discuss the same geographical region.

c. MRMS MESH

All verification will be conducted using the Multi-Radar Multi-Sensor Maximum Estimated Size of Hail (MRMS MESH hereafter, Witt et al. 1998; Lakshmanan et al. 2006; Smith et al. 2016) as a validation source. MRMS MESH data is available on a 1-km horizontal grid covering the full CONUS with 2-min temporal frequency. Use of this dataset admittedly has a number of drawbacks, including lesser skill delineating between hail with significantly severe (> 50 mm) and severe (between 25 and 50 mm) diameters (Ortega 2018) and determining hail occurrence over the southeast U.S. (Murillo and Homeyer 2019; Murillo et al. 2021). However, at this time the MRMS MESH dataset was the only radar-based hail size estimate available at sub-hourly resolutions. It has been found to successfully distinguish between sub-severe (< 25 mm) and severe (> 25 mm) diameter hail (Ortega 2018) and is preferable to public severe hail reports with underlying population biases (Allen and Tippett 2015). We refer readers to Wendt and Jirak (2021) for a full exploration of differences between hail climatologies generated by *Storm Data* storm reports and MRMS MESH. The full spatial coverage of MRMS MESH also allows object-based verification by hail swath as opposed to by singular report, a particularly important factor given recent research examining the evolution of a storm’s hail production over its lifecycle (Kumjian et al. 2021).

As in Adams-Selin et al. (2019), the MRMS MESH dataset was truncated at 19 mm (0.75 in) in deference to the original Witt et al. (1998) algorithm formulation only using hail reports of that size or larger. Because of this truncation, hail swath objects in the MRMS MESH field with maximum sizes larger than 25 mm were more frequent than objects with a maximum size between 19 and 25 mm. In the object-based verification method (detailed later in Section 2e), only matched hail swaths were evaluated to avoid penalizing where the model failed to predict convection. Performance diagrams, a frequently used method for evaluating convective event forecast skill, do not include correct forecasts of null events and therefore should only be used for relatively infrequent events.Thus, all object-based statistics in this study were calculated for a threshold of 38-mm (1.5-in) hail or larger, to allow for a large enough population of objects with peak hail sizes below that threshold. A larger threshold (e.g., 50 mm) was also considered, but

50 mm hail events did not occur frequently enough for regular subjective verification during the HWT. Further discussion of this decision is provided in Section 3b.

It should also be noted in both HAILCAST and MRMS MESH hailstones are assumed to be spherical. Such an assumption is likely invalid, particularly for larger hailstones (e.g., Shedd et al. 2021) and hailstone mass would be a better predictor. However, addressing this issue is beyond the scope of the current study.

d. Upscaling neighborhood configurations

Neighborhood verification of model hail forecasts was based on the upscaling smoothed neighborhood maximum ensemble probability (NMEP^{smooth}) method described in Schwartz and Sobash (2017); this method is also presented as the practically perfect forecast verification method by Hitchens et al. (2013). Both model forecast and MRMS MESH hail size datasets were prepared for this method by determining maximum size at each native grid point over all times during successive 12-12 UTC 24-h periods. This aggregation was accomplished using the Model Evaluation Tools (MET hereafter, Brown et al. 2021). After aggregation, to upscale the data, model and MRMS MESH data are each regridded to a coarser grid (Figs. 1a,b). In the results shown here many of the ensemble members are evaluated individually. In these cases the coarse grid is binary with the member either predicting hail occurrence of a specific size or not. For ensemble data, the coarse grid is an ensemble probability of hail occurrence of that size. After regridding, the data are then smoothed over a set neighborhood of points (Fig. 1c).

Several different versions of upscaling neighborhood verification exist in the literature, often with conflicting terminology. Schwartz and Sobash (2017, SS17 hereafter) reviewed many of these configurations for ensemble verification of any forecast type. For convective forecasts, the terminology and methods of Hitchens et al. (2013, HBK13 hereafter) are often used. Finally, the MET software itself via the *regrid_data_plane* command uses yet a third set of terminology. To provide clarification, explicit MET inputs will be discussed in the format of both SS17 and HBK13.

SS17 identify two controls on the generation of smoothed NMEP at a grid point i. The searching radius, x km, is the distance from i within which is searched for the occurrence of an event. After application of this radius, the resulting field contains a binary yes/no probability of if an event occurred within x km of grid point i (Fig 1b). (If an ensemble is being evaluated, the average of

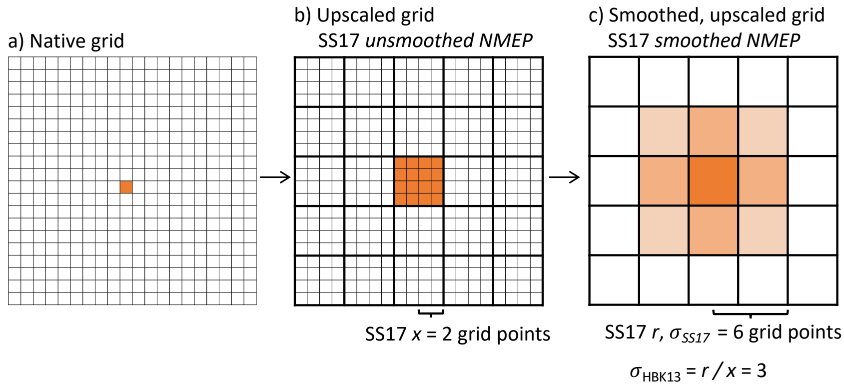


FIG. 1. Example of upscaling neighborhood configuration. (a) Example data on its native grid. Orange represents occurence of hail above our chosen threshold, expressed as a binary probability. (b) Data upscaled to a coarser grid, but still a binary probability. (c) Upscaled grid after smoothing. Orange shades are the smoothed probability field, and now also the forecast probability of the event.

the binary fields from all members can then be calculated.) This field, termed the unsmoothed NMEP by SS17, can remain in the native grid resolution or be converted to a coarser resolution (c.f. Figs. 1a,b of SS17 for examples of unsmoothed NMEP in coarse and native resolution). The smoothing radius, r km, is the spatial scale over which smoothing is performed (Fig. 1c). In many cases, including here, the smoothing is performed via a Gaussian standard deviation filter, σ. Hence SS17 states in these cases, r is "effectively replaced" by σ_{SS17}, resulting in a σ_{SS17} with units of km (e.g., Sobash et al. 2016). Conversely, HBK13 interpret σ as the spatial confidence one could have in a forecast of that event type. They combine the two radii of SS17 into one unitless σ_{HBK13} by calculating r/x, resulting in values around 0.75 to 3.0 with smaller values representing higher spatial confidence (smaller smoothing radius).

The *regrid_data_plane* MET tool takes three input arguments: *width*, *gaussian_dx*, and *gaussian_radius*. The *gaussian_radius* and *gaussian_dx* arguments are equivalent to the r and x values of SS17, and the ratio of the values (*gaussian_radius*/*gaussian_dx*) to σ_{HBK13}. The *width* value is number of native grid points that take part in the regridding of a given point (and therefore also the resolution in native grid points of the output unsmoothed NMEP field). The 24-h configuration follows the processes of Adams-Selin et al. (2019) and Gallo et al. (2021): the data is regridded to the 80-km NCEP 211 grid. During this process, the *width* argument was set to 27 for the 3-km model data and to 40 for the 1-km MRMS MESH. The maximum value within the box was used for the regridding. Both datasets were then set to a binary 1 or 0 value based on a threshold of 38 mm (1.5 in). A verification threshold of 38-mm hail was selected after evaluation at the 2019 SFE revealed larger hail sizes did not occur frequently enough for the desired complementary ongoing subjective evaluation.

The model data was further smoothed using a Gaussian filter with a Gaussian distance (*gaussian_dx*, SS17 x) of 81.271 km and Gaussian radii (*gaussian_radius*, SS17 r) of 81.271, 100, 120, 140, and 160 km. These values correspond to σ_{HBK13} of 1, 1.25, 1.5, 1.75, and 2 (using an x of 80 km instead of 81.271.) Because the regridding to a coarser dataset occurrs in MET before the smoothing, values of $\sigma_{HBK13} < 1$ could not be used as r could not be less than *width*. For the sake of clarity, future references to the Gaussian standard deviation filter in this text will use the definition of σ_{SS17}, or r. For verification of the HRRRE ensemble, the unsmoothed binary thresholded NMEP field on the NCEP211 grid for each ensemble member was averaged, to create a probability the ensemble would have predicted hail of at least the threshold size within that grid box, before the additional Gaussian smoothing was performed.

The observational dataset was not smoothed in agreement with the studies of Adams-Selin et al. (2019) and Gallo et al. (2021). After all regridding and smoothing processes were complete, verification occurred using MET's *grid_stat* to compute the reliability and other probabilistic-based statistics.

e. Object-based configurations

For the object-based verification, model data was left on its native 3-km domain. The MRMS data was regridded from its native 1-km grid-spacing by using a maximum value within a 1.5-km radius of each CLUE domain grid point, as in Adams-Selin et al. (2019). This method ensured the maximum hail size within each hail swath was preserved.

Three different spatiotemporal configurations for object-based matching were used. The 24-h configuration, consisting of hail swaths aggregated over a 24-h period (12-12 UTC) before verification, was designed to match hail swaths produced by supercell/multicell families or a single Mesoscale Convective Systems (MCS). This type of forecast was designed to be similar to what would be issued by the Storm Prediction Center as a Day 1 Convective Outlook. The 6-h configuration is designed to matched similarly sized swaths as the convective outlook configuration, but aggregated over a smaller time period (6 h); this verification attempts to mimic verification of a watch. Finally, the 1-h configuration is designed to validate forecasts that would be useful to forecasters issuing a warning, and is configured to match 1-h aggregated hail swaths produced by individual storm cells. In practice, the 6-h configuration produced results very similar to the 24-h configuration, so further discussion will be limited to the 24- and 1-h configurations. The similarity of the 6- and 24-h configuration verification results aligns with previous research that found most severe weather at a point occurs within a 4-h period (Krocak and Brooks 2020). Each of these configurations was developed using the Method for Object-based Diagnostic Evaluation (MODE; Davis et al. 2006a,b). Examples of forecast and observed hail swath objects using the 24- and 1-h configurations for a case in southern Texas on 28 May 2020 is provided in Fig. 2.

