Exploring the Usefulness of Downscaling Free Forecasts from the Warn-on-Forecast System

WILLIAM J. S. MILLER,[a,b] COREY K. POTVIN,[b,c] MONTGOMERY L. FLORA,[a,b] BURKELY T. GALLO,[a,d] LOUIS J. WICKER,[b] THOMAS A. JONES,[b,a] PATRICK S. SKINNER,[b,a] BRIAN C. MATILLA,[b,a] AND KENT H. KNOPFMEIER[b,a]

[a] *Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma*
[b] *NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma*
[c] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*
[d] *NOAA/NCEP/Storm Prediction Center, Norman, Oklahoma*

(Manuscript received 19 May 2021, in final form 20 November 2021)

ABSTRACT: The National Severe Storms Laboratory (NSSL) Warn-on-Forecast System (WoFS) is an experimental real-time rapidly updating convection-allowing ensemble that provides probabilistic short-term thunderstorm forecasts. This study evaluates the impacts of reducing the forecast model horizontal grid spacing Δx from 3 to 1.5 km on the WoFS deterministic and probabilistic forecast skill, using 11 case days selected from the 2020 NOAA Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (SFE). Verification methods include (i) subjective forecaster impressions; (ii) a deterministic object-based technique that identifies forecast reflectivity and rotation track storm objects as contiguous local maxima in the composite reflectivity and updraft helicity fields, respectively, and matches them to observed storm objects; and (iii) a recently developed algorithm that matches observed mesocyclones to mesocyclone probability swath objects constructed from the full ensemble of rotation track objects. Reducing Δx fails to systematically improve deterministic skill in forecasting reflectivity object occurrence, as measured by critical success index ($\mathrm{CSI}_{\mathrm{DET}}$), a metric that incorporates both probability of detection ($\mathrm{POD}_{\mathrm{DET}}$) and false alarm ratio ($\mathrm{FAR}_{\mathrm{DET}}$). However, compared to the $\Delta x = 3$ km configuration, the $\Delta x = 1.5$ km WoFS shows improved midlevel mesocyclone detection, as evidenced by its statistically significant (i) higher $\mathrm{CSI}_{\mathrm{DET}}$ for deterministic midlevel rotation track objects and (ii) higher normalized area under the performance diagram curve (NAUPDC) score for probability swath objects. Comparison between $\Delta x = 3$ km and $\Delta x = 1.5$ km reflectivity object properties reveals that the latter have 30% stronger mean updraft speeds, 17% stronger median 80-m winds, 67% larger median hail diameter, and 28% higher median near-storm-maximum 0–3-km storm-relative helicity.

KEYWORDS: Forecast verification/skill; Numerical weather prediction/forecasting; Probability forecasts/models/distribution; Short-range prediction; Mesoscale models; Model errors; Model evaluation/performance; Regional models

1. Introduction

As we work toward improving operational convection-allowing model (CAM) thunderstorm forecasts, strategic decisions must be made over how to allocate additional computing power that becomes available. Decreasing CAM horizontal grid spacing (Δx)[1] is one option to improve the resolution of convective motions. Using a horizontal grid spacing of $\Delta x \sim 4$ km is generally thought to be the coarsest permissible resolution needed for explicitly resolving clouds and larger-scale convective overturning motions (Weisman et al. 1997). In recent years CAMs have become routinely used in operational numerical weather prediction (NWP) to forecast warm season convective weather events, and their improved skill compared to that of coarser models has been well documented. For example, real-time forecasts using the Advanced Research version of the Weather Research and Forecasting Model (WRF-ARW; Skamarock et al. 2008) better represented the structure and convective mode of mesoscale convective systems when Δx was decreased from ~10 to 4 km and convective parameterizations were turned off (Done et al. 2004; Weisman et al. 2008). However, increasing model horizontal resolution carries a steep computational cost: reducing Δx by a factor of ½ requires a factor of 8 increase in computer processing power, even while not changing the vertical grid spacing. Therefore, the forecast skill gained from decreasing horizontal grid spacing must be carefully weighed against other potentially beneficial utilizations of increased computing power, such as improved physics scheme complexity or a larger ensemble size.

Several previous studies have examined the sensitivity of CAM-simulated convective structure and updraft intensity to variations in Δx over the sub 1–4-km range. Bryan et al. (2003) found significant differences in convective overturning patterns, cloud depth, system phase speed, and storm cell size in a simulated squall line when Δx was varied between 100 m and 1 km. They concluded that although only models with $\Delta x \leq 100$ m could accurately *resolve* turbulent convective flows, those with $\Delta x \sim 1$ km could still capture basic elements of squall line structure and provide value to operational forecasters. When increasing a simulated squall line's Δx from 1

[1] Hereafter, the use of "Δx" represents the horizontal grid spacing in "x" and "y" directions.

Corresponding author: William Miller, wmiller1@umd.edu

DOI: 10.1175/WAF-D-21-0079.1

to 4 km, Bryan and Morrison (2012) found that updraft speed decreased; similar results have been reported for simulated supercells when Δx was increased from 1 to 2 km (Adlerman and Droegemeier 2002; Noda and Niino 2003; Potvin and Flora 2015). These findings are consistent with the expectation that as Δx increases, nonhydrostatic accelerations should weaken in the wider resolved updrafts (Markowski and Richardson 2010) and spatial filtering should more substantially dampen local w maxima. On the other hand, Bryan and Morrison (2012) also showed how a larger Δx could reduce turbulent mixing and environmental air entrainment rates, thereby increasing cloud latent heat release and updraft accelerations. Additionally, it is worth noting that the $O(1)$-km length scale is a "gray zone" (termed the terra incognita by Wyngaard 2004) that falls in between the $O(10)$-km length scale for which many planetary boundary layer (PBL) physics schemes used in the current generation of CAMs have been "tuned" (Shin and Hong 2015) and the $O(100)$-m scales needed for running large eddy simulations. Given the important role that boundary layer processes play in the initiation and subsequent evolution of deep moist convection, it is possible that operational CAMs may suffer degraded performance as Δx approaches ~1 km unless some PBL scheme parameters are further tuned or made scale-aware (Bryan et al. 2003; Shin and Hong 2015). For example, Verrelle et al. (2015) found that simulated supercell updraft intensity *increased* with increasing Δx over the 0.5–2-km range. They attributed this trend to reduced entrainment on the coarser grids, which resulted from underproduction of subgrid-scale turbulent kinetic energy by the PBL physics scheme.

Keeping the above considerations in mind, we now turn our attention to the National Severe Storms Laboratory (NSSL) Experimental Warn-on-Forecast System (WoFS), a CAM ensemble that provides probabilistic forecast guidance on short $O(0–3)$-h time scales.[2] Development of WoFS is motivated by the National Oceanic and Atmospheric Administration (NOAA)'s strategic goal of increasing National Weather Service (NWS)-issued warning lead times for severe convective hazards, including tornadoes, damaging thunderstorm downburst winds, hail, and flash flooding (Stensrud et al. 2009, 2013). For example, the average lead time for verified NWS tornado warnings in 2015 was 8 min (Brooks and Correia 2018), clearly suboptimal for the protection of human lives. Ultimately, WoFS could provide forecasters responsible for issuing tornado, severe thunderstorm, and flash flood warnings with another decision-making tool that complements radar-derived products and storm spotter reports.

WoFS, formerly named the NSSL Experimental Warn-on-Forecast System for ensembles (NEWS-e; Wheatley et al. 2015), has been used experimentally since 2016 to generate real-time forecasts over a regional domain covering a portion of the continental United States where hazardous convective weather is expected to occur. Skinner et al. (2018, hereafter S18) applied a deterministic object-based verification methodology to real-time 2016 and 2017 WoFS forecasts, and they found that the WoFS can forecast thunderstorm occurrence at 0–3-h lead times with a reasonably high skill. WoFS has also demonstrated skill in its short-term prediction of thunderstorm flash flooding (Yussouf and Knopfmeier 2019) as well as the localized tornado, wind and rain hazards associated with landfalling tropical cyclones (Jones et al. 2019; Yussouf et al. 2020).

Due to computational limitations, WoFS is developed using $\Delta x = 3$ km. Although 3-km horizontal grid spacing is far too coarse for resolving tornadoes, Potvin and Flora (2015) showed that idealized $\Delta x = 3$ km simulations could capture low-level mesocyclone tracks reasonably well. Their finding is encouraging, given that (i) mesocyclones are a necessary precursor for tornadogenesis in supercell thunderstorms (Markowski and Richardson 2010); and that (ii) supercells spawn the majority of deadly U.S. tornadoes (Schoen and Ashley 2011). Although only ~25% of all mesocyclones detected by Doppler radar produce a tornado, the probability of tornado association increases substantially to ~40% for low-level mesocyclones, i.e., those with cloud bases detected below 1 km above ground level (AGL; Trapp et al. 2005). Furthermore, recent modeling studies have shown that tornadogenesis likelihood is well correlated with low-level mesocyclone strength (e.g., Mashiko 2016a,b; Roberts et al. 2016; Yokota et al. 2018). Updraft helicity (UH; Kain et al. 2008)—the height integral of upward vertical velocity multiplied by the vertical component of relative vorticity ζ—can serve as a proxy variable for detecting mesocyclones in model output. S18 found that WoFS tended to over-forecast rotational storm objects defined using 2–5 km AGL layer UH (hereafter $UH_{2–5}$). Perhaps more significantly, they reported little difference between the WoFS skill in forecasting the occurrence of 0–2 km AGL layer UH (hereafter $UH_{0–2}$) objects and that for $UH_{2–5}$ objects, which implies that the $\Delta x = 3$ km WoFS may struggle to resolve the storm-scale processes that control low-level mesocyclone development and dissipation. Lawson et al. (2021, hereafter L21) found that reducing the WoFS Δx to 1 km improved detection of the most intense storms, i.e., those with high radar reflectivity or strong low- to midlevel rotation, when using a novel probabilistic object-based information gain metric that rewarded successful prediction of rare events. Their study was limited to four spring 2019 cases characterized by environments with high vertical wind shear (VWS) and low convective available potential energy (CAPE). Therefore, further studies are needed to better understand the sensitivity of WoFS forecast skill to Δx.

In this study, we evaluate the impacts of reducing Δx from 3 to 1.5 km on the WoFS short-range (i.e., 0–3 h) skill in forecasting the occurrence of mesocyclones and severe thunderstorms. We select quasi-operational WoFS forecasts generated for 11 case days in support of the 2020 NOAA Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (SFE; Clark et al. 2012; Gallo et al. 2017; Clark et al. 2021). Starting from a set of $\Delta x = 3$ km WoFS analyses, $\Delta x = 3$ km [hereafter, denoted as "Realtime") WoFS free forecasts are compared to downscaled (hereafter, denoted as "HIRES") free forecasts that activate a $\Delta x = 1.5$ km nest inside of the parent Realtime

[2] Although the WoFS mission primarily focuses on 0–3-h forecasts, the experimental WoFS is now routinely run for a 6-h period.

TABLE 1. Summary of the 2020 severe weather case days analyzed for this study. For each case, the period covered by WoFS forecasts, maximum SPC risk level from that day's 1630 UTC outlook within the WoFS real-time domain, number of SPC archived tornado reports within the real-time domain over the forecast period, primary states affected, primary storm mode, and peak WoFS-forecast mixed-layer CAPE in regions impacted by severe weather are given. Note that the 3 Mar early morning event is listed using the previous day because its associated WoFS forecast period occurred during the 2 Mar SPC outlook period. Forecasts from bolded case days were subjectively evaluated by Spring SFE participants.

Day	Forecast period (UTC)	SPC risk level	No. of tornado reports	Primary states impacted	Primary storm mode	MLCAPE (J kg^{-1})
2 Mar	0000–0800	Slight	18	KY, MO, TN	Supercell	~1000
28 Apr	2000–0400	Moderate	3	AR, MO, OK, TX	Linear	>3000
4 May	2000–0400	Enhanced	1	AR, KS, MO, OK	Mixed	>3000
7 May	2000–0400	Slight	0	OK, TX	Mixed	2000–3000
13 May	2000–0400	Enhanced	1	OK, TX	Mixed	3000
15 May	2000–0400	Enhanced	1	OK, TX	Linear	3000
20 May	2000–0400	Enhanced	0	MT, WY	Linear	1000–2000
22 May	2000–0400	Enhanced	8	AR, OK, TX	Mixed	>3000
26 May	2000–0400	Slight	6	IA, IL, MN, WI	Mixed	1000–2000
27 May	2000–0400	Enhanced	1	TX	Mixed	2000–3000
29 May	2000–0400	Slight	1	MD, NY, PA, VT	Mixed	1000–2000

grid. Our term "free forecasts" emphasizes the fact that the model integration is not interrupted with any data assimilation update cycles. We evaluate WoFS forecast skill separately in *deterministic* and *probabilistic* frameworks using object-based methods developed by S18 and Flora et al. (2019, hereafter F19), respectively. The deterministic verification method, applied independently to each ensemble member, defines forecast storm objects as discrete contiguous regions where a proxy variable for storm intensity exceeds a predetermined threshold and matches them to observed objects. By contrast, the probabilistic method uses the full ensemble to generate forecast probability swath objects, each of which represents the likelihood that a given storm event produces a mesocyclone, and matches them to observed storm objects. Our study addresses the following questions. First, to what extent (if at all), does the HIRES WoFS improve deterministic forecasts of severe thunderstorm and mesocyclone occurrence, relative to the Realtime WoFS? Second, does reducing the WoFS Δx to 1.5 km improve probabilistic skill of mesocyclone forecasts? And finally, how sensitive are WoFS-forecast storm-scale processes to doubling the horizontal grid resolution?