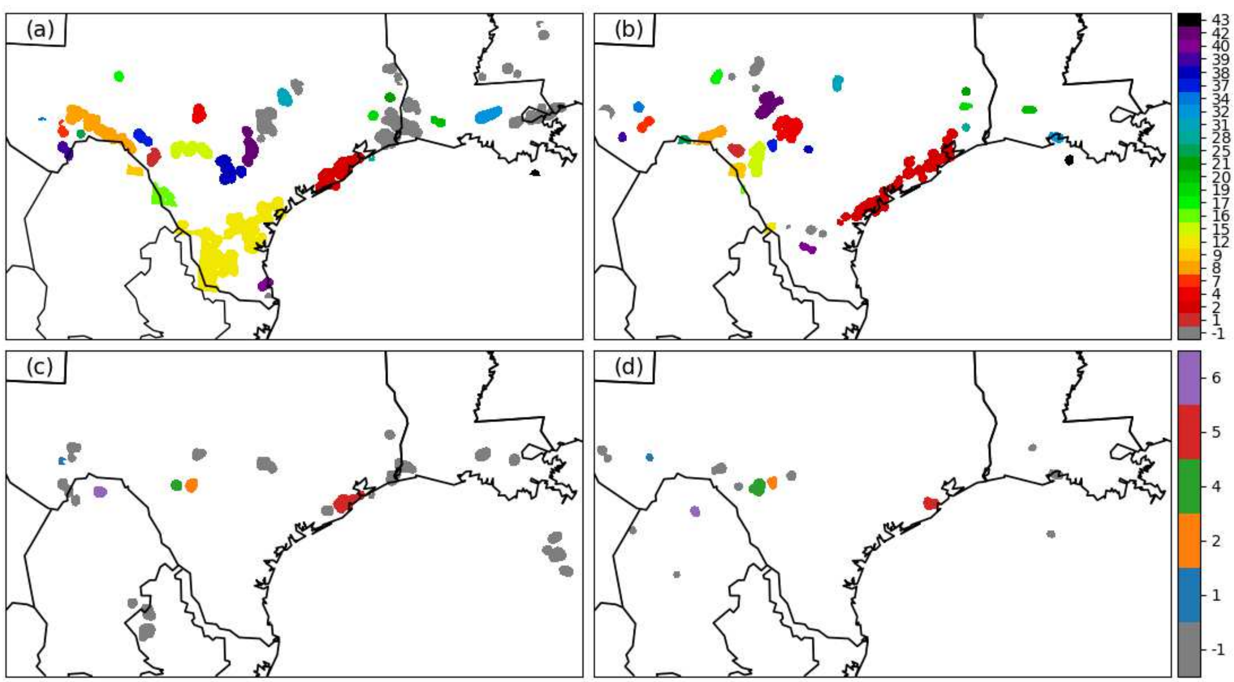


FIG. 2. Identified objects from the (a,b) 24-h configuration for a 24-h period ending 12 UTC 29 May 2020 and (c,d) 1-h configuration for a 1-h period ending 22 UTC 28 May 2020. Left column is from FV3-HAILCAST forecasts; right column from MRMS MESH. Non-matched objects are shown in grey (-1 on color bar); matched objects are shown via matching colors in each row. A total of 24 matched objects were identified in (a,b) and 5 in (c,d). Note that the numbers identifying the objects are not consecutive; matched pairs that do not meet the required interest threshold are removed from final matched object output by MODE.

The MET tools have been augmented to include a suite of use cases to demonstrate MET tools usage, in a framework called METplus (Brown et al. 2021). A METplus hail verification use case was developed that applies the 1-, 6-, or 24-hour configurations, and can be customized via user-selected time period(s). If an ensemble is being validated, each member can be verified individually or the ensemble maximum as a whole. The verification results from MODE are used to calculate contingency table statistics from the matched objects, displayed visually via performance diagram as in the next section.

In MODE, objects are identified in the forecast and observation fields using a convolution radius of 4 grid points and a convolved field threshold of 12.5 mm (Adams-Selin et al. 2019). Objects smaller than 4 grid points are omitted from analysis. The forecast and observation objects are matched using MODE's fuzzy logic function, with emphasis on distance between objects and their respective areas and orientations. Additional configuration information is provided in Table 3. The difference between 1- and 24-h configurations was primarily achieved through increased object merging in the 24-h configuration but suppressing it entirely in the 1-h configuration. Performance diagrams are computed from matched pairs; unmatched pairs are omitted to avoid penalizing where the model failed to predict convection.

Table 3. MODE configurations.

Configuration option	24-h	1-h
Convolution radius	4 gridpoints	4 gridpoints
Convolution threshold	12.5 mm	12.5 mm
Area threshold	4 gridpoint	4 gridpoints
Max distance between centroids	400 km	400 km
Merging threshold	0.5	0
Total interest threshold	0.7	0.5

3. Identification of a "good" hail forecast

a. Subjective evaluation of verification methods

Participants of the 2020 HWT SFE were surveyed to understand internal attitudes about convective hazard forecasting skill. Forty-one unique participants provided answers. The HWT SFE is designed to be a collaboration among forecasters, researchers, and model/algorithm developers, with its primary goal a two-way exchange of information and products between research and operations (e.g., Kain et al. 2003; Clark et al. 2012b; Gallo et al. 2017b). The information exchange is intentionally both subjective, via discussion, and objective, via statistical evaluation, to encourage dialog about the usefulness of products. At the 2020 SFE, 17 of the 41 participants that answered our survey were identified as forecasters, 18 researchers, and 11 developers, thereby representing a cross-section of representative interests from the severe convective hazard field. The following questions were asked:

1. (1.1) What do you mean when you say a 1.5-in hail forecast is "good"? (1.2) Do you think any of these figures successfully capture your opinion of the skill of the two different 1.5-in hail forecasting methods over the course of the week? Why or why not?

2. (2.1) Do you think validating hail forecasts over different time/spatial scales is helpful? (2.2) How effective at capturing hail forecast performance over the different time/spatial scales do you feel the three pairs of figures are?

The figures referenced in these questions are shown in Fig. 3, and consist of a variety of methods validating 38-mm hail forecasts over the course of one week during the SFE. Verification of 38-mm hail over a week period was selected after the 2019 SFE revealed 50-mm hail frequency was not high enough for evaluation on a daily basis; lowering the threshold and extending the verification period provided enough forecasts to evaluate. Six total figures were provided for evaluation of the hail forecasts each week.

Participants expressed a range of opinions about the contents of a "good" hail forecast. The total number of responses received to Question 1.1 was 44. (Three participants answered the questions twice, but on different days.) "Correct location" was noted most frequently, in 30 of 44 responses. Half as many responses (16) included size, and only 6 responses also noted timing. Of the participants concerned with hail size, several noted they would consider a hail size forecast of within 0.5 in (12.5 mm) of the observed reports as "good".

All responses to Question 1.1 included mention of correct hail size and/or location as important ingredients in a "good" hail forecast (participants could provide multiple ingredients in a single response). These answers were further analyzed for overlap among responses. "Correct location" could be divided into two groups of emphasis: accurate forecast of individual storm location,

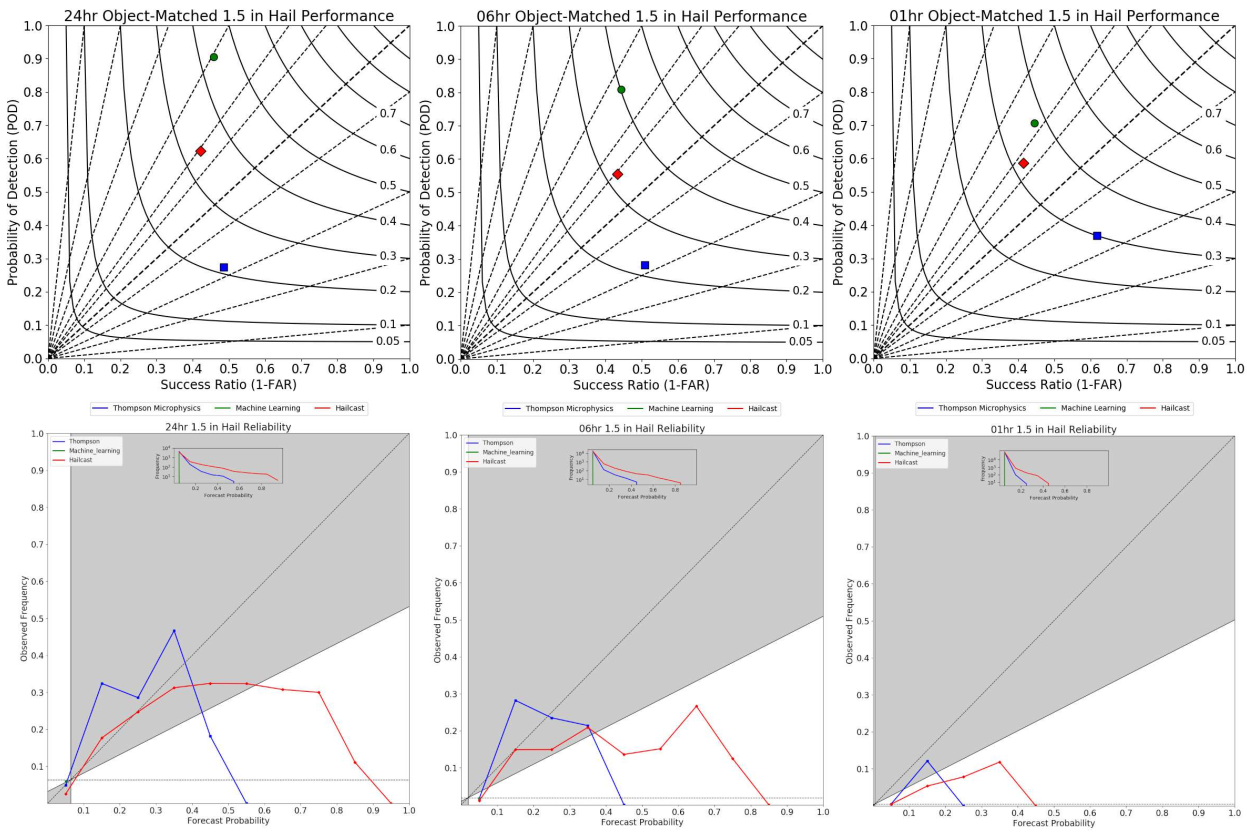


FIG. 3. Reproduction of sample evaluation figure shown the 2020 HWT SFE participants on the Friday of each week. The top row shows performance diagrams, calculated as in Section 2e, for the verification of 38-mm (1.5-in) hailfall forecasts produced within the HRRRE over each full week. Solid curves are constant Critical Success Index (CSI). Dashed lines are lines of constant bias, with a bias of 1 occurring along the diagonal, underforecast bias below, and overforecast bias above. The bottom row contains reliability diagrams, calculated as in Section 2d. 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. Inset plots showing the frequency of forecasts in each probability bin. The columns show a range of spatial and temporal scales: the 24-h (left), 6-h (middle), and 1-h (right) configurations described in Section 2e above.

or accurate forecast placement of Gaussian-smoothed neighborhood probabilities of 38-mm hail. Responses focusing on individual storm location often also provided what they considered to be a reasonable spatial error threshold: for example, "within 2 or 3 counties" or "within 25-50 miles", although it was noted that negative public response to even small spatial forecast errors within densely populated areas could be significant. Distinguishing hail-producing ability among multiple CAM-forecasted convective cells was also desired. Responses concerned with accuracy of storm location, as opposed to probabilities, were mostly also concerned with accuracy of forecasted hail size (8 of 12 responses).

Conversely, responses focusing on the Gaussian-smoothed neighborhood probabilities wanted to see a high probability of detection (POD) with a small area of false alarm to focus attention on regions with the highest probability of hail. This group of responses was largely concerned with model-predicted regions of high forecasted probability of hail, on the order of 100-200 km, in which hail did not occur. Only 3 of 18 “correct location of probability” responses also mentioned accuracy of size in their response.

A total of 13 responses to Question 1.2 were collected, all of which found the figures helpful. Several (5) participants found the performance diagrams conveyed skill more clearly, mentioning ease at determining over- and under-forecasting; a few requested displays of additional size thresholds. The responses noted a “lack of signal” from the reliability diagrams. Interestingly, the responses favoring the performance diagrams were not limited to those who considered either accurate storm location or probabilities more important; participants with different ideas of what constituted a “good” hail forecast still found the performance diagrams helpful.

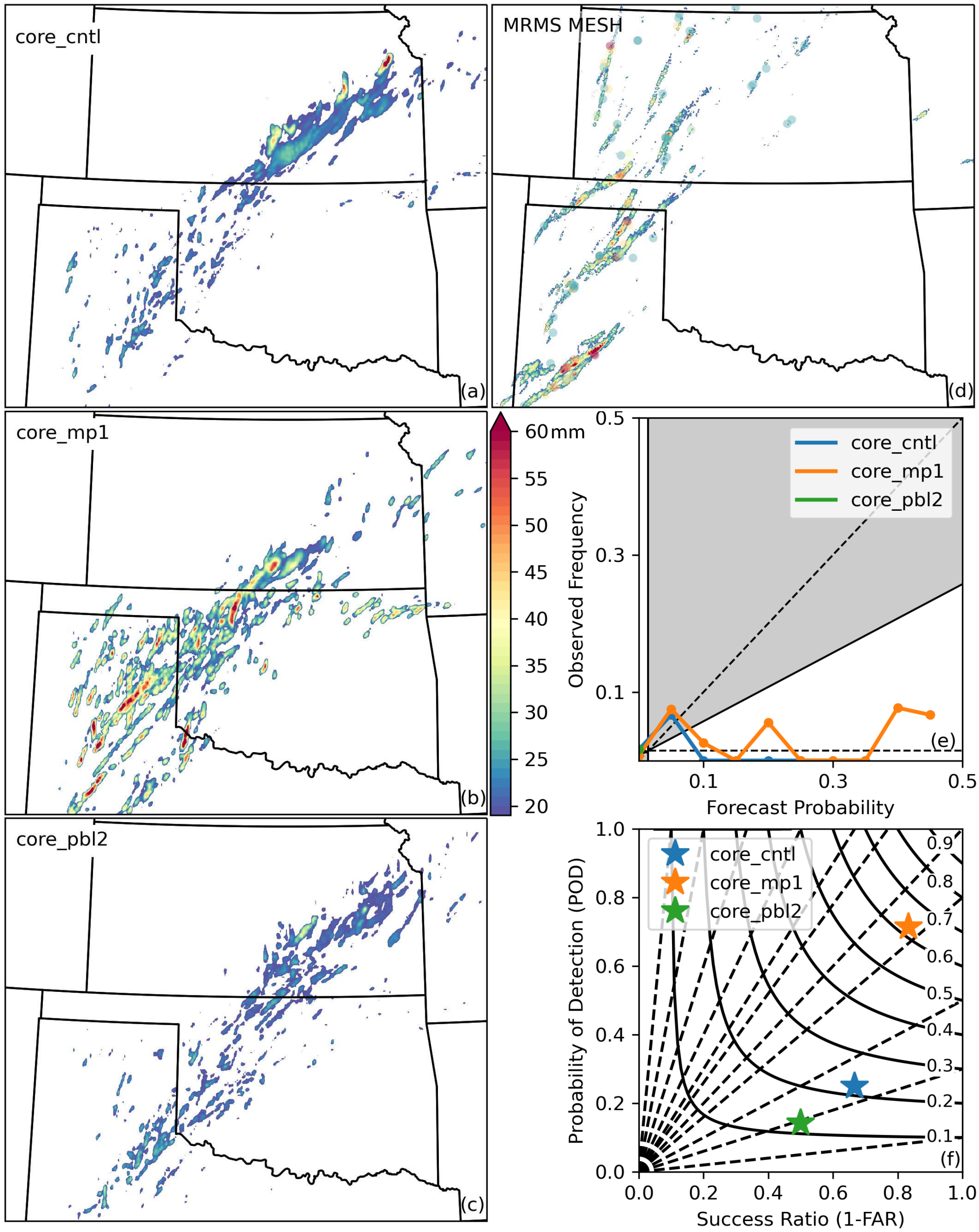
Results from Question 2.1 were overwhelmingly in favor of verification statistics calculated over a range of spatial and temporal resolutions with no responses opposed. Participants liked having verification conducted over 24-h time periods to understand the full storm system as an event, as well as periods smaller than 24 h to understand the model’s effectiveness at forecasting the evolution of the storm system. Such responses suggest more may have been interested in correct timing as part of a “good” hail forecast than explicitly stated in their answer to Question 1.1. Many responses (8) suggested 4 h as a preferred resolution as opposed to the 6 and 1 h shown here; a few commented that expecting accuracy on a 1-h timescale is too unrealistic for 24-36 h forecasts. All responses to Question 2.2 (23 in total) found the varying spatiotemperoal scale verification figures helpful for understanding model performance. Again, a few respondents (4) expressed preference for the performance diagrams citing faster interpretation; none expressed preference for the reliability diagrams.

b. An example case study verification method comparison

To further explore the idea of a "good" hail forecast and the effectiveness of different verification methods, three example FV3-HAILCAST hail size forecasts covering 12 UTC 23 - 12 UTC 24 May 2019 are provided in Fig. 4 along with radar-estimated hail size data and *Storm Data* storm reports. Verification results from the upscaling neighborhood configuration (Section 2d) and the object-matching method (Section 2e) are also included; for a description of these diagrams as used for hail forecasting reference Adams-Selin et al. (2019). Forecast and observed hail was aggregated over the full 24-h period as described above. Immediately evident is the wide variability of skill among FV3 members, which will be discussed further in Section 4. In fact, member *pert_sfcl1*, not shown, produced no hail of 38 mm or larger. This date was selected for case study examination as it is roughly representative of each member's performance over the full 2019 SFE.

The event produced several extended hail swaths in western Texas and the Texas and Oklahoma panhandles with peak observed hail sizes in the swaths estimated above 50 mm (Fig. 4d). Shorter swaths were also evident in western Kansas, with smaller peak sizes around 40 mm. The three hail forecasts shown each have a range of advantages and drawbacks. Member *core_cntl*, while incorrectly predicting that more severe hail will occur in eastern Kansas instead of western Texas and the Oklahoma panhandle, does correctly capture that the more intense hail will occur in swaths from single cells. The *core_mp1* forecast better places the location of the severe hail, but forecasts too wide of coverage with several cells with at least 40 mm hail simulated in eastern Oklahoma. Finally, *core_pbl2* produces only a few small hail swaths with sizes larger than 38 mm but also has the least amount of false alarm.

The reliability diagram in Fig. 4e indicates an overforecast of 38-mm hail for all forecast probabilities of *core_mp1* larger than 5%, and almost no skill overall. The mismatched placement of the forecast and observed hail swaths in the central Texas panhandle, beyond the distance of the smoothing radius, contributed to the poor skill as did the extensive false alarm in Oklahoma. The widespread coverage of the severe hail in the *core_mp1* member resulted in high forecast probabilities using the Gaussian smoothing method, despite the two concepts not necessarily being related. The reliability curve of *core_cntl* is surprisingly similar to that of *core_mp1*, despite the latter displaying improved placement and number of the 38-mm hail swaths. *Core_pbl2* does not produce a non-zero reliability curve given the few locations where > 38-mm hail was evident.



418 FIG. 4. Verification case study using FV3-HAILCAST hail size forecasts (mm) over the period from 12 UTC 23 - 24 May 2019 from the *core_cntl* (a), *core_mp1* (b), and *core_pbl2*. (c) CAPS FV3 members from the 2019 SFE. MRMS MESH estimated hail size is in (d) along with *Storm Data* storm reports shown as partially transparent large dots. The reliability diagram (e), calculated as in Section 2d, and the performance diagram (f), calculated as in Section 2e, are for 38-mm (1.5-in) hail for this 24-h period only.

Such results encapsulate the strengths and weakness of upscaling methods. *Core_cntl* is correctly penalized for the large area of false alarm, but perhaps not correctly rewarded for the spatially offset hail swaths in the Texas panhandle similar in appearance to the MESH estimations. *Core_cntl* and *Core_pbl2* show almost no skill per the reliability diagrams. Such results, while truthful, do not provide any additional helpful information such as the peak hail sizes in the incorrectly placed hail swaths in *core_cntl* better capturing the hail-producing potential of the Southern Plains environment as opposed to that of *core_pbl2*.