The remainder of this paper is organized as follows. The next section describes the 11 case days, model configuration, observations, and verification methods. Section 3 presents our results comparing the Realtime and HIRES WoFS forecasts in terms of subjective forecaster impressions, object-based deterministic and probabilistic skill verification, and model representations of storm-scale processes. A summary and conclusions are given in the final section.

2. Datasets and methodology

a. Summary of cases

Table 1 summarizes the 11 spring 2020 severe thunderstorm case days analyzed herein. The most destructive of these events occurred overnight on 2–3 March, when a long-tracked supercell developed in northwestern Tennessee around 0400 UTC just south of a mesoscale convective system (MCS) that had been tracking northeastward along a stationary front. This supercell yielded several significant tornadoes as it moved eastward across northern Tennessee over the next five hours, including an enhanced Fujita scale (EF) 3 that killed 5 people (220 injured) in the Nashville metropolitan area, as well as a violent EF4 that killed 19 people (87 injured) near Cookeville (NOAA/NCEI 2021). The Tennessee supercell developed while a 700-hPa shortwave moved overhead; the latter helped to enhance low-level vertical wind shear, enabling 0–1-km storm-relative helicity (SRH$_{0\text{–}1}$) to locally exceed 400 m^2 s^{-2}. Warm-sector mixed-layer CAPE of ~1000 J kg^{-1} resulted from the advection of an elevated mixed layer by southwesterly winds ahead of an upper-level trough anchored over the Central Plains atop seasonally high low-level moisture that the MCS and its associated surface low helped to converge along the warm front.

Otherwise, spring 2020 was rather notable for its small number of Great Plains tornadoes associated with discrete supercells. The remaining ten cases selected for this study generally have linear or mixed-mode dominant convective patterns. Among them, 28 April was the most impactful. On that day, an intense squall line developed southwestward along a cold front in central and eastern Oklahoma ahead of an upper-level trough, and it surged southeastward, generating widespread damaging winds and a few tornado reports. Other notable cases include 7 May, which featured an intense right-moving long-tracked supercell over the Texas panhandle that produced very large hail, and 22 May, when several tornadic supercells explosively developed in a high-CAPE environment near an outflow boundary draped through the Red River valley.

b. Model configuration

The Realtime WoFS configuration used for the 2020 HWT SFE consists of a 36-member WRF-ARW version 3.9.1

(Skamarock et al. 2008) ensemble that is cycled every 15 min from 1500 to 0300 UTC the following day.[3] The model uses a 3-km Δx and 51 vertical levels. The High-Resolution Rapid Refresh Ensemble (HRRRE; Dowell et al. 2016) provides initial and lateral boundary conditions to the fixed 900 km × 900 km WoFS domain; the latter's daily position is chosen in collaboration with the Spring Forecasting Experiment. The Gridpoint Statistical Interpolation (GSI) ensemble Kalman filter (EnKF) assimilates the following observation types: conventional; Multi-Radar Multi-Sensor (MRMS) reflectivity and radial velocity; cloud water path retrievals, atmospheric motion vectors, and clear-sky radiances from the *GOES-16* imager; and any available Oklahoma Mesonet observations. PBL and radiation physics schemes are varied among the WoFS members to help increase ensemble spread and improve reliability, given the tendency for small EnKF ensembles to be underdispersive compared to the meteorological "errors of the day" (Houtekamer and Zhang 2016). Each member uses one of six unique parameterization combinations, which feature either the Yonsei University (YSU; Hong et al. 2006), Mellor–Yamada–Janjić (MYJ; Mellor and Yamada 1982; Janjić 2002), or Mellor–Yamada–Nakanishi–-Niino (MYNN; Nakanishi and Niino 2004, 2006) PBL schemes and either (i) the Dudhia longwave and Rapid Radiative Transfer Model (RRTM) shortwave schemes, or (ii) the Rapid Radiative Transfer Model-Global (RRTMG) longwave and shortwave schemes; see S18 Table 1 for further details. All members use the NSSL two-moment cloud microphysics scheme (Mansell et al. 2010) and the RUC land surface model (Smirnova et al. 2016). The reader can find additional details of the WoFS in Wheatley et al. (2015), Jones et al. (2016), and Jones et al. (2020).

For each hour (half hour) beginning at 1700 UTC and running through 0300 UTC the next day, 6-h (3-h) Realtime WoFS free forecasts are initialized from cycled analyses members 1–18. Additionally, 3-h HIRES free forecasts are initialized hourly over the 2000–0100 UTC period from cycled analysis members 1–9. The 9-member HIRES configuration activates a fixed 690 km × 690 km nest with $\Delta x = 1.5$ km inside of the Realtime analysis domain at $t = 0$ h and two-way nesting is used through the entire forecast period; all other HIRES model settings—including the vertical level configuration—are identical to the Realtime settings. WoFS forecast output for all 2020 SFE cases is available for public display at https://wof.nssl.noaa.gov/realtime. Examples of HIRES output fields will be shown in section 3b. For the objective analysis methods described in sections 3b–d herein, we only consider Realtime forecasts that can be directly compared to a HIRES forecast initialized from the same WoFS analysis, namely, 3-h forecasts initialized hourly between 2000 and 0100 UTC, members 1–9.

c. Subjective evaluation criteria

Subjective impressions from participants in the 2020 SFE were collected during next day evaluations, where participants were asked for feedback on the Realtime and HIRES ensemble forecasts of UH. Specifically, participants were asked to rate three initializations of each ensemble (2000, 2200, and 0000 UTC) with the following question: "Please rate the performance of the following ensembles initialized at XXXX UTC on a scale of 1 (Very Poor) to 10 (Very Good) using the hourly products (labeled 1-h in the dropdown product selection menu). Consider the ability of the ensemble to provide useful guidance to a forecaster trying to issue a forecast of severe convective storms." Then, participants were asked specifically about different aspects of the forecast that pertain to operational forecaster concerns: convective mode and initiation. Finally, participants were asked at each initialization time whether the HIRES ensemble provided additional useful information compared to the Realtime ensemble. For verification, available local storm reports (LSRs) were overlaid on the forecast output. Since the evaluations took place the day after event occurrence, the dataset of LSRs available to participants was incomplete as reports frequently take a week or two after an event to be fully compiled. However, both ensembles were evaluated against the same set of reports. See Clark et al. (2021) for further details.

Two other important differences to note between the subjective and objective analyses completed herein pertain to the case list and the number of ensemble members used to generate the probabilistic fields that are being evaluated. While the objective analyses matched the members between the ensembles (i.e., a subset of 9 Realtime members was compared to the HIRES members), the subjective evaluation focused more on the question of whether a smaller ensemble with finer grid spacing could provide information beyond what a larger ensemble with coarser grid spacing could. As such, the SFE participants' subjective analyses compare the full 18-member Realtime ensemble to the 9-member HIRES ensemble. Finally, a smaller difference is in the number of cases. As mentioned previously, the 2 March case was run retrospectively, so SFE participants did not subjectively rate it. For two additional cases (28 April and 29 May), the data were unavailable for participants to make comparisons as part of the SFE. Thus, the subjective evaluation encompasses eight cases (Table 1, bolded) compared to the objective evaluation's 11, with 65 responses from SFE participants across those cases.

d. Verification dataset

The NSSL MRMS gridded dataset (Smith et al. 2016) is used for verifying the location and timing of WoFS-forecast storms. MRMS is a real-time 0.01° × 0.01° analysis on a latitude–longitude grid covering the contiguous United States, updated every 2 min, that composites observations from the WSR-88D Doppler radar network. When considering all thunderstorms, we use MRMS composite reflectivity (REFLCOMP) interpolated to the WoFS grid as a proxy for storm intensity. We separately verify WoFS forecasts of low-level and midlevel mesocyclones using the MRMS azimuthal

[3] The 2 March case was an exception since it was an overnight event that occurred prior to the 2020 SFE period. WoFS was cycled retrospectively starting from the 1200 UTC 2 March HRRRE initial conditions and continuing through 0800 UTC 3 March.

TABLE 2. Object identification variable intensity threshold values used for this study. Realtime and HIRES values are listed as pairs delimited by the "/" symbol, with the latter given in italic font.

Dataset	REFLCOMP (dBZ)	$UH_{2–5}$ ($m^2\ s^{-2}$) or $AWS_{2–5}$ (s^{-1})	$UH_{0–2}$ ($m^2\ s^{-2}$) or $AWS_{0–2}$ (s^{-1})
WoFS YSU	45.9/*48.6*	63.2/*138.2*	15.9/*35.4*
WoFS MYJ	45.5/*48.3*	64.0/*145.0*	15.6/*36.2*
WoFS MYNN	46.0/*48.6*	67.3/*149.7*	16.3/*36.9*
MRMS	39.8/*40.9*	0.0039/*0.0041*	0.0036/*0.0038*

wind shear product computed over the 0–2 km AGL layer (hereafter $AWS_{0–2}$) and 2–5 km AGL layer (hereafter $AWS_{2–5}$), respectively. Following Miller et al. (2013) and S18, we interpolate the MRMS-derived $AWS_{0–2}$ and $AWS_{2–5}$ data to the WoFS grid every 5 min and then construct "rotation tracks" by computing the maximum $AWS_{0–2}$ and $AWS_{2–5}$ at each gridpoint over the previous 30-min period. To mitigate the impacts of spurious noise, extensive quality control is performed when generating the $AWS_{0–2}$ and $AWS_{2–5}$ analyses. For example, only the WSR-88D radial velocities collocated with quality-controlled reflectivity exceeding 20 dBZ are used in the azimuthal wind shear calculation; see S18 for further details. MRMS rotation tracks are compared against 30-min swaths of WoFS-forecast gridpoint maximum $UH_{0–2}$ and $UH_{2–5}$. To simplify the deterministic storm object matching algorithm (described in section 2f), separate MRMS verification datasets are used for Realtime and HIRES forecasts. MRMS fields are interpolated from their native latitude–longitude grid to the Realtime domain and HIRES nest using a Cressman scheme with a 3-km radius of influence. Visual comparison of the $\Delta x = 3$ km and $\Delta x = 1.5$ km interpolated MRMS datasets (not shown) reveals that differences are quite small and unlikely to influence the deterministic object-based Realtime and HIRES WoFS verifications described in section 3b.

e. Deterministic storm object identification

CAM ensemble forecasts of thunderstorms, particularly those in cellular and mixed convective modes, present a unique verification challenge due to the large volume of output data and the large number of features needing to be tracked. While subjective human evaluation will always have value, it should be complemented with objective verification methods that can provide quantitative analysis of model errors and their statistical significance. Traditional point-to-point verification techniques, such as a root-mean-square error (RMSE) computed over a forecast domain, unduly penalize small forecast storm position and/or timing errors even when the model resolves a storm's structure well (Potvin et al. 2017). Furthermore, severe thunderstorms are by nature rare events that typically occupy only a small fraction of a forecast grid.

Therefore, like other recent WoFS forecast verification studies (S18; F19; Potvin et al. 2020; L21), we use an automated object-based verification technique to evaluate the WoFS deterministic and probabilistic forecast skill. We use the Python Scikit-image software (Van der Walt et al. 2014) to identify "storm objects" as contiguous regions on a two-dimensional WoFS forecast or MRMS analysis grid where field values of a variable measuring storm intensity exceed a predetermined threshold. Forecast storm objects can then be matched to observed storm objects. The major advantages of using an object-based framework over traditional gridpoint-based forecast verification techniques include the former's focus on rare but significant events and its tolerance for small, operationally acceptable errors in forecast storm timing and location. However, one disadvantage of object-based verification techniques is their sensitivity to tunable object identification and matching parameters.