The performance diagram (Fig. 4f) and the object-matching method of Section 2e both indicate an underforecasting bias of *core_cntl* and *core_pbl2*. Specifically, many of the matched hail swath objects from these members have forecast peak hail sizes below 38 mm but larger observed peak sizes. *Core_cntl* shows slightly higher skill (as determined by Critical Success Index; CSI) than *core_pbl2*, as was also shown by Fig. 4e. *Core_mp1* shows the highest skill using this verification method. The hail swaths objects in the Texas panhandle were matched, eliminating any skill penalty due to spatial offsets. However, because only matched observed and forecast hail swath objects were evaluated, the erroneously produced convection and severe hail by that member in eastern Oklahoma did not reduce the determined skill.

Evaluation of this case study further underscores the recommendation from HWT SFE particpants that multiple methods are necessary to truly understand the skill of a convective hazard forecast.

4. CAM-HAILCAST performance over 24-h periods

a. Upscaling neighborhood verification

The upscaling neighborhood verification reveals the difficulty of forecasting 38-mm (1.5-in) hail using any of the methods evaluated herein (Fig. 5). Such a result is unsurprising, given previous poor verification results in the literature of 50-mm hail predictions (e.g., Gagne et al. 2017, 2019; Adams-Selin et al. 2019). Comparison among the different forecasting methods across the years is still instructive, particularly when comparing performance of WRF-based and FV3-based methods and different Gaussian smoothing (σ) values.

In 2019 (Fig. 5a), the smaller magnitudes of forecast probabilities, across all members, is evident. (Note the zoomed-in horizontal axis in Fig. 5a). None of the four displayed members produced a probability of the occurrence of $>$ 38 mm hail larger than 0.45. The result reveals one of the drawbacks of using neighborhood verification methods. Spatially larger forecast areas of > 38-mm hail, or even simply forecasts that occurred across boundaries of the coarser grid, are translated into a higher magnitude probability of occurrence.

Core_mp1 strongly overpredicts the occurrence of this size hail (Fig. 5a). Per this verification method, the resulting forecast was largely even worse than a climatological forecast. Increasing the length of the smoothing radius (σ) shows only slight improvement in verification of higher probabilities, largely by shifting them to lower probabilities. For the other three members, increasing the smoothing radius simply reduced the number of forecast higher probabilities to be verified, resulting in lesser skill and underforecast occurrence of that hail size. Even using a smaller σ value, however, *core_cntl* still shows underforecasting relative to the other members. The four members, in sum, show a wide variety of performance of FV3-HAILCAST across different physics configurations during the 2019 SFE, although all lack in certainty.

The FV3-HAILCAST configurations running during the 2020 and 2021 SFE both show an increase in certainty and occurrence of forecast probabilities larger than 0.5 relative to 2019. Changes in the σ smoothing value do not greatly shift the subsequently calculated reliability curve at lower probability values (e.g., < 0.5) but results in large changes at higher probability values, suggesting only a few high probability forecast events. This conclusion is confirmed by the inset frequency plots in Figs. 5b,c. Per Fig. 5b the HRRRE-HAILCAST forecasts during the 2020 SFE are more skillful than the HRRRE-Thompson or FV3-HAILCAST methods. (HRRRE verification statistics were calculated in real time for subjective evaluation at the 2020 SFE therefore additional σ thresholds could not be tested.) Whether the improvement of HRRRE-HAILCAST over FV3-HAILCAST is due to the forecasts being sourced from an ensemble instead of a single member is not clear.

b. Object-based verification

The upscaling neighborhood verification discussed in the previous section provided information about a member's tendency toward over or underforecasting of 38-mm hail occurrence, but did not separate that tendency from an over or underforecasting of convection in general. While the *core_mp1* member significantly overforecast 38-mm hail per Fig. 5a, the 24-h configuration in Fig. 6a shows that member did the best job of identifying 38-mm hail among storms where

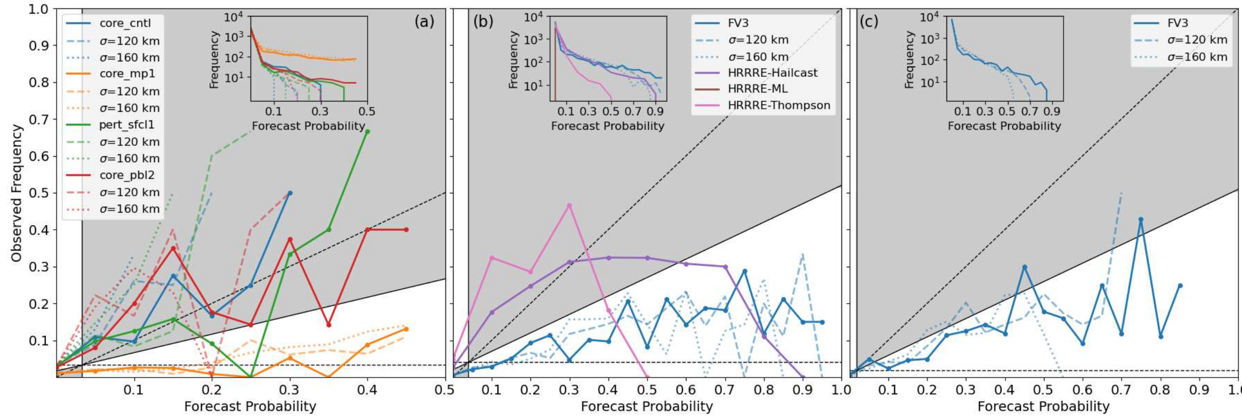


Fig. 5. Reliability diagrams for 38 mm (1.5 in) hail forecasts from the (a) 2019, (b) 2020, and (c) 2021 SFEs via the "convective outlook" configuration. Solid colored lines use a σ smoothing value of 80 km, fainter dashed and dotted lines 120 and 160 km, respectively. Note the zoomed in x axis of (a). The gridded HRRRE-ML probabilities were unavailable during the 2020 SFE.

hail did actually occur. That is, *core_mp1* simply overforecasts convection in general; where its convective forecasts were successful it was most skillful among the members at predicting 38-mm hail occurrence. That analysis is similarly displayed in Fig. 4b: the hail swaths of member *core_mp1* look most like those that actually occurred, there are simply too many of them. *Core_cntl*, *core_pbl2*, and *pert_sfcl1*, while showing higher skill values in Fig. 5a, underforecast hail size when convection is correctly forecast per Fig. 6a.

For the 2020 SFE, FV3-HAILCAST showed the least biased skill when distinguishing hail swaths that produced 38-mm hail. HRRRE-HAILCAST and HRRRE-ML displayed higher values of CSI, but were increasingly biased toward overforecasting, a trend that also appeared in Fig. 5b. The HRRRE-Thompson method, conversely, underforecast 38-mm hail both where convection was correctly simulated (Fig. 6b) as well as overall (Fig. 5b).

The 2021 SFE FV3-HAILCAST showed skill equivalent to the 2020 SFE FV3-HAILCAST; a somewhat surprising result given that the underlying model physics configuration changed between the years (Table 1). The overforecasting of 38-mm hail evident in Fig. 5c appears to be due to an overforecast of convection in general, as the member showed a slight underforecasting bias of 38-mm hail where convection was simulated correctly (Fig. 6c).

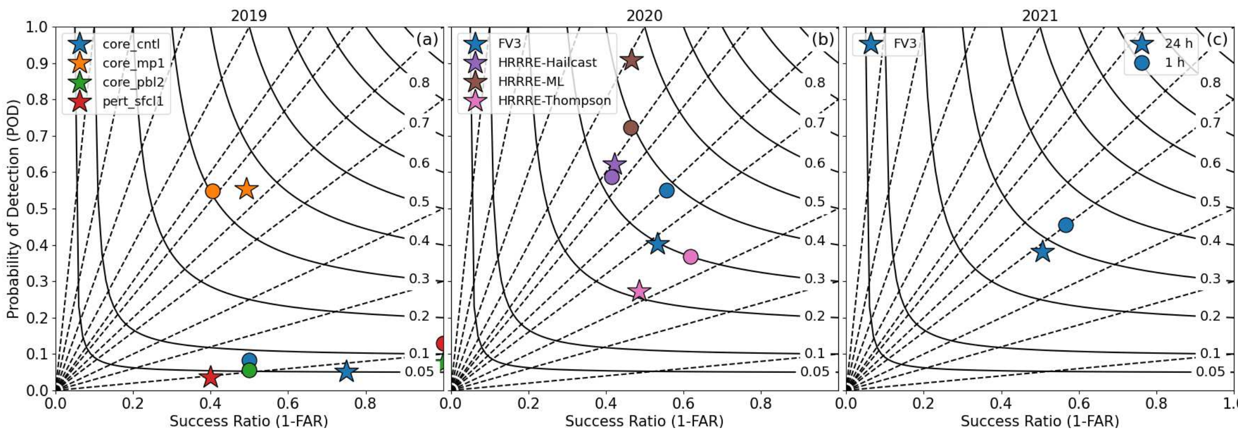


FIG. 6. Performance diagrams for 38 mm (1.5 in) hail forecasts from the (a) 2019, (b) 2020, and (c) 2021 SFEs via the 24-h (stars) or 1-h (circles) configuration.

c. Verification by size distribution

To further analyze the wide variability of FV3-HAILCAST performance among the 2019 SFE CAPS ensemble, the forecast hail distribution among 12.5-mm (0.5 in) size bins is shown in Fig. 7a. Given that MRMS MESH does not show skill at distinguishing among storms producing surface hail at 12.5-mm intervals (e.g., Ortega 2018; Murillo and Homeyer 2019) we are not using the distribution of MRMS MESH in Fig. 7a for verification, but instead as a rough baseline for CAPS member intercomparison. Notably, *core_mp1* produces more hail of all sizes than any of the other members or the MESH estimates. Such a result agrees with the analysis of the previous two subsections that *core_mp1* overproduced convection in general. Conversely, *pert_sfcl1* underproduced larger hail sizes compared to the other members and MESH, but was more comparable at small hail sizes. Such results suggest it produced a more appropriate amount of convection than *core_mp1*, as was similarly suggested by its more skillful appearance in Fig. 5a. However, FV3-HAILCAST produced less skillful hail forecasts within that convection, as indicated by the minimal large hail sizes for *pert_sfcl1* in Fig. 7a and strong underforecasting bias in Fig. 6a.