Storm objects defined using the composite reflectivity (REFLCOMP) field are treated as proxies for all types of thunderstorms. Our object identification methodology is based upon the assumption that a perfect forecast produces an identical areal footprint in the forecast and verification fields. Model-output reflectivity is sensitive to the microphysics parameterization and other sources of model bias; therefore, REFLCOMP object boundary thresholds are defined separately for Realtime, HIRES, and MRMS datasets. REFLCOMP object thresholds are also defined separately for WoFS members using different PBL physics parameterizations given the WoFS forecast sensitivity to the latter (Potvin et al. 2020). After collecting REFLCOMP values from all eleven 2020 cases (Table 1) at all grid points and 5-min verification times, we set REFLCOMP object boundary thresholds to the 99th percentile value. Midlevel (low-level) rotation track objects boundaries are set to the 99.95th percentile of the 2020 WoFS $UH_{2–5}$ and MRMS $AWS_{2–5}$ ($UH_{0–2}$ and $AWS_{0–2}$) field climatologies, respectively. Our Realtime REFLCOMP, $UH_{2–5}$ and $UH_{0–2}$ object identification thresholds generated from the 2020 seasonal climatology (Table 2) show good agreement with those computed for the 2016 and 2017 WoFS by S18 using a similar method (see their Fig. 3). HIRES REFLCOMP thresholds exceed Realtime REFLCOMP thresholds by ~3 dBZ, perhaps a result of enhanced condensation and freezing in the stronger HIRES updrafts (see section 3d). The stronger HIRES updrafts also partially account for the significantly higher $UH_{0–2}$ and $UH_{2–5}$ thresholds in HIRES.

Quality control checks are performed on all WoFS and MRMS storm objects prior to object matching. Only Realtime and HIRES storm objects found inside of a 20-km-wide buffer zone surrounding the $\Delta x = 1.5$ km nest boundaries are retained. Additional checks are designed to reject spurious MRMS features or excessively small forecast objects not likely to produce severe weather. A minimum area threshold of 144 (100) km^2 is applied to REFLCOMP (rotation track) storm objects, and any storm objects separated by less than 10 km are grouped as a single object. Additionally, rotation track objects are checked to ensure that they are built from 5-min data "elements" that

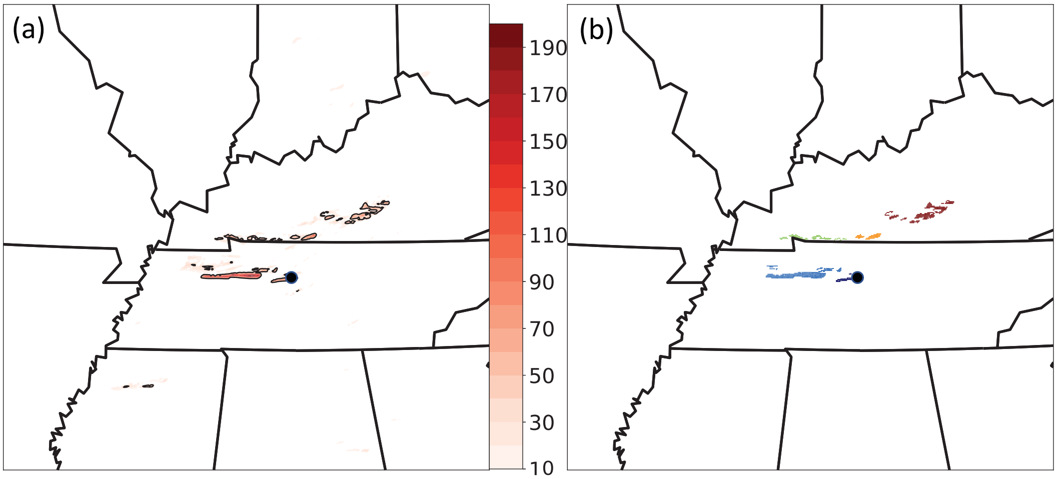

FIG. 1. (a) Maximum UH_{0-2} values ($m^2 s^{-2}$; shaded) taken from the t = 30–60-min forecast period for the HIRES ensemble member 1 forecast initialized at 0500 UTC 3 Mar. Black contours enclose regions where UH_{0-2} exceeds the rotation track object identification threshold (Table 2). (b) Quality-controlled low-level rotation track objects, differentiated by color shading, derived from the UH_{0-2} field shown in (a). Black dot symbols in (a) and (b) denote Nashville's location.

come from at least two consecutive verification times. Any rotation track object composed of at least two elements with eccentricity exceeding 0.9 and length exceeding 25 km is also rejected based on the expectation that it is an advancing linear gust front feature rather than a mesocyclone. Imposing this criterion resulted in removal of numerous 28 April UH_{0-2} objects with linear morphology that were not associated with any tornado reports. Figure 1 shows a set of quality-controlled low-level rotation track objects taken from a 2 March HIRES forecast ensemble member.

f. Deterministic object-based verification

Forecast storm objects are matched to observed storm objects using S18's algorithm, which they adapted from the Method for Object-based Diagnostic Evaluation software tool (MODE; Davis et al. 2006a,b). The algorithm is applied to each ensemble member at every 5-min forecast verification time in a two-step process. First, each observed storm object is compared against all forecast storm objects taken from a ±20-min surrounding time window. For each possible forecast–observed storm object pair, the S18 total interest (S18 TI) score is computed as

$$\text{S18 TI} = \left[\frac{\left(\frac{\text{cd}_{\max} - \text{cd}}{\text{cd}_{\max}}\right) + \left(\frac{\text{md}_{\max} - \text{md}}{\text{md}_{\max}}\right)}{2}\right]\left(\frac{t_{\max} - t}{t_{\max}}\right), \quad (1)$$

where cd is their centroid distance; md is their minimum distance; t is their time difference; and $\text{cd}_{\max}$ = 40 km, $\text{md}_{\max}$ = 40 km, and $t_{\max}$ = 20 min are the maximum allowed centroid displacement, minimum displacement, and time difference, respectively. All object pairs with S18 TI scores exceeding 0.2 are catalogued. If any observed objects are matched to multiple forecast objects, the matched pair with the highest S18 TI score is selected. Each forecast object from within the time window can be matched only once. Out of the total number of observed objects $N_{\text{OBS,TOTAL}}$, the number of matched observed objects $N_{\text{OBS,MATCHED}}$ (i.e., "hits") are tallied; "misses" are the residual. The second step repeats the process, except that now each forecast storm object is compared to all observed storm objects taken from within a ±20-min surrounding time window, yielding the ensemble member's total number of forecast storm objects $N_{\text{FC,TOTAL}}$, number of matched forecast storm objects $N_{\text{FC,MATCHED}}$, and number of "false alarms," equivalent to $N_{\text{FC,TOTAL}} - N_{\text{FC,MATCHED}}$ Deterministic probability of detection (POD_{DET}), false alarm ratio (FAR_{DET}), bias (BIAS_{DET}), and critical success index (CSI_{DET}) scores are then computed as

$$\text{POD}_{\text{DET}} = \frac{N_{\text{OBS,MATCHED}}}{N_{\text{OBS,TOTAL}}}, \quad (2a)$$

$$\text{FAR}_{\text{DET}} = \frac{N_{\text{FC,TOTAL}} - N_{\text{FC,MATCHED}}}{N_{\text{FC,TOTAL}}}, \quad (2b)$$

$$\text{BIAS}_{\text{DET}} = \frac{N_{\text{FC,TOTAL}}}{N_{\text{OBS,TOTAL}}}, \quad (2c)$$

$$\text{CSI}_{\text{DET}} = \frac{N_{\text{FC,MATCHED}}}{N_{\text{FC,TOTAL}} + N_{\text{OBS,TOTAL}} - N_{\text{OBS,MATCHED}}}. \quad (2d)$$

g. Probabilistic object-based verification

Probabilistic CAM ensemble forecast guidance has traditionally been interpreted in terms of the likelihood of an event occurring within a prescribed neighborhood around each gridpoint, commonly referred to as the neighborhood maximum ensemble probability (NMEP; Schwartz and Sobash 2017). "Likelihood" is typically defined by the ensemble probability—

the number of ensemble members forecasting a binary outcome (e.g., did total accumulated rainfall exceed 1 in.?) at a point or within a neighborhood divided by the ensemble size—[Schwartz and Sobash 2017; see their Eq. (4)]. For next-day CAM ensemble output, computing forecast probabilities using this neighborhood-based "spatial" approach and subsequent spatial smoothing helps to correct for (i) expected loss of gridscale accuracy at longer lead times and for (ii) the well-documented problem of poor CAM ensemble reliability, where insufficient ensemble spread yields forecast likelihoods of storm-related events that exceed their conditional event frequencies (F19). However, F19 showed how spatial probabilities may not be appropriate on their own for short-range WoFS forecasts because the spatial smoothing removes finer-scale details that may be of interest to forecasters at 0–3-h lead times. Additionally, F19 pointed out how *forecasting the probability that a given thunderstorm will produce a mesocyclone*, a key WoFS operational objective, is conceptually quite different from forecasting the probability that *a given gridpoint will experience a mesocyclone event*.

Here, we adopt the alternative probabilistic verification approach developed by F19 that is based on "event" probabilities rather than traditional spatial probabilities. Essentially, event probabilities in this study predict the likelihood that a given simulated thunderstorm will produce a mesocyclone within the uncertainty of storm location predicted by the ensemble. Unlike the probabilistic CAM verification methods reviewed by Schwartz and Sobash (2017), this method does not use a prescribed isotropic neighborhood or smoothing technique. Instead, the "neighborhood" is determined anisotropically by the WoFS ensemble spread—more ensemble spread results in a larger region of storm location uncertainty. The trade-off for the event-based method, though, is the inability to account for missed observations where the WoFS does not predict the occurrence of a storm (Flora et al. 2021).

Quality-controlled forecast rotation track objects (i.e., 30-min $\mathrm{UH}_{2\text{–}5}$ and $\mathrm{UH}_{0\text{–}2}$ swaths; section 2e) from individual members are transformed into forecast probability swath objects in a two-step process. First, raw gridscale two-dimensional mesocyclone probability fields are generated for the set of overlapping 30-min forecast periods staggered every 15 min, i.e., t = 0–30 min, t = 15–45 min, t = 30–60 min, …, t = 150–180 min. For each gridpoint i and ensemble member j, binary probabilities BP_{ij} are defined in terms of whether i belongs to the set of gridpoints S_j contained within member j's 30-min rotation track objects:

$$\mathrm{BP}_{ij} = \begin{cases} 1, & \text{if} \quad i \in S_j \\ 0, & \text{if} \quad i \notin S_j \end{cases}. \tag{3}$$

A two-dimensional ensemble probability field EP_i is then defined as the fraction of the N = 9 ensemble members that produce a rotation track overlapping with gridpoint i:

$$\mathrm{EP}_i = \frac{1}{N}\sum_{j=1}^{N}\mathrm{BP}_{ij}\,. \tag{4}$$

Although no further postprocessing is applied to the Realtime EP fields, the HIRES EP fields are "upscaled" to the 3-km grid by applying a NMEP filter with a 3 × 3 gridpoint box; the latter step ensures a fairer comparison between the Realtime and HIRES output, in part because it compensates for the fact that mesocyclone tracks are less likely to overlap a single gridpoint when grid spacing is reduced. Grid staggering effects should be negligible, given that our focus is on contiguous objects rather than on gridpoint-based verification. In this study, we identify the probability objects from the EP field using the two-step watershed algorithm developed in Flora et al. (2021), which is an updated version of the algorithm presented in F19 (see Flora et al. 2021 section 3a for a detailed description of the probability object identification algorithm and its parameter settings). Each probability swath object is assigned a single forecast probability value corresponding to the maximum EP within its boundaries.

To compare Realtime and HIRES WoFS probabilistic mesocyclone forecast skill, sets of $\mathrm{UH}_{2\text{–}5}$ and $\mathrm{UH}_{0\text{–}2}$-derived probability swath objects are matched to their layer-equivalent quality-controlled observed rotation track objects (i.e., 30-min $\mathrm{AWS}_{2\text{–}5}$ or $\mathrm{AWS}_{0\text{–}2}$ swaths; section 2e) as for the deterministic verification (section 2f). However, for the probability objects, $\mathrm{cd}_{\max}$ and $\mathrm{md}_{\max}$ are set to 0 km—thus providing a more conservative probabilistic skill estimate (F19)—and $t_{\max}$ = 15 min. Probability swath object matching applied separately to Realtime and HIRES forecasts yields total numbers of "hits," "misses," and "false alarms" for each probability threshold p; these quantities are then used for computing the probabilistic contingency table metrics $\mathrm{POD}_{\mathrm{PROB}}$, $\mathrm{FAR}_{\mathrm{PROB}}$, $\mathrm{BIAS}_{\mathrm{PROB}}$, and $\mathrm{CSI}_{\mathrm{PROB}}$ via a set of equations analogous to Eqs. (2a)–(2d).

We also test the Realtime and HIRES probabilistic mesocyclone forecasts for reliability, where a forecast ensemble is considered reliable if it generates forecast probabilities that are reasonably consistent with their conditional event frequencies. To compute the reliability of the forecast ensemble probabilities, the probabilities are separated into bins based on their number of ensemble members (e.g., p = 1/9, 2/9, …, 9/9) from which we compute the mean forecast probabilities and conditional event frequencies, where the latter is the fraction of probabilistic objects for a given bin that are matched to an observed swath. For confidence intervals, the set of forecast probabilities and observed data (i.e., whether a forecast probability swath is matched to an observed track) are bootstrapped (N_{boot} = 1000) and the mean forecast probabilities and conditional event frequencies are recomputed. We consider probabilistic mesocyclone forecasts at threshold p to be *reasonably reliable* (*probably unreliable*) if the mean value of the N_{boot}-sized conditional observed frequency distribution falls inside (outside) of a range of "reliable" conditional frequencies known as *consistency bars*, defined using the method of Bröcker and Smith (2007). Consistency bars account for N_p-dependent uncertainties in the conditional observed frequencies computed from the forecast and observation datasets. They span the 2.5%–97.5% quantiles of a *surrogate observed frequency* distribution generated by a second bootstrap resampling (N_{boot} = 1000) of the N_p probability swath objects. For each resampling iteration, an N_p-sized set of surrogate event observations $\hat{Y}_i$, i = 1, …, N_p is generated, using

$$\hat{Y}_i = \begin{cases} 1, & \text{if} \quad Z_i < p \\ 0, & \text{else} \end{cases}, \tag{5}$$

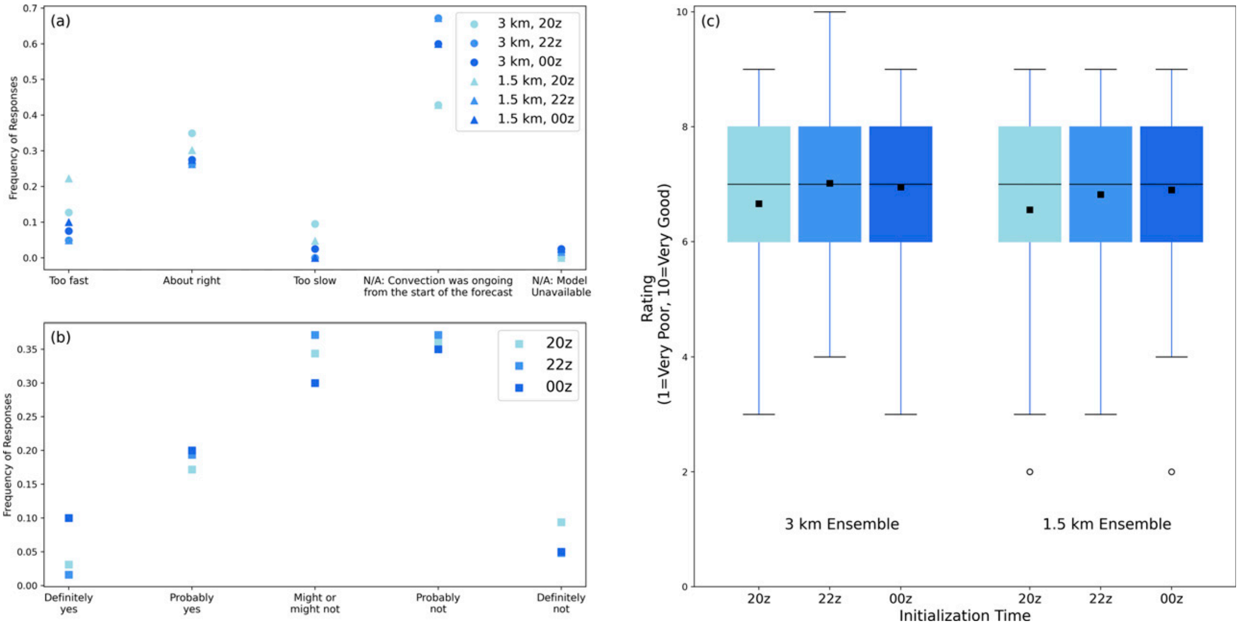

FIG. 2. Results from subjective evaluation conducted by participants in the 2020 SFE, including (a) responses to the question: "How do the following ensembles initialized at XX00 UTC depict convective initiation?"; (b) responses to the question "At XX00 UTC, does the experimental 1.5-km ensemble provide additional useful information compared to the Realtime 3.0-km ensemble?," and (c) ratings of each ensemble's forecast at three different initializations on a scale of 1–10.

where Z_i is a series of uniformly distributed random variables ranging between zero and one, yielding a surrogate observed frequency $\sum_{i=1}^{N_p} \hat{Y}_i/N_p$.

3. Results

a. Subjective forecaster impressions

SFE participant impressions of the HIRES versus Realtime WoFS were mixed. The largest difference between the ensembles from the targeted questions asked regarding convective initiation, timing, and mode, was in the realm of convective initiation. Participants noted that the 1.5-km ensemble at any given initialization tended to initiate convection sooner than the corresponding 3-km initialization, which was more frequently rated "about right" or "too slow" in terms of convective initiation. However, for most cases convection was already ongoing (Fig. 2a). Overall, 1–10 ratings of the forecasts were very similar between the 18-member Realtime and 9-member HIRES ensembles (Fig. 2c), with larger differences occurring between initialization times of the same ensemble compared to between the ensembles at the same initialization time. Thus, it is unsurprising that the participants responded to the question "At XX00 UTC, does the experimental 1.5-km ensemble provide additional useful information compared to the Realtime 3.0-km ensemble?" most frequently with "Might or might not" or "Probably not" (Fig. 2b). However, additional input collected via open-ended comments reveals important differences between the ensembles at specific times. For example, on both 7 and 13 May, participants mentioned that the Realtime ensemble could capture storms that the HIRES missed, and on 26 May a participant noted that the increased structure in the HIRES storms provided better indication of storm severity.

In their open-ended comments, participants also frequently mentioned the higher intensities provided by the HIRES compared to the Realtime ensemble, specifically in maximum $UH_{2–5}$ values and wind speeds, and that they did not know how much of these intensity differences resulted from the different model climatologies arising from the differing grid spacings versus case-dependent differences in the strength of the storms depicted by each ensemble. As such, it will be important in future subjective comparisons to carefully consider the display and contouring of output from ensembles with different grid spacings, particularly since current operational CAMs such as the HREF ensemble and its members use a 3-km grid spacing. Given that forecasters are currently most familiar with the range of UH values produced by a 3–4-km model Δx, they may have difficulty determining what a "high" UH value would be for HIRES output. Using percentile thresholds compared to fixed values and carefully considering color curves will help ensure comparisons that look beyond stronger magnitudes provided by higher horizontal grid resolutions to differences in storm placement, structure, and intensity.

b. Deterministic object-based verification

Figure 3 compares Realtime and HIRES reflectivity object POD_{DET}, FAR_{DET}, CSI_{DET}, and $BIAS_{DET}$ time series, where thin "spaghetti" lines show data for a single ensemble member averaged over all 11 cases. In general, Realtime and HIRES reflectivity object POD_{DET}, FAR_{DET}, and CSI_{DET}

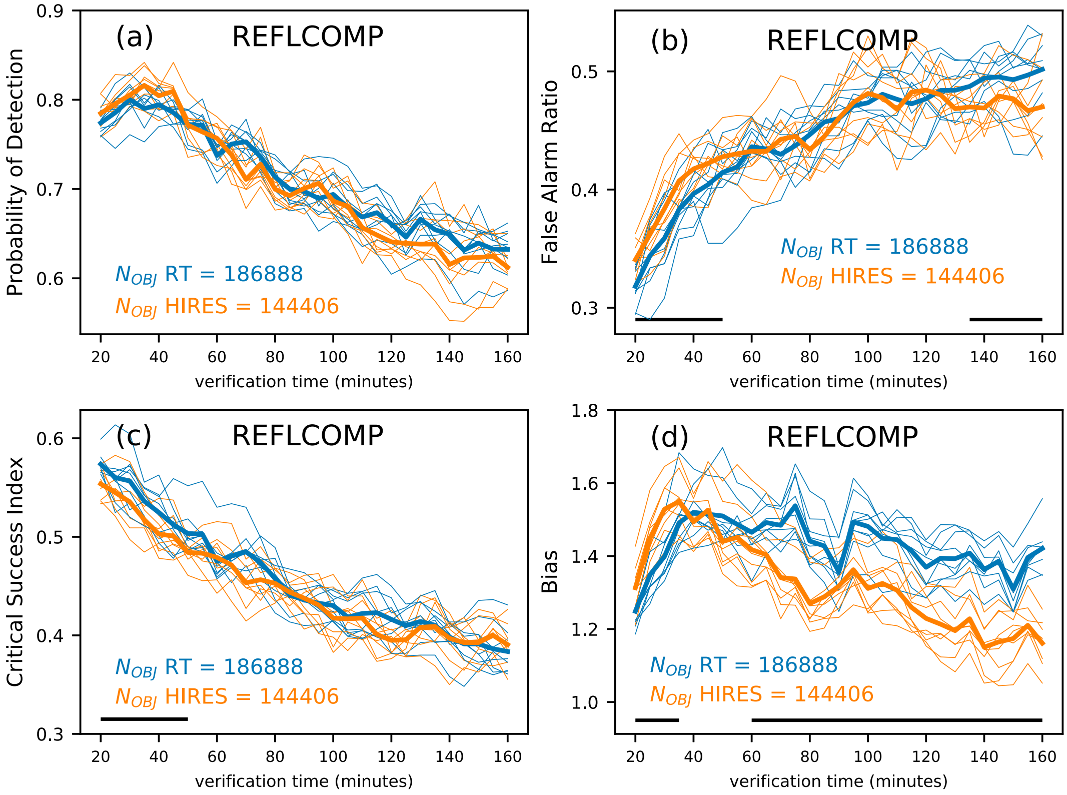

FIG. 3. Forecast time series of (a) $\mathrm{POD_{DET}}$, (b) $\mathrm{FAR_{DET}}$, (c) $\mathrm{CSI_{DET}}$, and (d) $\mathrm{BIAS_{DET}}$ for Realtime (blue) and HIRES (orange) reflectivity objects. Data for each ensemble member are averaged among the eleven 2020 cases shown in Table 1. Thin lines show individual members and thick lines show the ensemble mean. Plots are annotated with the total numbers of Realtime and HIRES reflectivity objects generated for all 11 cases at all 5-min model output times. Black horizontal lines above the abscissa denote time intervals where the Realtime and HIRES ensemble mean differences are statistically significant, as determined by a bootstrap resampling method (see text for details). Note that the deterministic object contingency table metrics are not computed for the first and last 20 forecast minutes because the object matching algorithm uses a $t = \pm 20$-min time window (section 2d).

show little difference (Figs. 3a–c). However, we do find a modest, albeit statistically significant, ~0.02 magnitude reduction in HIRES $\mathrm{CSI_{DET}}$ compared to Realtime over the t = 20–50-min period (Fig. 3c) which results from a greater HIRES over-forecasting bias (Figs. 3b,d); recall that the SFE participants noted a tendency for HIRES to initiate convection more quickly compared to Realtime (section 3a). Here we consider differences between the HIRES and Realtime forecasts to be statistically significant if both (i) the HIRES ensemble mean falls outside of a 95% confidence interval generated by bootstrap resampling ($N_{\mathrm{boot}} = 1000$; Wilks 2011) of the Realtime data over all members and cases, using a 15-min data binning interval; and (ii) a permutation test run on the Realtime and HIRES bootstrapped distributions yields a p value of <0.05. Interestingly, HIRES forecasts begin showing a statistically significant ~15% lower (i.e., improved) ensemble mean $\mathrm{BIAS_{DET}}$ after t = 60 min (Fig. 3d). Table 3 shows that Realtime and HIRES reflectivity object time-averaged ensemble mean $\mathrm{CSI_{DET}}$ are quite similar for most individual cases. As previously stated, many of our 11 cases are mixed-mode dominant (Table 1), and so our set of Realtime and HIRES reflectivity objects includes a significant contribution from linear structures (not shown) that should be more predictable than cellular objects. We should keep in mind that our reflectivity object $\mathrm{CSI_{DET}}$ verification considers only their *occurrence*. It is possible that reducing the WoFS Δx to 1.5 km may improve representation of finer-scale structures within these reflectivity objects important to severe weather generation, a topic that we shall explore in section 3d.

Turning to the midlevel rotation track objects (Fig. 4), we find more notable improvement in HIRES skill relative to Realtime forecasts. HIRES ensemble mean midlevel rotation track object $\mathrm{POD_{DET}}$ exceeds that of Realtime forecasts by ~0.1 throughout the verification period—which is statistically significant (Fig. 4a), whereas the two configurations show little difference in midlevel rotation track object $\mathrm{FAR_{DET}}$ (Fig. 4b). Thus, the HIRES ensemble mean midlevel rotation track object $\mathrm{CSI_{DET}}$ exceeds that of Realtime forecasts by a statistically significant ~0.05 throughout the verification period (Fig. 4c). Both Realtime and HIRES midlevel rotation track object $\mathrm{CSI_{DET}}$ scores are substantially lower than their respective reflectivity object $\mathrm{CSI_{DET}}$ scores, similar to what S18 found.

TABLE 3. Ensemble mean $\mathrm{CSI_{DET}}$, averaged over all 5-min verification times (t = 20–160 min for reflectivity objects and t = 50–160 min for rotation track objects), for each of the 2020 SFE case days. Realtime and HIRES values are listed as pairs delimited by the "/" symbol, with the latter given in italic font.