Several recent studies in the literature have examined how convection-allowing models with FV3 or WRF-ARW dynamical cores can show similar skill in forecasting convective features at multiple scales (Harris et al. 2019; Zhang et al. 2019; Snook et al. 2019; Gallo et al. 2021). Zhang et al. (2019) in particular examined the skill of 10 different 2018 SFE CAPS FV3 ensemble members at producing hourly accumulated precipitation. They found members with the Thompson microphysics scheme produced significantly more precipitation than the NSSL scheme, particularly at higher amounts; differences caused by boundary layer scheme changes were not as large (see Fig. S3, Zhang et al. 2019). While no hail or convective updraft information was included in that study, a similar difference in convective updraft and therefore hail forecasts could reasonably be expected to follow.

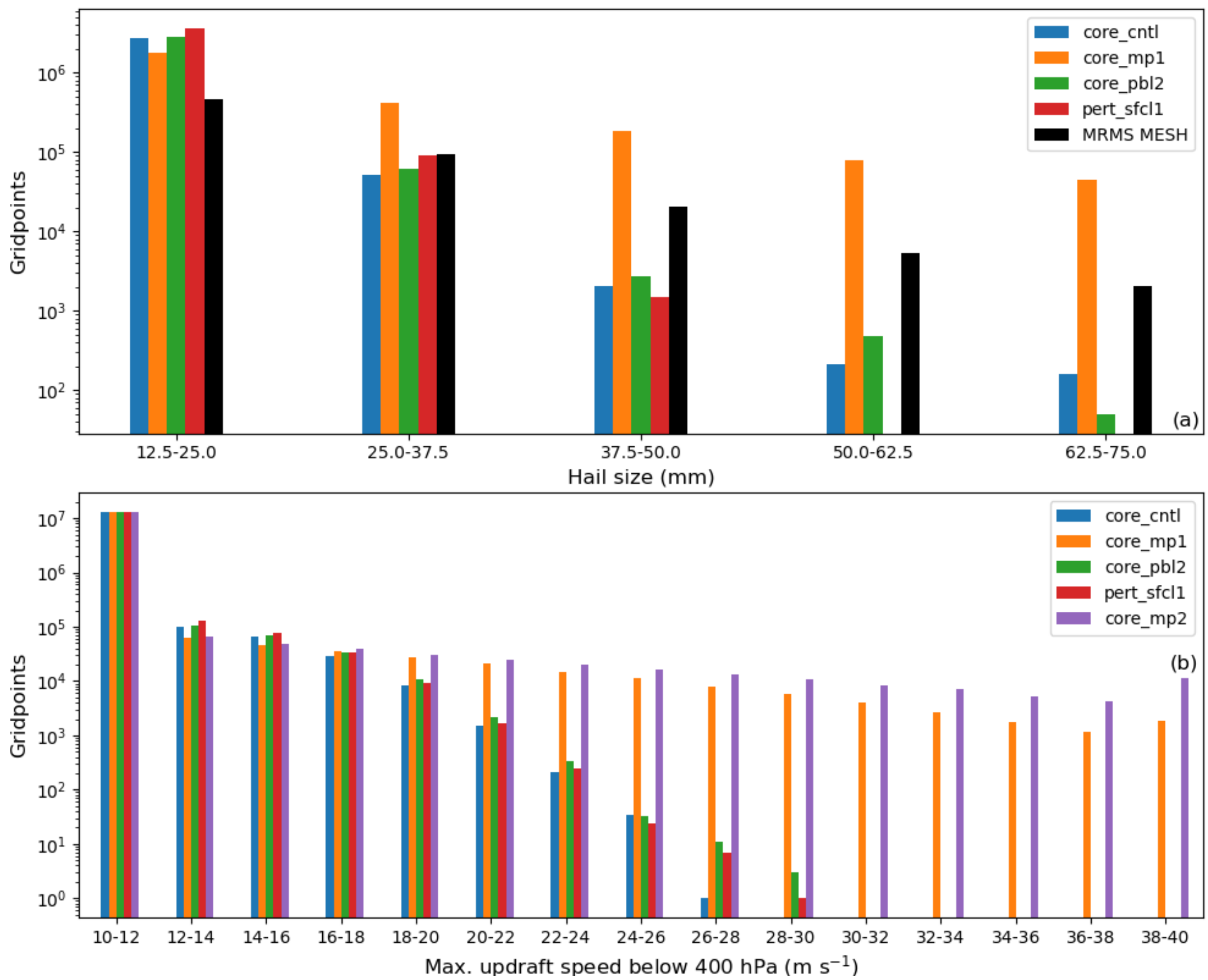


FIG. 7. Distribution of 24-h maximum hail size (a) and column-maximum updraft speed below 400 hPa (b) at every domain gridpoint during the 2019 SFE. MRMS MESH data is regridded to the SFE domain following the method outlined in Sec. 2d. MRMS MESH data shown for comparison only; MESH does not show skill at distinguishing among storms producing surface hail at 12.5-mm intervals (e.g., Ortega 2018; Murillo and Homeyer 2019).

Distribution of column-maximum updraft velocities across the subset of CAPS FV3 members during the 2019 SFE are shown in Fig. 7b. An additional member, *core_mp2*, is shown; this member has the same configuration as *core_mp1* except uses the Morrison microphysics parameterization (Morrison et al. 1997). Much like Zhang et al. (2019), a change in the microphysics parameterization (c.f., *core_cntl*, *core_mp1*, *core_mp2*) has a bigger impact than a change in the boundary layer parameterization (c.f., *core_cntl* and *core_pbl2*). A change in the surface layer scheme also has a smaller impact (c.f., *core_cntl* and *pert_sfcl1*). Unlike the 2018 SFE results or Zhang et al. (2019), in the 2019 SFE both the NSSL and Morrison members showed a larger distribution of higher updraft speeds compared to the Thompson members. CAPS FV3 members with identical microphysics configurations but different initial conditions still showed similar results (not shown). A potential possibility for the change in relative performance among the members with Thompson, Morrison, and NSSL microphysics is the switch from the custom CAPS implementation of the Common Community Physics Package Zhang et al. (CCPP; 2018) schemes used in 2018, to the NOAA Environmental Modeling Systems (NEMS) GFS CCPP implementation in 2019. The NSSL microphysics parameterization was also upgraded between 2018 and 2019 with increased snow and ice crystal fallspeeds along with larger maximum collection efficiency of graupel and hail collection of raindrops; these increases would enhance total precipitation and, potentially, system updraft speed (*T. Mansell, personal communication*). Whatever the cause, it is clear that the wide distribution of updraft speeds among CAPS FV3 members translates directly to the wide distribution of FV3-HAILCAST hail sizes. Members with larger updraft speeds corresponded to the members with higher amounts of larger hailstones, a reasonable result.

More skillful performance was seen by FV3-HAILCAST during the 2020 and 2021 SFEs compared to the 2019 forecasts, as also noted previously. The *sarfv3-ICs02* run, part of the 2020 SFE, used the Thompson microphysics parameterization as in *core_cntl* in the CAPS FV3. The *NSSL FV3-LAM*, part of the 2021 SFE, used the NSSL microphysics parameterization as in *core_mp1*. Because the FV3 dynamical core configuration used during these years was in flux, a specific reason for these changes is not readily identifiable. For example, the number of vertical levels used in the model shifted from 64 in 2019 up to 81 in 2020, before returning to 64 in 2021. The amount of explicit diffusion used also varied, increasing from 2019 to 2020, which would have a stabilizing effect on the model. However, it is evident that both the dynamical core configuration and the performance of FV3-HAILCAST slowly stabilized between 2019 and 2021, as is evidenced by the change in microphysics parameterization between 2020 and 2021 with no accompanying extreme change in skill like that seen among the 2019 CAPS members.

5. Time- and space-dependent verification

As discussed in Section 3a, participants in the 2020 SFE found hail forecast verification at a variety of time and spatial scales helpful. Comparison of the star (24-h) and circle (1-h) symbols in Fig. 6 reveals changes in forecast skill when shifting from the 24-h to the 1-h configurations across all three SFE years. In 2019, the results of *core_cntl*, *pert_sfcl1*, and *core_pbl2* do show some large shifts in False Alarm Rate (FAR) with small simultaneous changes in Probability of Detection (POD) or overall CSI. Given the small number of of 38-mm hail swath objects (< 5) produced by these three members in both the 24- and 1-h configurations, we do not consider these changes in skill significant. However, *core_mp1* produces many 38-mm hail swath objects at both 24- and 1-h configurations (Figs. 8a,b). Given the difficulty in successfully forecasting convective-scale features at 1-h intervals 12-36 hours in advance, it is unsurprising that the overall skill decreases from the 24- to 1-h configuration for *core_mp1*. The magnitude of the reduction in CSI is not large, however, suggesting that FV3-HAILCAST in this member can roughly simulate the timing of 38-mm hail development *if* the underlying convection is also correctly forecast.

Each 2019 SFE CAPS FV3 member showed a different peak in the diurnal cycle of all hail-producing convection. *Core_mp1* showed the largest number of hail swath objects of all sizes at 2100 UTC, followed by *core_pbl2* and *pert_sfcl1* at 2200 UTC, and finally *core_cntl* at 2300 UTC. MRMS MESH hail swath objects did not peak until 0000 UTC. Despite its unrealistically early peak in overall hail swath objects, the number of large (38-mm) hail swaths within *core_mp1* did not peak until 2200 UTC, only an hour before the MESH-estimated peak. This fairly successful capture of the temporal evolution of hail size within the objects was reflected by the still high CSI score in the Fig. 6a.