Day	REFLCOMP deterministic objects	$\mathrm{UH_{2–5}}$ deterministic objects	$\mathrm{UH_{0–2}}$ deterministic objects
2 Mar	0.50/*0.50*	0.28/*0.35*	0.33/*0.36*
28 Apr	0.43/*0.40*	0.35/*0.41*	0.28/*0.27*
4 May	0.62/*0.64*	0.35/*0.43*	0.26/*0.29*
7 May	0.40/*0.38*	0.43/*0.40*	0.46/*0.43*
13 May	0.47/*0.46*	0.08/*0.13*	0.08/*0.14*
15 May	0.53/*0.54*	0.28/*0.38*	0.19/*0.30*
20 May	0.35/*0.27*	0.10/*0.13*	0.05/*0.07*
22 May	0.32/*0.32*	0.25/*0.31*	0.28/*0.33*
26 May	0.44/*0.45*	0.00/*0.00*	0.01/*0.06*
27 May	0.51/*0.51*	0.29/*0.32*	0.25/*0.28*
29 May	0.45/*0.44*	0.08/*0.10*	0.03/*0.08*

Differences between Realtime and HIRES low-level rotation track forecasts show a similar pattern (Fig. 5), with HIRES having a statistically significant higher ensemble mean $\mathrm{CSI_{DET}}$ prior to t = 120 min (Fig. 5c), driven by a higher ensemble $\mathrm{POD_{DET}}$ (Fig. 5a). The modest improvement in low-level mesocyclone detection attained when WoFS Δx is reduced from 3 to 1.5 km is consistent with Potvin and Flora (2015), who found that 1-km or smaller grid spacing was necessary for properly capturing the timing and intensity of low-level mesocyclones in their idealized simulated supercells. However, reducing WoFS Δx to 1.5 km also tends to increase the low- and midlevel mesocyclone over-forecasting bias (Figs. 4d, 5d).

Figure 6 shows performance diagrams (Roebber 2009) comparing Realtime and HIRES t = 120-min reflectivity and rotation track object contingency table metrics computed separately for each member and case day. No cases show any notable improvement in reflectivity object forecast $\mathrm{CSI_{DET}}$ at t = 120 min (Fig. 6a). The 7, 13, 26, and 27 May cases appear to drive the improved (i.e., lower) HIRES reflectivity object over frequency bias at later forecast times (Fig. 3d; Table 4). For midlevel rotation track objects, the 2 March, 28 April, and 4, 13, 15, and 22 May cases' $\mathrm{CSI_{DET}}$ scores show the most improvement with reduced horizontal grid spacing (Fig. 6b; Table 3). Of these cases, 13, 15, and 22 May show the most improvement in their low-level rotation track object CSI scores when Δx is reduced (Fig. 6c; Table 3). Interestingly, although the 0300 UTC 3 March initialized HIRES 3-h forecast shows a stronger (as compared to Realtime) signal for the development of the supercell in northwestern Tennessee

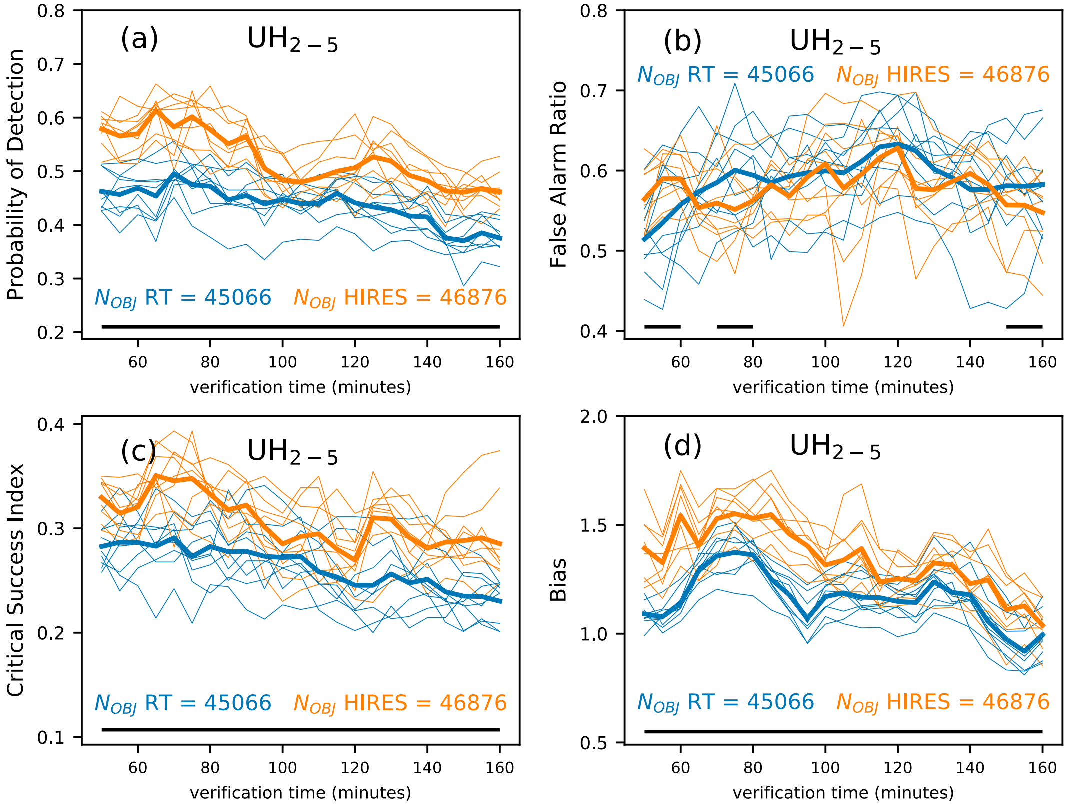

FIG. 4. As in Fig. 3, but for midlevel (2–5 km AGL) rotation track objects. Since rotation track objects are built using forecast data from the prior 30 min and a t = ±20-min time window is used for object matching (section 2d), deterministic contingency table metrics are not computed prior to t = 50 min.

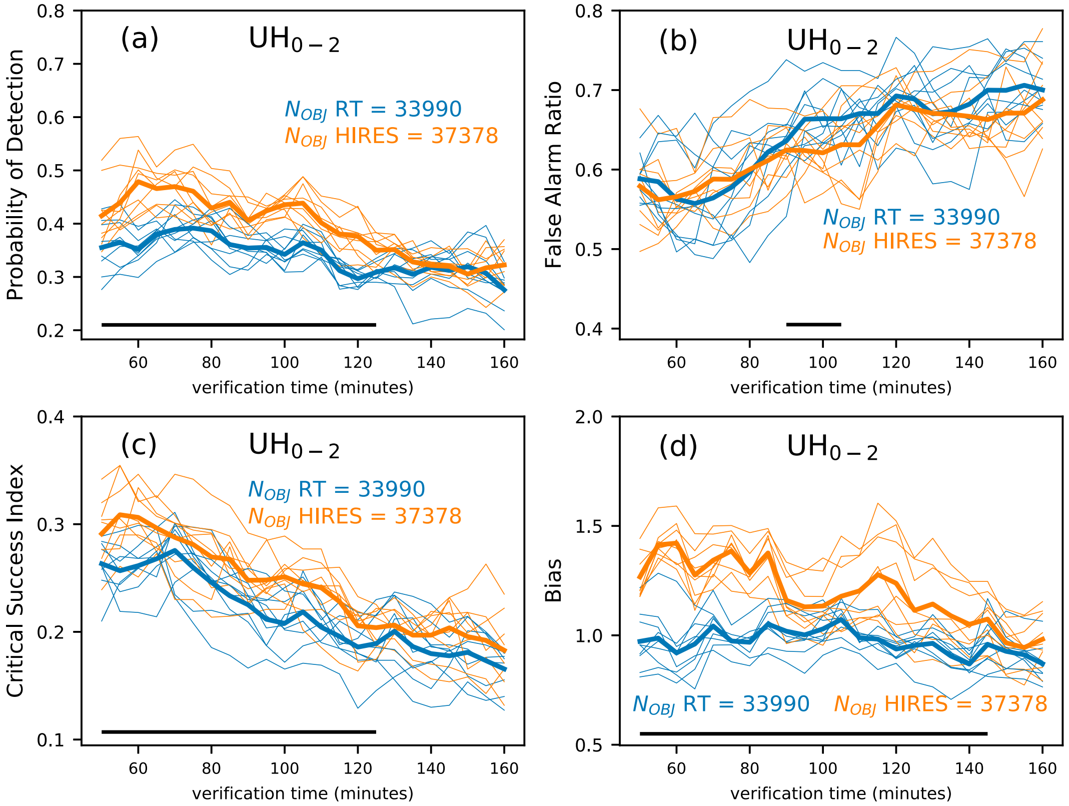

FIG. 5. As in Fig. 3, but for low-level (0–2 km AGL) rotation track objects.

(Figs. 7a–c) that produces the Nashville tornado (section 2a), the Realtime forecast initialized an hour later captures this supercell track quite well and the HIRES forecast shows little further improvement (Figs. 7d–f). The spurious southeasterly moving supercell forecast by the 0300 UTC initialized HIRES ensemble (Fig. 7b) also illustrates an example of the tendency for HIRES to over-forecast low-level rotation (Fig. 5d).

No obvious patterns in the CAPE regimes (high versus low) or convective mode characteristics appear to distinguish the cases where reducing Δx to 1.5 km improved the time-averaged ensemble mean $\mathrm{CSI_{DET}}$ score (cf. Tables 1 and 3). We should note, however, that except for 2–3 March, our cases tend to be mixed-mode dominant; further work is needed to more fully assess the impact of reduced Δx on more "classic" Plains region supercells. The extremely low midlevel and low-level rotation track object $\mathrm{CSI_{DET}}$ scores for 26 and 29 May (Figs. 6b,c; Table 3) reflect the fact that these low-CAPE cases tended to produce weak updrafts (not shown) and, thus, very few of their forecast storm UH values exceeded the object identification thresholds tied to the 2020 seasonal case climatology.

c. *Probabilistic object-based verification*

Figures 8 and 9 show performance diagrams for midlevel and low-level UH probability swath objects, respectively. Realtime and HIRES probability swath objects generated from all case days and initialization times are separately aggregated into early-verification (t = 0–90 min) and late-verification (t = 90–180 min) forecast period "batches." For each batch, all N_p probability swath objects sharing the same ensemble probability p (e.g., 1/9, 2/9, …) are matched to observed rotation track objects (section 2g), yielding an N_p-length vector of binary outcomes where 1 and 0 denote a match and nonmatch, respectively. The binary outcome vector is resampled with replacement N_{boot} = 1000 times, and for each bootstrap iteration, a set of equations analogous to Eqs. (2a)–(2d) are used to assign $\mathrm{POD_{PROB}}$, $\mathrm{FAR_{PROB}}$, $\mathrm{CSI_{PROB}}$ and $\mathrm{BIAS_{PROB}}$ scores to the N_p objects. Figures 8 and 9 show the mean performance diagram curve and its surrounding 95% confidence interval. In addition, the normalized area under the performance curve (NAUPDC) and normalized CSI (NCSI) are provided for both the Realtime and HIRES ensembles (Flora et al. 2021). NAUPDC and NCSI are defined as

$$\mathrm{NAUPDC} = \frac{\mathrm{AUPDC} - c}{1 - c}, \qquad (6a)$$

$$\mathrm{NCSI} = \frac{\mathrm{CSI} - c}{1 - c}, \qquad (6b)$$

where c is the "skew" (Boyd et al. 2012; Lampert and Gançarski 2014; Flora et al. 2021), which in this case is equal

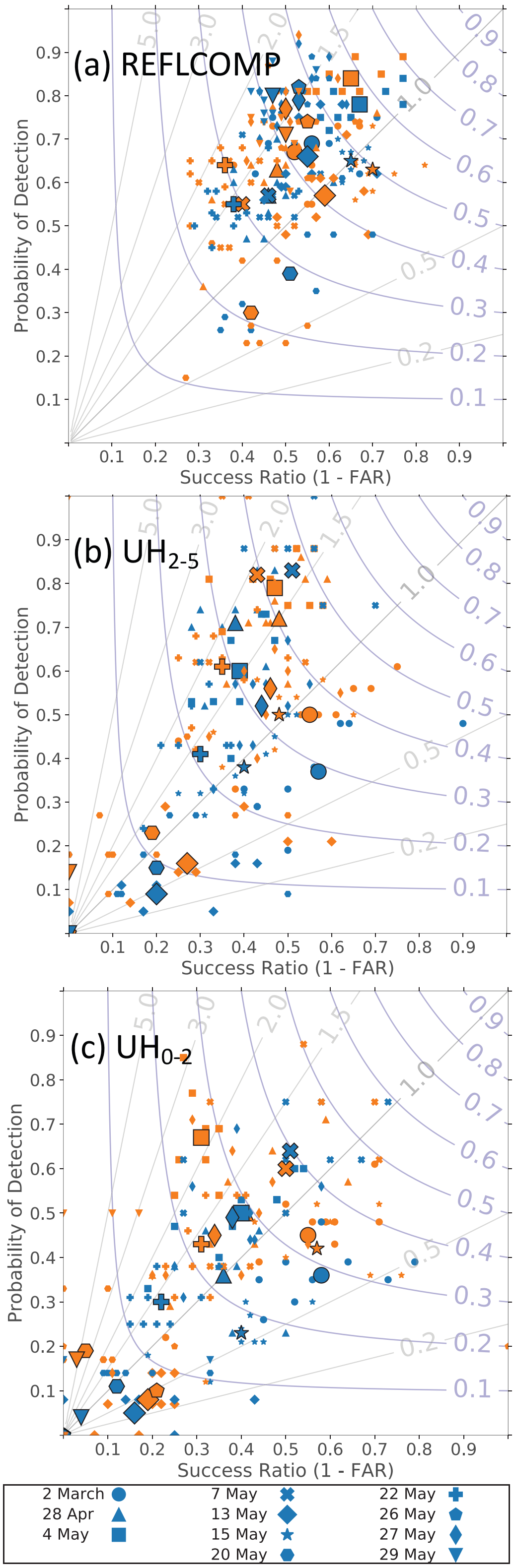

FIG. 6. (a) Performance diagram of t = 120-min forecast reflectivity object POD_{DET} and success ratio ($SR_{DET} = 1 - FAR_{DET}$), where each symbol denotes a different case day.