In the 2020 SFE, HRRRE-ML had a relatively large decrease in skill as calculated by CSI, but a similarly large reduction in bias (Fig. 6b). HRRRE-Thompson and FV3-HAILCAST showed a relatively large increase in skill, while HRRRE-HAILCAST's skill remained unchanged. Unfortunately the total 1-h object counts from the HRRRE methods were not archived, so to examine

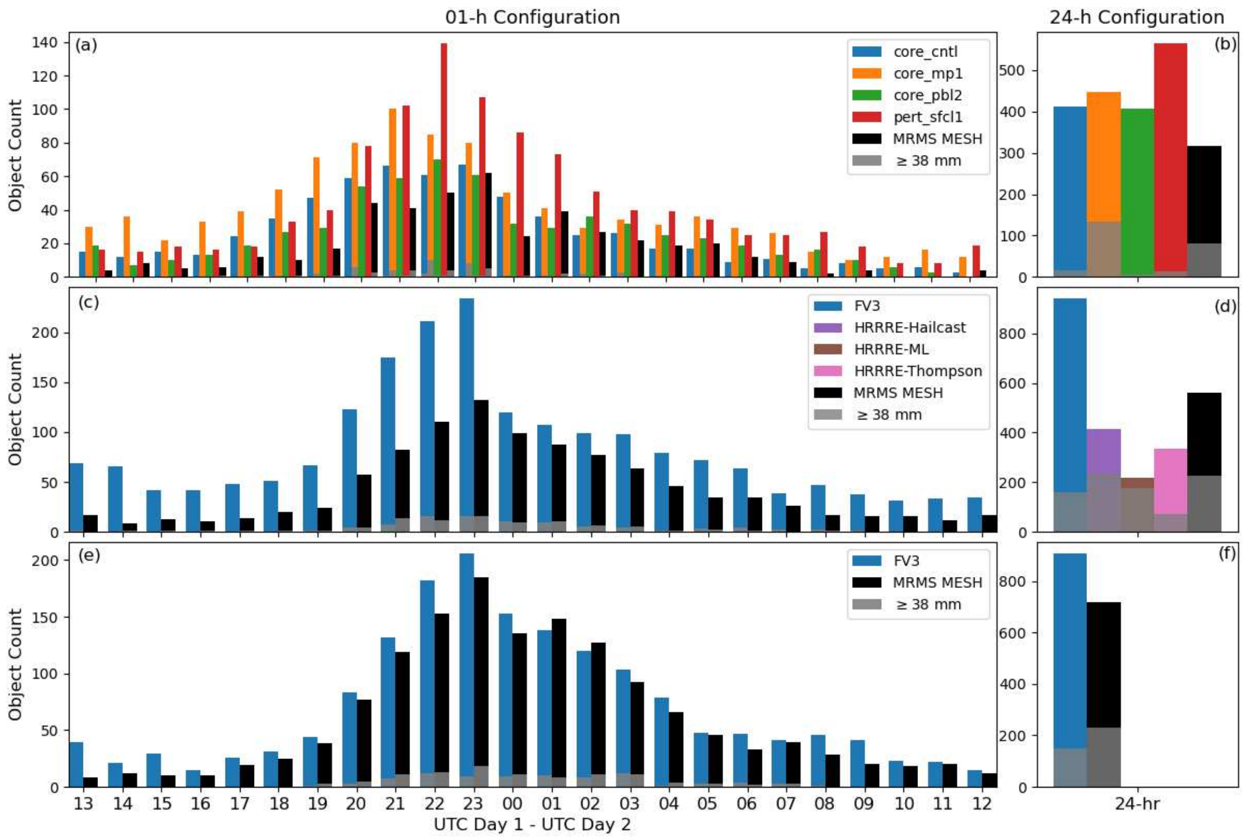


FIG. 8. Number of identified hail swath objects of all sizes (colors), and number containing hail at least 38 mm or larger (grey overlay). Model member or MRMS MESH identified in legends. Results from 2019 (a, b), 2020 (c, d), and 2021 (e, f) SFEs. Left column is 1-h configuration (a, c, e), right column is 24-h configuration (b, d, f). Note hourly HRRRE objects were not archived.

potential reasons behind these changes in skill Fig. 9, difference in peak sizes between matched forecast and observed hail swaths, is presented. As stated before, MRMS MESH is unable to skillfully differentiate between hail sizes at 5 mm intervals. Figure 9 is used to compare general bias in size distribution for matched objects.

FV3-HAILCAST produced more hail swath objects than the HRRRE methods or MRMS MESH in the 24-h configuration, but a roughly comparable number of ≥38 mm hail swaths (Fig. 8d). Such a result suggests an underforcasting of hail size, agreeing with the negative bias of FV3-HAILCAST in Fig. 6b. The size difference distribution of the 24-h matched objects (Fig. 9a) further confirms this result, showing a 5-10 mm underforecast between matched hail swaths to occur most frequently. The distribution of size differences is more evenly spread between a -10 to 10 mm difference for

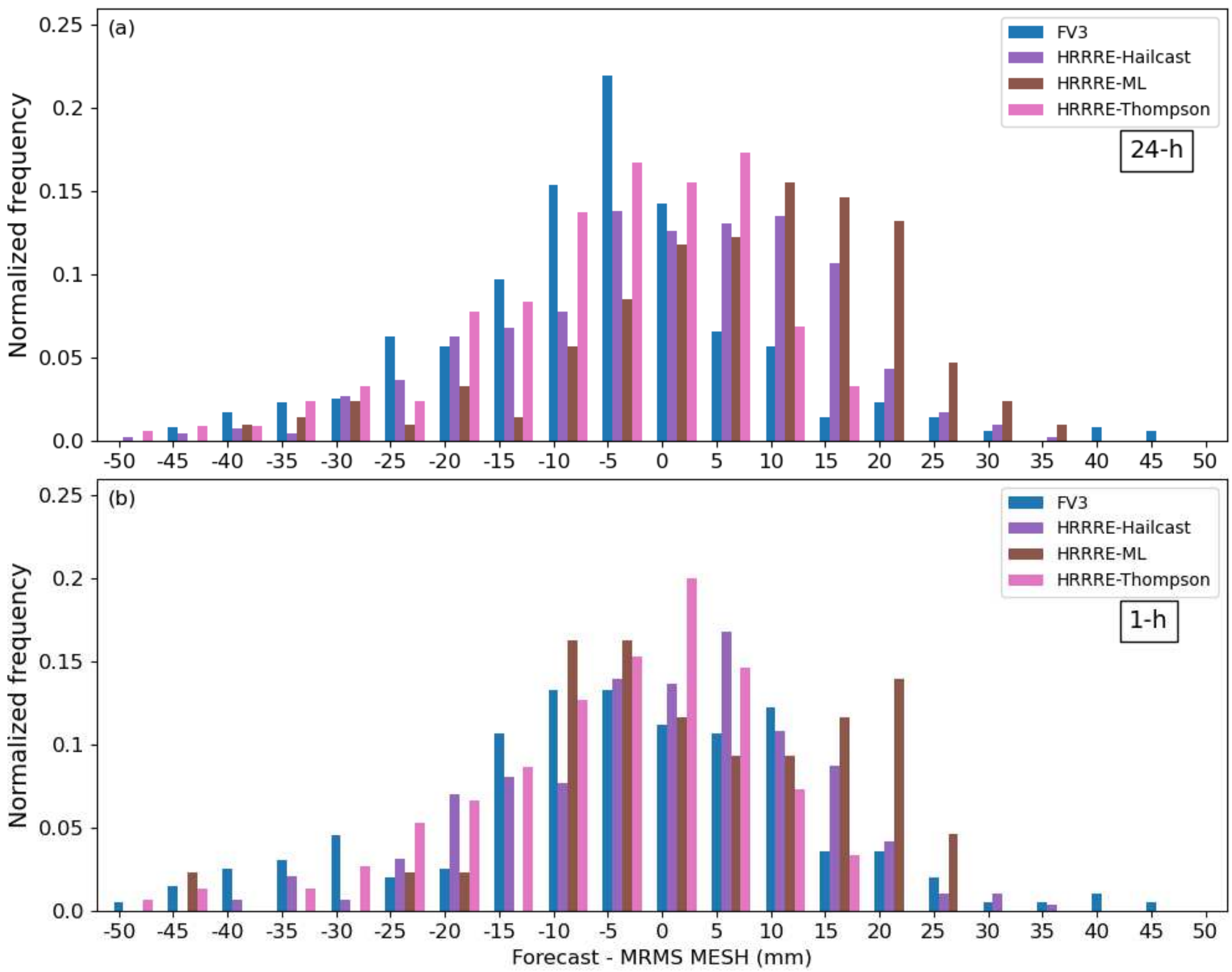


FIG. 9. Frequency of differences between the maximum hail size value from all matched forecast and observed (MRMS MESH-estimated) hail swath objects. Matched hail swath objects are identified from the 2020 SFE using the 24-h (a) and 1-h (b) configurations; results from 1-h configuration are summed over all forecast hours. Frequency of each bin is normalized by the total number of hail swath objects from the 2020 SFE from that model or algorithm (Figs. 8c,d). Note MRMS MESH data are shown for comparison only; MESH does not show skill at distinguishing among storms producing surface hail at 5-mm intervals (e.g., Ortega 2018; Murillo and Homeyer 2019).

the 1-h configuration. From Fig. 8c it is evident that while FV3-HAILCAST overproduces smaller magnitude hail swath objects from 20-23 UTC, this overproduction lessens after 00 UTC. It is possible that the more stringent matching criteria for the 1-h configuration screened out these overproduced smaller magnitude hail swath objects, improving the 1-h configuration skill scores.

The HRRRE-Thompson method similarly saw the peak of the difference distribution shift from 5 mm down for 24-h to 0 mm for the 1-h configuration (Fig. 9a,b). The HRRRE-ML distribution of differences shifted most dramatically, from a 10-mm peak difference in the 24-h configuration to -5 mm in the 1-h configuration. The HRRRE-HAILCAST size differences, conversely, were minimal between the two configurations. These differences suggest that the HRRRE-ML method was more skillful at identifying the temporal evolution of hail size within forecast objects, while the HRRRE-HAILCAST method was more skillful at identifying systems that would contain larger (i.e., 38 mm) hail.

The 2021 SFE FV3-HAILCAST also presented slightly improved skill at the 1-h configuration compared to the 24-h configuration, just like the 2020 FV3-HAILCAST results. The magnitude of increase in CSI is slightly less in 2021 compared to 2020, however. Figure 8e,f reveals that while FV3-HAILCAST still had an overproduction of hail swath objects during the 21-23 UTC hours, the overproduction was lessened compared to the 2020 results (Fig. 8c). Evaluation of the difference distributions for the 24-h and 1-h configurations (not shown) showed a most frequent difference of -5 mm for both, with a more narrow 24-h configuration.