TABLE 4. As in Table 3, but for ensemble mean $BIAS_{DET}$.

Day	REFLCOMP deterministic objects	$UH_{2–5}$ deterministic objects	$UH_{0–2}$ deterministic objects
2 Mar	1.44/*1.40*	0.55/*0.88*	0.71/*0.90*
28 Apr	1.08/*1.05*	2.13/*1.90*	1.25/*1.04*
4 May	1.43/*1.47*	1.89/*1.90*	1.60/*1.89*
7 May	1.24/*1.13*	1.74/*2.24*	1.49/*1.50*
13 May	1.25/*1.00*	0.37/*0.55*	0.39/*0.67*
15 May	1.30/*1.14*	0.90/*0.97*	0.50/*0.67*
20 May	0.93/*0.77*	0.68/*1.02*	1.29/*2.05*
22 May	1.74/*1.73*	1.13/*1.36*	1.20/*1.17*
26 May	2.00/*1.75*	0.80/*0.93*	0.28/*0.54*
27 May	1.48/*1.40*	1.21/*1.22*	1.10/*1.07*
29 May	1.84/*1.73*	0.81/*1.58*	0.35/*1.46*

to the fraction of all probability swath objects that are matched. AUPDC is computed with the following formula:

$$\mathrm{AUPDC} = \sum_{k=1}^{K} (\mathrm{POD}_k - \mathrm{POD}_{k-1})\mathrm{SR}_k, \qquad (7)$$

where $k = 1, \ldots, K$ indexes the ensemble probability thresholds (i.e., 1/9, 2/9, etc.) and $SR_k = 1 - FAR_k$. The normalization is with respect to a low-skill baseline, and thus NAUPDC and NCSI are skill scores similar to the Brier skill score where positive (negative) values indicate performance better (worse) than the baseline (values near zero indicate similar performance). Note that our decision to use AUDPC as a probabilistic forecast skill verification metric rather than the commonly used receiver operating characteristic (ROC) curve is based on the former being more appropriate for the prediction of rare events (Davis and Goadrich 2006; Saito and Rehmsmeier 2015).

Comparing the Realtime and HIRES t = 0–90-min midlevel UH probability swath object forecasts (Fig. 8a), we find a larger HIRES CSI_{PROB} for all ensemble probability thresholds, particularly for the p = 3/9, 6/9, 7/9, and 8/9 probabilities (see annotated circles in Fig. 8). At lower probabilities (e.g., p = 3/9), an increased HIRES POD_{PROB} drives the CSI_{PROB} improvement relative to the Realtime WoFS, whereas for higher probabilities (e.g., p = 7/9 and p = 8/9), the CSI_{PROB} improvement is more driven by HIRES's reduced FAR_{PROB}. Realtime and HIRES t = 0–90-min midlevel UH probability swath object CSI_{PROB} are both maximized at p = 4/9, where their respective mean values are ~0.38 and ~0.4.

⟵

Blue and orange colors are used for Realtime and HIRES forecasts, respectively. Labeled curved contours and straight lines denote constant CSI_{DET} and $BIAS_{DET}$, respectively. Small and large symbols denote individual members and the ensemble mean, respectively. (b) As in (a), but for midlevel (2–5 km AGL) rotation track objects. (c) As in (a), but for low-level (0–2 km AGL) rotation track objects.

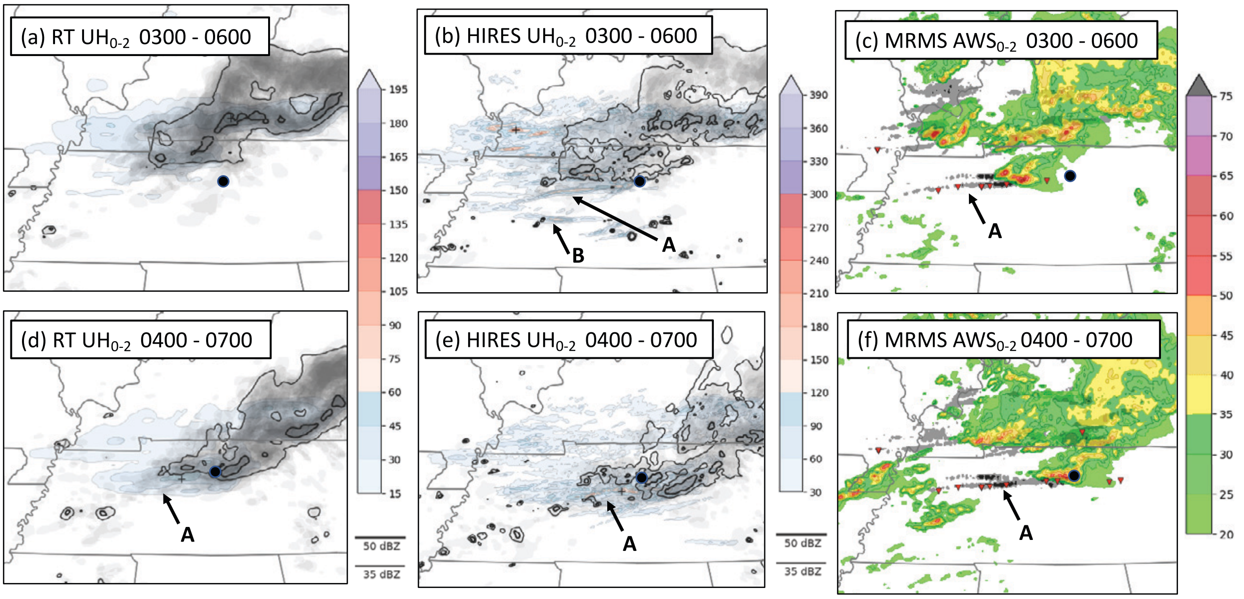
(a) RT UH$_{0-2}$ 0300 - 0600
(b) HIRES UH$_{0-2}$ 0300 - 0600
(c) MRMS AWS$_{0-2}$ 0300 - 0600
(d) RT UH$_{0-2}$ 0400 - 0700
(e) HIRES UH$_{0-2}$ 0400 - 0700
(f) MRMS AWS$_{0-2}$ 0400 - 0700
A
B
195
180
165
150
135
120
105
90
75
60
45
30
15
390
360
330
300
270
240
210
180
150
120
90
60
30
75
70
65
60
55
50
45
40
35
30
25
20
50 dBZ
35 dBZ

FIG. 7. (a) Realtime (RT) WoFS-forecast UH$_{0–2}$ 90th percentile swaths (color shaded; m^2 s^{-2}) from the 0300 UTC 3 Mar initialized forecast, generated using data collected from all 18 ensemble members over the t = 0–3-h forecast period. "Paintball" objects where composite reflectivity exceeds 40 dBZ (gray shading; darkness increases with number of overlapping ensemble members), and probability-matched mean (PMM; Ebert 2001) composite reflectivity (dBZ; black contours), both from t = 3 h, are also shown. (b) As in (a), but for the 0300 UTC 3 Mar initialized 9-member HIRES WoFS forecast. (c) MRMS composite reflectivity (color shaded; dBZ) from 0600 UTC 3 Mar and AWS$_{0–2}$ (0.004 and 0.008 s^{-1} in gray and black shading, respectively) temporal maximum-value swaths taken from the 0300–0600 UTC period. (d),(e) As in (a) and (b), but for the 0400 UTC 3 Mar initialized Realtime and HIRES forecasts, respectively. (f) As in (c), but for MRMS composite reflectivity valid at 0700 UTC 3 Mar and AWS$_{0–2}$ swaths from 0400 to 0700 UTC 3 Mar. Red triangles in (c) and (f) mark locations of SPC local tornado reports recorded during their respective 3-h AWS$_{0–2}$ compositing windows. Letters A and B label the track of the Nashville, TN, supercell and a spurious southeastward-moving supercell, respectively; both are discussed in the text. Black dot symbols in (a)–(f) denote Nashville's location.

For most probability thresholds, Realtime and HIRES midlevel UH CSI$_{\mathrm{PROB}}$ becomes smaller at t = 90–180 min relative to the previous 90-min period (cf. Figs. 8a, 8b), similar to the trend found in F19; this is not surprising, given our expectation for mesocyclone predictability to decrease at longer forecast lead times. However, we find considerably less HIRES improvement over the Realtime WoFS for midlevel UH CSI$_{\mathrm{PROB}}$ over the t = 90–180-min period (Fig. 8b). Given that our event-based mesocyclone probability framework is contingent on the parent thunderstorm having already formed (section 2g), it is possible that (i) the increased positional displacement of the same storm resolved in different members and (ii) the dissipation of some storms at later lead times renders the increased grid resolution less effective in improving probabilistic midlevel UH forecast skill after t = 90 min. Low-level UH probability swath objects, on the other hand, show no improvement in t = 0–90-min CSI$_{\mathrm{PROB}}$ with smaller Δx for most ensemble probabilities (Fig. 9a).

Using a permutation test (Wilks 2011), we find that compared to the Realtime ensemble, HIRES has a statistically significant (p value ~0.001): (i) higher (i.e., improved) NAUPDC and NCSI for midlevel UH probability swaths over the full forecast period; (ii) higher (i.e., improved) NCSI for t = 90–180-min low-level UH probability swaths; and (iii) lower (i.e., degraded) NAUPDC and NCSI for t = 0–90-min low-level UH probability swaths. F19 found lower CSI$_{\mathrm{PROB}}$ values (below 0.3 for all ensemble probabilities) for 2017 and 2018 3-km WoFS UH probability swath objects matched to observations using a similar method; differences between their results and those for the 2020 Realtime WoFS shown in Figs. 8 and 9 could result in part from the lack of time difference tolerance used by F19 in their probabilistic object matching.

Next, we compare the reliability of Realtime and HIRES forecasts. Figure 10a shows that both Realtime and HIRES midlevel UH probability swath object t = 0–90-min forecasts are generally overconfident, which could be attributable to WoFS underdispersive behavior also noted by F19. Recall from section 2g that the consistency bars provide a range of conditional event frequencies indicative of a plausibly reliable forecast. As explained in Bröcker and Smith (2007), consistency bar length tends to be inversely proportional to sample size, which is why the Realtime and HIRES consistency bars become longer at higher probability thresholds where fewer midlevel UH probability swath objects are generated, as shown in Fig. 10a.

However, both ensembles' UH probability swath forecasts generally become more reliable for later forecast lead times (t = 90–180 min; Fig. 10b), perhaps due to high-probability

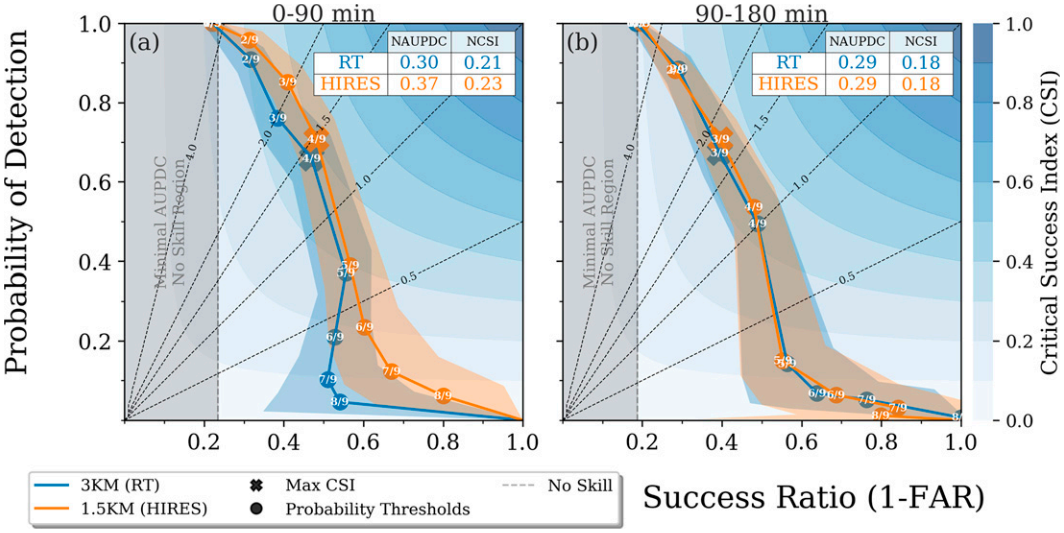

FIG. 8. (a) Performance diagram for the set of WoFS-forecast midlevel (2–5 km AGL) UH probability swath objects generated from the 0–30-, 15–45-, 30–60-, 45–75-, and 60–90-min forecast time intervals. Each subset of midlevel probability swath objects sharing a common probability threshold (e.g., 1/9, 2/9, …) is resampled using bootstrapping, yielding a mean $\mathrm{POD_{PROB}}$ and $\mathrm{SR_{PROB}}$, denoted by a circle annotated with the threshold value, and a 95% confidence interval surrounding the mean (shaded area). Blue and orange colors show Realtime (RT) and HIRES forecasts, respectively. Blue shaded contours and dotted black lines plot $\mathrm{CSI_{PROB}}$ and $\mathrm{BIAS_{PROB}}$, respectively. The X symbols denote the ensemble probabilities with maximum $\mathrm{CSI_{PROB}}$. The gray region shows a hypothetical minimal-skill forecast AUPDC described by Eq. (6); the vertical line bounding it to the right marks the base rate. (b) As in (a), but for midlevel probability swath objects generated from the 75–105-, 90–120-, 105–135-, 120–150-, 135–165-, and 150–180-min forecast time intervals.

objects becoming rarer as different member forecasts of the same storm event become more spatially separated; F19 reported a similar trend for 2017/18 WoFS midlevel UH probabilistic forecasts (see their Fig. 7). The positive Brier skill score (BSS; Wilks 2011) for both Realtime and HIRES midlevel UH probabilistic forecasts over the full forecast period (Figs. 10a,b) suggests that both ensembles have sufficient skill and reliability to outperform a minimally skilled climatological forecast; these differences are statistically significant (p value ~ 0.001 using a permutation test). A reliability diagram comparing Realtime and HIRES low-level UH probability swaths (not shown) resembles Fig. 10.

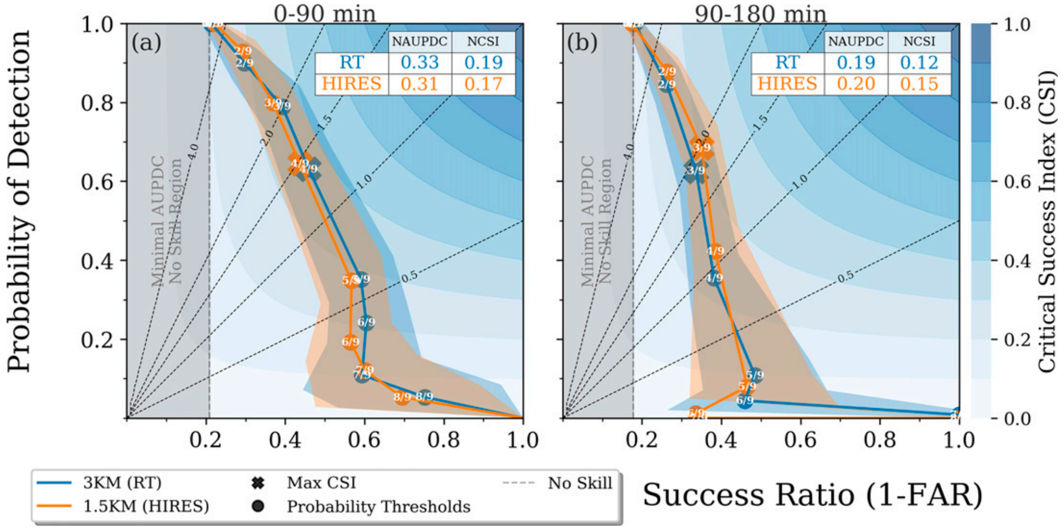

FIG. 9. As in Fig. 8, but for low-level (0–2 km AGL) UH probability swath objects.

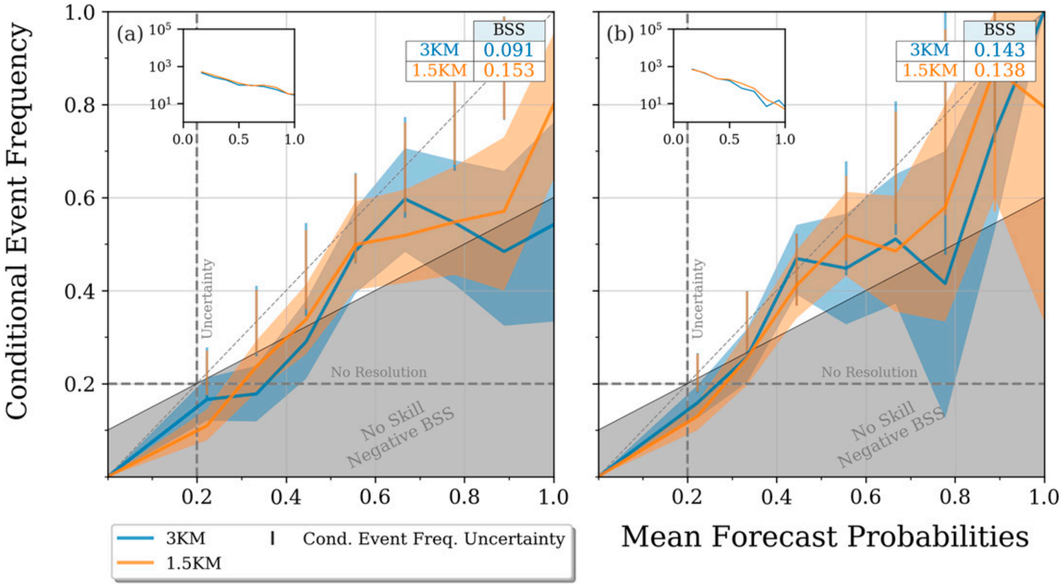

FIG. 10. (a) Reliability diagram for midlevel (2–5 km AGL) UH probability swath objects aggregated over 30-min forecasts every 15 min with valid times up to 90 min (e.g., 0–30, 15–45, … , 60–90 min). Blue (orange) curves plot mean values of the Realtime (HIRES) bootstrapped conditional event frequency distributions on the *y* axis, computed for each ensemble probability threshold (e.g., 1/9, 2/9, …), as shown on the *x* axis. Blue (orange) shading shows the Realtime (HIRES) bootstrapped conditional event frequency distribution's 95% confidence interval surrounding the mean. The dashed diagonal line represents a perfectly reliable forecast, and the blue (orange) vertical lines surrounding the diagonal show Realtime (HIRES) consistency bars computed using the method of Bröcker and Smith (2007). The base rate (BR ~0.2 for both Realtime and HIRES) is also plotted on the *x* and *y* axes, and the rectangular region where x > BR and y > BR indicates a positive BSS (i.e., some degree of forecast skill compared to event climatology). Blue (orange) lines in the inset figure plot the number of probability swath objects as a function of ensemble probability threshold. (b) As in (a), but for forecasts with valid times between 90 and 180 min (e.g., 90–120, 105–135, … , 150–180 min).

d. Storm-scale characteristics

Finally, the sensitivity of WoFS forecast storm characteristics to reducing Δx by a factor of 2 will be briefly explored. Figure 11a compares Realtime and HIRES forecast time series of column-maximum updraft speed (hereafter w_{MAX}) averaged over reflectivity objects from all 11 cases. Both Realtime and HIRES reflectivity objects tend to experience a w_{MAX} spike over the first 15 forecast minutes, likely a result of storm-scale model spinup. After t = 30 min, object-

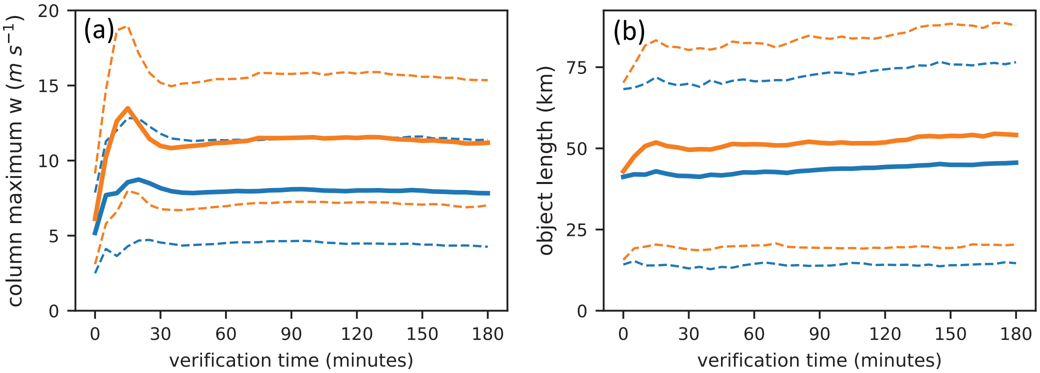

FIG. 11. (a) Time series of Realtime (blue) and HIRES (orange) forecast column-maximum w (m s^{-1}), averaged first over each reflectivity object and then among reflectivity objects from all 11 cases generated at each 5-min model output time (thick solid lines). Thin dashed lines bound the ±1 standard deviation range of Realtime (blue) and HIRES (orange) reflectivity object-averaged column-maximum w surrounding their respective mean values, when considering the forecast time-dependent distribution of reflectivity objects generated from all 11 cases. (b) As in (a), but for reflectivity object major axis length (km).

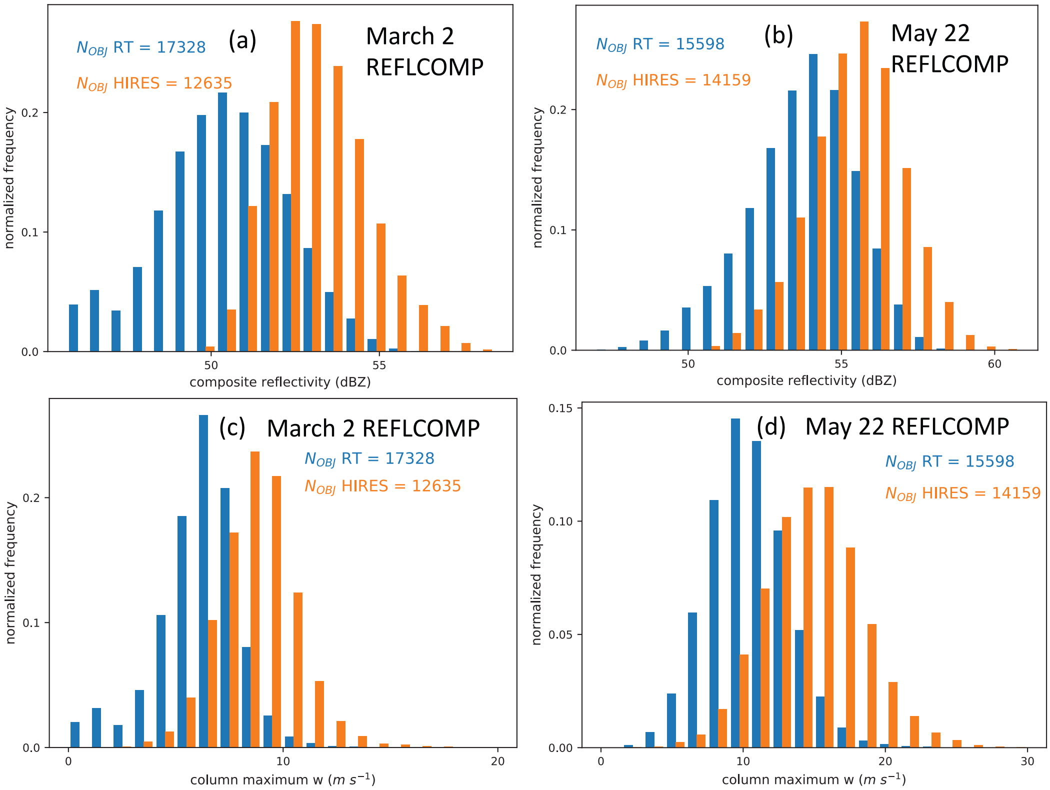

FIG. 12. (a) Histograms of composite reflectivity (dBZ) spatially averaged over each reflectivity object identified at all 3 Mar Realtime (blue) and HIRES (orange) 5-min model output times. For each histogram, bin frequencies are normalized to keep the area under the probability distribution function traced by the histogram equal to 1. (b) As in (a), but for 22 May forecasts. (c),(d) As in (a),(b), but for column-maximum w (m s^{-1}) averaged over each reflectivity object.

averaged $w_{\rm MAX}$ reaches a steady state intensity that is ~30% higher for HIRES objects compared to Realtime objects; these differences are statistically significant at the 99% confidence level via a Student's t test. Bryan and Morrison (2012) and Potvin and Flora (2015) similarly found, for squall lines and supercells, respectively, an increase in simulated updraft speed as Δx was reduced from 4 to 1 km, which these studies attributed to a better model representation of nonhydrostatic processes at smaller horizontal grid spacing. For example, other factors being unchanged, coarser grid cells could resolve anomalously wide updrafts with an anomalously strong buoyancy perturbation pressure gradient force (PGF$_b$) pointing downward through their cores; once a hypothetical updraft becomes sufficiently wide its PGF$_b$ approaches the hydrostatic PGF (Markowski and Richardson 2010). The first 15 forecast minutes also feature a sharp, statistically significant increase in HIRES reflectivity object length (Fig. 11b) and eccentricity (not shown) compared to those of Realtime reflectivity objects, perhaps due to accelerated cell mergers in HIRES forecasts.