6. Discussion and conclusions

In this study the performance of CAM-HAILCAST, within the HRRR-E and three implementations of the LAM FV3 over multiple spatiotemporal scales during the 2019, 2020, and 2021 NOAA SFEs, was used to explore the concept of a "good" hail forecast and the effectiveness of multiple verification methods. During the 2020 SFE these verification methods were subjectively evaluated in conjunction with a survey about the ingredients of a "good" hail forecast.

Survey participants differed in their idea of a "good" hail forecast and even in their definition of what a hail forecast consists. Approximately half considered a hail forecast to be similar to a single CAM model convective forecast, including identification of individual hail swaths. This section of respondents considered both the location and hail size of the forecast important, but did still consider a forecast with some spatial error to be "good". The other half of respondents considered a hail forecast to consist of broader probabilistic swaths of occurrence of a specific hail size. These respondents were most concerned with incorrect location of the forecast probabilities, particularly large regions of false alarm. Such results suggest that before verifying a forecast of hail or any convective hazard, investigators need to first determine the type of forecast desired by their users. Are they interested in localized, specific CAM output, or broader probabilistic information? The answer should contribute to the appropriate choice of verification technique.

As part of the survey, two verification techniques were examined to determine the effectiveness of each at assessing how "good" a variety of hail forecasts were. Upscaling neighborhood and object-matching methods were selected due to their frequent use in the literature for convective hazard verification (e.g., Hitchens et al. 2013; Schwartz and Sobash 2017; Gagne et al. 2017; Skinner et al. 2018; Flora et al. 2021; Miller et al. 2021; Gallo et al. 2021). In this analysis, the object-matching method was modified to only verify hail forecasts among matched forecast and observed objects, separating hail forecast skill from underlying general convective forecast skill. Both upscaling neighborhood and modified object-matching techniques can be performed with the MET or METplus software package (Brown et al. 2021) as described herein. Survey participants expressed preference for the object-matching method if their idea of a hail forecast focused on identification of individual hail swaths. Conversely, participants expressed preference for upscaling neighborhood methods if their idea of a hail forecast was a broader region of probabilities. All survey participants recognized the usefulness of verifying forecasts over multiple spatial and temporal scales.

Additional analysis was conducted examining the strengths and weaknesses of these two verification methods in evaluating CAM-HAILCAST forecasts from the three SFEs. Evaluation of FV3-HAILCAST hail forecasts found significant variability in skill among members of the CAPS FV3 multi-physics ensemble in the 2019 SFE. During the 2020 and 2021 SFE, however, the skill variability among physics options lessened and FV3-HAILCAST forecasts improved. Both upscaling neighborhood and object-matching methods were necessary to understand these results. For example, the upscaling neighborhood method found the 2019 SFE CAPS FV3 member with the NSSL microphysics scheme overproduced convection. However, where the convective forecast was correct, the object-matching method determined this member's FV3-HAILCAST hail size forecast was most skillful. Conversely, the CAPS FV3 member with the Thompson microphysics scheme produced a more realistic amount of convection, but hail size forecasts where convection was correctly forecast were poor. Subsequent years' forecasts with both of these microphysics parameterizations improved in overall convective distribution although hail forecasting performance remained steady. Given the underlying configuration of FV3 dynamic core was in flux during those three years but the FV3-HAILCAST algorithm remained fixed, such a result is not unexpected. Verification over different spatiotemporal ranges was also useful in understanding skillfulness of the FV3 core in simulating the diurnal convective cycle, as well as FV3-HAILCAST skill at simulating the hail temporal development within that convection. During the 2020 SFE, FV3-HAILCAST and HRRRE WRF-HAILCAST skill in identifying which convective cells would produce sizeable hail, over 24- and 1-h periods, were roughly comparable (Fig. 6).

In sum, it is recommended that future evaluation of convective hazard forecasts consider the forecast type expected by the end user and make use of multiple types of verification methods. A combination of upscaling neighborhood methods including different smoothing radii, object-matching methods that retain only matching forecast and observed objects to isolate convective hazard forecast performance from NWP performance, and verification over varying spatial and temporal scales are all recommended to gain a comprehensive picture of the performance of a forecast method and its perception by those using the resulting product.

Acknowledgments. This work was supported by NOAA Grant NA18OAR4590388 and NSF PREEVENTS Grant ICER-1855050. The Developmental Testbed Center (DTC) is funded by NOAA, the U.S. Air Force, the National Center for Atmospheric Research (NCAR), and the National Science Foundation (NSF). NCAR is a major facility sponsored by the National Science Foundation under Cooperative Agreement 1852977. The comments of Barry Bowers and two anonymous reviews helped to clarify and organize the results and text.

Data availability statement. All FV3-HAILCAST forecasts from the 2019, 2020, and 2021 SFEs have been archived by the authors and are available upon request. MET and METplus software is publicly available via `https://dtcenter.org/community-code/model-evaluation-tools-met/download`. MRMS MESH data was accessed through the Iowa Environmental Mesonet Data Archive (`https://mesonet.agron.iastate.edu/archive/`). FV3-HAILCAST software is available through the UFS weather model Github repository (`https://github.com/ufs-community/ufs-weather-model`).

References

Adams-Selin, R., A. Clark, C. Melick, S. Dembek, I. Jirak, and C. Ziegler, 2019: Verification of WRF-HAILCAST during the 2014-2016 NOAA/Hazardous Weather Testbed Spring Forecasting Experiments. *Wea. Forecasting*, **34**, 61–79.

Adams-Selin, R., and C. Ziegler, 2016: Forecasting hail using a one-dimensional hail growth model within WRF. *Mon. Wea. Rev.*, **144**, 4919–4939.

Alexander, C., and Coauthors, 2020: Rapid Refresh (RAP) and High Resolution Rapdi Refresh (HRRR) model development. *26th Conf. on Numerical Weather Prediction*, Boston, MA, Amer. Meteor. Soc., 8C.1, https://rapidrefresh.noaa.gov/pdf/Alexander_AMS_NWP_2020.pdf.

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

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Weather Rev.*, **144**, 1669–1694, https://doi.org/10.1175/MWR-D-15-0242.1.

Black, T. L., and Coauthors, 2021: A limited area modeling capability for the Finite-Volume Cubed-Sphere (FV3) dynamical core and comparison with a global two-way nest. *J. Adv. Modeling Earth Systems*, **13**, https://doi.org/10.1029/2021MS002483.

Britt, K. C., P. S. Skinner, P. L. Heinselman, and K. H. Knopfmeier, 2020: Effects of horizontal grid spacing and inflow environment on forecasts of cyclic mesocyclogenesis in NSSL's warn-on-Forecast system (WoFS). *Wea. Forecasting*, **35**, 2423–2444, https://doi.org/10.1175/WAF-D-20-0094.1.

Brown, B., and Coauthors, 2021: The Model Evaluation Tools (MET): More than a decade of community-supported forecast verification. *Bull. Amer. Meteor. Soc.*, **102**, E782–E807, https://doi.org/10.1175/BAMS-D-19-0093.1.

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

Chen, F., and J. Dudhia, 2001: Coupling an Advanced Land Surface–Hydrology Model with the Penn State–NCAR MM5 Modeling System. Part I: Model Implementation and Sensitivity. *Mon. Wea. Rev.*, **129 (4)**, 569–585, https://doi.org/10.1175/1520-0493(2001)129⟨0569:CAALSH⟩2.0.CO;2.

Clark, A., J. Kain, P. Marsh, J. C. Jr., M. Xue, and F. Kong, 2012a: Forecasting tornado path lengths using a three-dimensional object identification algorithm applied to convection-allowing forecasts. *Wea. Forecasting*, **27**, 1090–1113.

Clark, A. J., and Coauthors, 2012b: An overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bull. Am. Meteorol. Soc.*, **93**, 55–74, https://doi.org/10.1175/BAMS-D-11-00040.1.

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.*, *In Review*.

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.

Davis, C., B. Brown, and R. Bullock, 2006b: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. *Mon. Wea. Rev.*, **134**, 1785–1795.

Dowell, D., 2020: HRRR Data-Assimilation System (HRRRDAS) and HRRRE Forecasts. Tech. rep., NOAA/ESRL/GSL, 8 pp. URL https://rapidrefresh.noaa.gov/internal/pdfs/2020_Spring_Experiment_HRRRE_Documentation.pdf.

Faust, E., M. Bove, and A. Radler, 2021: Thundestorms, hail and tornadoes: Localised but extremely destructive. Accessed 10 Jan 2021, https://www.munichre.com/en/risks/natural-disasters-losses-are-trending-upwards/thunderstorms-hail-and-tornados.html.

Flora, M. L., C. K. Potvin, P. S. Skinner, S. Handler, and A. McGovern, 2021: Using machine learning to generate storm-scale probabilistic guidance of severe weather hazards in the Warn-on-Forecast system. *Mon. Wea. Rev.*, **149**, 1535–1557, https://doi.org/10.1175/MWR-D-20-0194.1.

Gagne, D., S. Haupt, D. Nychka, and G. Thompson, 2019: Interpretable deep learning for spatial analysis of severe hailstorms. *Mon. Wea. Rev*, **147**, 2827–2845.

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.

Gallo, B. T., and Coauthors, 2017a: Breaking new ground in severe weather prediction: The 2015 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Wea. Forecasting*, **32**, 1541–1568.

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

Gallo, B. T., and Coauthors, 2021: Exploring Convection-Allowing Model Evaluation Strategies for Severe Local Storms Using the Finite-Volume Cubed-Sphere (FV3) Model Core. *Wea. Forecasting*, **36**, 3–19, https://doi.org/10.1175/WAF-D-20-0090.1.

Han, J., M. L. Witek, J. Teixeira, R. Sun, H.-L. Pan, J. K. Fletcher, and C. S. Bretherton, 2016: Implementation in the NCEP GFS of a Hybrid Eddy-Diffusivity Mass-Flux (EDMF) Boundary Layer Parameterization with Dissipative Heating and Modified Stable Boundary Layer Mixing. *Wea. Forecasting*, **31**, 341–352, https://doi.org/10.1175/WAF-D-15-0053.1.

Harris, L. M., S. L. Rees, M. Morin, L. Zhou, and W. F. Stern, 2019: Explicit prediction of continental convection in a skillful variable-resolution global model. *J. Adv. Modeling Earth Systems*, **11**, 1847–1869, https://doi.org/10.1029/2018MS001542, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2018MS001542.

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.

Iacono, M. J., J. S. Delamere, E. J. Mlawer, M. W. Shephard, S. A. Clough, and W. D. Collins, 2008: Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *J. Geophys. Res.*, **113**, D13103, https://doi.org/10.1029/2008JD009944.

Kain, J. S., P. R. Janish, S. J. Weiss, M. E. Baldwin, R. S. Schneider, and H. E. Brooks, 2003: Collaboration between forecasters and research scientists at the NSSL and SPC: The Spring Program. *Bull. Am. Meteorol. Soc.*, **84**, 1797–1806, https://doi.org/10.1175/BAMS-84-12-1797.

Kalina, E. A., I. Jankov, T. Alcott, J. Olson, J. Beck, J. Berner, D. Dowell, and C. Alexander, 2021: A progress report on the development of the High-Resolution Rapid Refresh Ensemble. *Wea. Forecasting*, **36 (3)**, 791–804, https://doi.org/10.1175/WAF-D-20-0098.1.

Krocak, M. J., and H. E. Brooks, 2020: An analysis of subdaily severe thunderstorm probabilities for the United States. *Wea. Forecasting*, **35**, 107–112.

Kumjian, M. R., and K. Lombardo, 2020: A hail growth trajectory model for exploring the environmental controls on hail size: model physics and idealized tests. *J. Atmos. Sci.*, **77 (8)**, 2765–2791, https://doi.org/10.1175/JAS-D-20-0016.1.

Kumjian, M. R., K. Lombardo, and S. Loeffler, 2021: The evolution of hail production in simulated supercell storms. *J. Atmos. Sci.*, **78 (11)**, 3417–3440, https://doi.org/10.1175/JAS-D-21-0034.1.

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 (5)**, 802–823.

Lin, Y., and M. R. Kumjian, 2022: Influences of CAPE on hail production in simulated supercell storms. *J. Atmos. Sci.*, **79 (1)**, 179–204, https://doi.org/10.1175/JAS-D-21-0054.1.

Long, P., 1986: An economical and compatible scheme for parameterizing the stable surface layer in the medium range forecast model. URL http://www.lib.ncep.noaa.gov/ncepofficenotes/files/01408602.pdf, nCEP Office Note 321, 24 pp.

Long, P., 1989: Derivation and suggested method of the application of simplified relations for surface fluxes in the medium-range forecast model: Unstable case. URL http://www.lib.ncep.noaa.gov/ncepofficenotes/files/0140893E.pdf, nCEP Office Note 356, 53 pp.

Mansell, E. R., C. L. Ziegler, and E. C. Bruning, 2010: Simulated Electrification of a Small Thunderstorm with Two-Moment Bulk Microphysics. *J. Atmos. Sci.*, **67 (1)**, 171–194, https://doi.org/10.1175/2009JAS2965.1.

Marsh, P. T., J. S. Kain, V. Lakshmanan, A. J. Clark, N. M. Hitchens, and J. Hardy, 2012: A method for calibrating deterministic forecasts of rare events. *Wea. Forecasting*, **27 (2)**, 531–538, https://doi.org/10.1175/WAF-D-11-00074.1.

Milbrandt, J. A., and M. K. Yau, 2006: A multimoment bulk microphysics parameterization. Part III: Control simulation of a hailstorm. *J. Atmos. Sci.*, **63**, 3114–3136, https://doi.org/10.1175/JAS3816.1.

Miller, W. J. S., and Coauthors, 2021: Exploring the usefulness of downscaling free forecasts from the Warn-on-Forecast system. *Wea. Forecasting*, **-1**, https://doi.org/10.1175/WAF-D-21-0079.1.

Morrison, H., G. Thompson, and V. Tatarskii, 1997: Impact of cloud microphysics on the development of trailing stratiform precipitation in a simulated squall line: Comparison of one- and two-moment schemes. *Mon. Wea. Rev.*, **137**, 991–1007.

Murillo, E. M., and C. R. Homeyer, 2019: Severe Hail Fall and Hailstorm Detection Using Remote Sensing Observations. *Journal of Applied Meteorology and Climatology*, **58 (5)**, 947–970, https://doi.org/10.1175/JAMC-D-18-0247.1.

Murillo, E. M., C. R. Homeyer, and J. T. Allen, 2021: A 23-Year Severe Hail Climatology using GridRad MESH Observations. *Mon. Wea. Rev.*, **-1**, https://doi.org/10.1175/MWR-D-20-0178.1.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8 (2)**, 281–293, https://doi.org/10.1175/1520-0434(1993)008⟨0281:WIAGFA⟩2.0.CO;2.

Nakanishi, M., and H. Niino, 2009: Development of an improved turbulence closure model for the atmospheric boundary layer. *Journal of the Meteorological Society of Japan*, **87**, 895–912, https://doi.org/10.2151/jmsj.87.895.

Olson, J., J. Kenyon, W. Angevine, J. Brown, M. Pagowski, and K. Sušelj, 2019: 61. A description of the MYNN–EDMF scheme and coupling to other components in wrf-arw. 42 pp.

Olson, J. B., T. Smirnova, J. S. Kenyon, D. D. Turner, J. M. Brown, W. Zheng, and B. W. Green, 2021: 67. A Description of the MYNN Surface-Layer Scheme. URL https://repository.library.noaa.gov/view/noaa/30605, 26 pp.

Ortega, K., 2018: Evaluating multi-radar, multi-sensor products for surface hailfall diagnosis. *E-Journal of Severe Storms Meteorology*, **13 (1)**.

Potvin, C. K., and Coauthors, 2020: Assessing systematic impacts of PBL schemes on storm evolution in the NOAA Warn-on-Forecast system. *Mon. Wea. Rev.*, **148**, 2567–2590, https://doi.org/10.1175/MWR-D-19-0389.1.

Putman, W. M., and S.-J. Lin, 2007: Finite-volume transport on various cubed-sphere grids. *Journal of Computational Physics*, **227 (1)**, 55–78, https://doi.org/10.1016/j.jcp.2007.07.022.

Roberts, B., A. J. Clark., B. T. Gallo, I. Jirak, C. Schwartz, and K. Knopfmeier, 2022: Sensitivity of model performance to driving model vs. model core during the 2020 NOAA HWT Spring Forecasting Experiment. *Wea. Forecasting*, *in review*.

Roberts, B., B. T. Gallo, I. L. Jirak, A. J. Clark, D. C. Dowell, X. Wang, and Y. Wang, 2020: What does a convection-allowing ensemble of opportunity buy us in forecasting thunderstorms? *Wea. Forecasting*, **35**, https://doi.org/10.1175/WAF-D-20-0069.1.

Schwartz, C. S., and R. A. Sobash, 2017: Generating probabilistic forecasts from convection-allowign ensebmles using neighborhood approaches: A review and recommendations. *Mon. Wea. Rev.*, **145**, 3397–3418.

Shedd, L., M. R. Kumjian, I. Giammanco, T. Brown-Giammanco, and B. R. Maiden, 2021: Hailstone Shapes. *Journal of the Atmospheric Sciences*, **78**, 639–652, https://doi.org/10.1175/JAS-D-20-0250.1.

Skinner, P. S., and Coauthors, 2018: Object-based verification of a prototype Warn-on-Forecast system. *Wea. Forecasting*, **33**, 1225–1250, https://doi.org/10.1175/WAF-D-18-0020.1.

Smirnova, T. G., J. M. Brown, S. G. Benjamin, and J. S. Kenyon, 2016: Modifications to the Rapid Update Cycle Land Surface Model (RUC LSM) Available in the Weather Research and Forecasting (WRF) Model. *Mon. Wea. Rev.*, **144**, 1851–1865, https://doi.org/10.1175/MWR-D-15-0198.1.

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.

Snook, N., F. Kong, K. A. Brewster, M. Xue, K. W. Thomas, T. A. Supinie, S. Perfater, and B. Albright, 2019: Evaluation of convection-permitting precipitation forecast products using WRF, NMMB, and FV3 for the 2016–17 NOAA Hydrometeorology Testbed Flash Flood and Intense Rainfall Experiments. *Wea. Forecasting*, **34**, 781–804, https://doi.org/10.1175/WAF-D-18-0155.1.

Sobash, R., C. Schwartz, G. Romine, K. Fossell, and M. Weisman, 2016: Severe weather prediction using storm surrogates from an ensemble forecasting system. *Wea. Forecasting*, **31**, 255–271.

Thompson, G., and T. Eidhammer, 2014: A Study of Aerosol Impacts on Clouds and Precipitation Development in a Large Winter Cyclone. *Journal of the Atmospheric Sciences*, **71 (10)**, 3636–3658, https://doi.org/10.1175/JAS-D-13-0305.1.

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.

Wheatley, D. M., K. H. Knopfmeier, T. A. Jones, and G. J. Creager, 2015: Storm-scale data assimilation and ensemble forecasting with the NSSL experimental Warn-on-Forecast System. Part I: Radar data experiments. *Wea. Forecasting*, **30 (6)**, 1795–1817, https://doi.org/10.1175/WAF-D-15-0043.1.

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*, **413**, 286–303.

Zhang, C., and Coauthors, 2019: How well does an FV3-based model predict precipitation at a convection-allowing resolution? Results from CAPS forecasts for the 2018 NOAA Hazardous Weather Test bed with different physics combinations. *Geophys. Res. Lett.*, **46**, 3523–3531, https://doi.org/10.1029/2018GL081702.

Zhang, M., G. Firl, L. Bernardet, and V. Kunkel, 2018: Scientific and technical documentation for parameterizations in the Common Community Physics Package (CCPP). Amer. Meteor. Soc, Denver, CO, URL https://ams.confex.com/ams/29WAF25NWP/webprogram/Paper345517.html.

Zhou, L., S.-J. Lin, J.-H. Chen, L. M. Harris, X. Chen, and S. L. Rees, 2019: Toward convective-scale prediction within the Next Generation Global Prediction System. *Bulletin of the American Meteorological Society*, **100**, 1225–1243, https://doi.org/10.1175/BAMS-D-17-0246.1.