Figure 12a shows histograms of composite reflectivity spatially averaged over each reflectivity object for the 2 March Realtime and HIRES forecasts. The histogram left-tail cutoffs are sensitive to the separate Realtime and HIRES climatology-based reflectivity thresholds chosen for defining the reflectivity object boundaries (Table 2). Mindful that these histograms only include points from the most intense convective cells (but also those most likely to produce severe weather), we find a ~3-dBZ increase in median HIRES reflectivity over median Realtime reflectivity. Comparison of Realtime and HIRES 2 March histograms generated in the same manner but for $w_{\rm MAX}$ (Fig. 12c) reveals a ~2 m s^{-1} stronger median $w_{\rm MAX}$ for the latter. Compared to those of 2 March, the 22 May Realtime and HIRES reflectivity object REFLCOMP and $w_{\rm MAX}$ histograms (Figs. 12b,d) are shifted toward higher values, consistent with the higher 22 May environmental CAPE (Table 1). Bryan and Morrison (2012) showed how the relationship between precipitation rate and CAM Δx can be complex and potentially nonmonotonic, given the competing effects of enhanced condensation and enhanced cloud evaporation as Δx decreases. Here, we find that HIRES reflectivity objects have higher median REFLCOMP compared to their Realtime counterparts, both for the low-CAPE 2 March case (Fig. 12a), the high-CAPE 22 May case (Fig. 12b), and the others (not shown); this suggests that for the most intense WoFS-forecast cells likely to produce severe weather, the condensation effect may outweigh the evaporation effect over the Δx = 3–1.5-km range. It is possible, however, that cloud evaporation could increase more strongly with

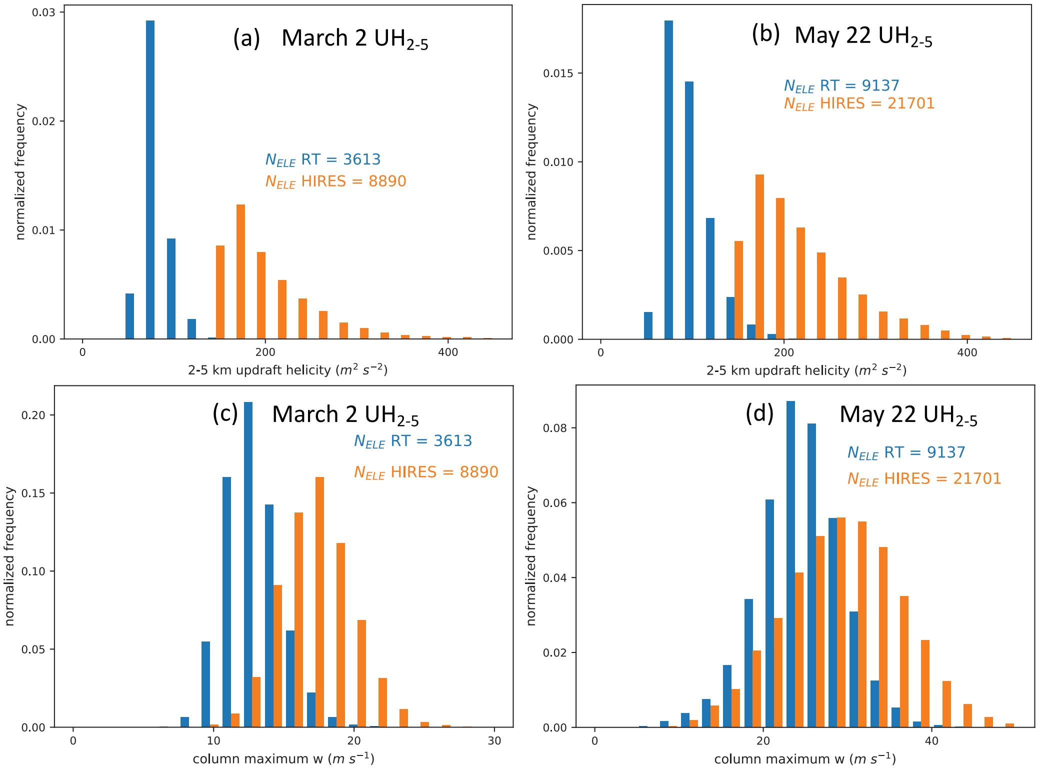

FIG. 13. As in Fig. 12, but for variables averaged spatially over the 5-min $UH_{2–5}$ elements used in constructing midlevel rotation track objects. Element-averaged $UH_{2–5}$ ($m^2\ s^{-2}$) and column-maximum w (m s^{-1}) histograms are shown in (a) and (b) and (c) and (d), respectively.

reduced WoFS Δx for drier environments or when different microphysics parameterizations are used.

Figures 13 and 14 compare Realtime and HIRES histograms of variables associated with midlevel and low-level rotation track objects, respectively. As for the reflectivity objects, reducing Δx from 3 to 1.5 km results in stronger simulated rotating storms on 2 March, as evidenced by the higher median HIRES $UH_{2–5}$ (Fig. 13a), w_{MAX} (Fig. 13c), $UH_{0–2}$ (Fig. 14a), and low-level ζ (Fig. 14c), as compared to Realtime storms. The 22 May forecasts show similar results (Figs. 13b,d, 14b,d), although like other high-CAPE cases (e.g., 28 April and 4 May, not shown), their median updraft intensities are less sensitive to Δx (cf. 13c,d). Given the complex dependencies between supercell updraft intensity, updraft width, CAPE, and environmental shear found in idealized simulations (Peters et al. 2019, 2020), it is not surprising to find variability in rotating storm w_{MAX} sensitivity to Δx among different 2020 case days. Also notable is the broader distribution of HIRES rotating storm w_{MAX} (Figs. 13c,d) and low-level ζ (Figs. 14c,d) compared to the Realtime WoFS, which suggests that the smaller Δx helps the model to resolve a greater range of mesocyclone intensities.

Figure 15 compares Realtime and HIRES boxplots (Wilks 2011) showing distributions of local horizontal-maximum values of a few additional variables extracted from reflectivity objects and their near-storm environments (NSEs), defined here as 120 km × 120 km domains centered on the objects following Potvin et al. (2020). We find a notable 80 m AGL wind speed[4] distribution shift to higher values when Δx is reduced to 1.5 km, with the median increasing from 20.4 to 23.8 m s^{-1} (Fig. 15a). This result is encouraging, given the tendency for the Realtime WoFS to underpredict near-surface wind gust intensity (Flora et al. 2021); however, further work is necessary to better understand the grid resolution dependence of WoFS-forecast near-surface winds and validate these results against observations. Reduced Δx also shifts the maximum hail diameter, as predicted by the one-dimensional HAILCAST model coupled to WRF (Adams-Selin and Ziegler 2016; Adams-Selin et al. 2019), toward higher values (Fig. 15b). This is not a surprising result given the tendency for HIRES to generate stronger reflectivity object updrafts compared to Realtime (Fig. 11a). Interestingly, median 0–3-km storm-relative helicity ($SRH_{0–3}$)—the vertically integrated dot product of the horizontal relative vorticity vector component with the storm-relative environmental wind vector

[4] As mentioned in Flora et al. (2021), the WRF-output instantaneous 80 m AGL wind speed has traditionally been used as a proxy for the maximum 10 m AGL wind gust recorded over the time intervals between model output timestamps.

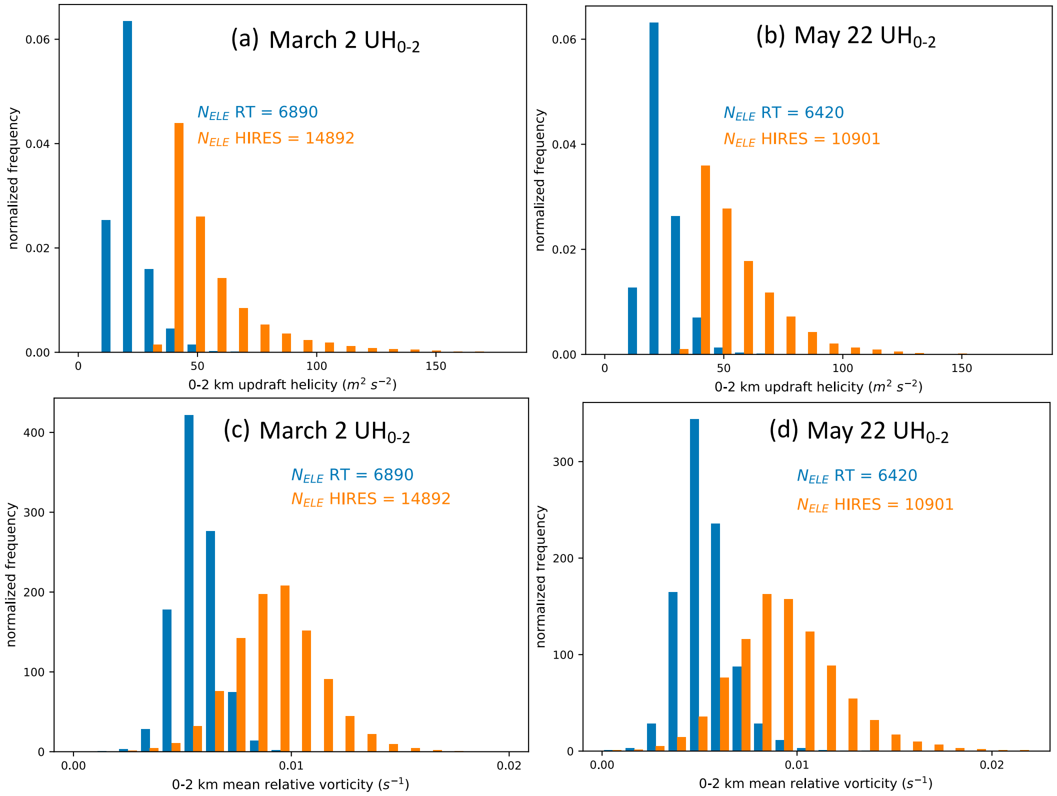

FIG. 14. As in Fig. 12 but for 5-min $UH_{0–2}$ elements used in building low-level rotation track objects. Element-averaged $UH_{0–2}$ ($m^2\ s^{-2}$) and 0–2 km AGL mean ζ (s^{-1}) histograms are shown in (a) and (b) and (c) and (d), respectively.

(Markowski and Richardson 2010), increases from 567 to 725 $m^2\ s^{-2}$ when Δx is reduced to 1.5 km (Fig. 15c). This implies that the smaller grid spacing may enhance low-level storm-relative inflow and VWS in the NSE through storm-environment interactions, a topic that invites further investigation. Median reflectivity object NSE mixed-layer significant tornado parameter (MLSTP; Thompson et al. 2003), a composite index measuring an environment's capability to support significant (EF2+) tornadoes that incorporates MLCAPE, SRH, VWS, liquid condensation level height, and convective inhibition, increases from 1.2 to 1.6 when Δx is reduced to 1.5 km (Fig. 15d).

Figures 16a and 16b compare Realtime and HIRES reflectivity object NSE bivariate distributions of $UH_{2–5}$ and MLCAPE, shown as kernel density estimates (KDEs; Scott 1992). Unsurprisingly, MLCAPE shows little dependency on Δx. These plots show how the tendency for $UH_{2–5}$ to increase with smaller grid spacing becomes most pronounced for storms in high-CAPE environments, an intriguing result that invites further investigation. KDEs comparing $UH_{0–2}$ with 0–1-km storm relative helicity ($SRH_{0–1}$; Figs. 16c,d) show an increased correlation of NSE low-level VWS and/or storm-relative inflow with storm $UH_{0–2}$ in HIRES compared to Realtime forecasts, providing further evidence that reduced Δx may enhance storm-environment interactions favorable for severe weather.

4. Summary and conclusions

This study addresses the question of whether reducing the WoFS forecast horizontal grid spacing from $\Delta x = 3$ km to $\Delta x = 1.5$ km sufficiently improves its deterministic and probabilistic forecast skill to justify the increased computational cost. It also seeks to better understand the sensitivity of WoFS-forecast storm and NSE characteristics to the reduction in Δx. For 11 case days selected from the 2020 HWT SFE, we have compared 9-member WoFS ensemble forecasts run using the pseudo-operational $\Delta x = 3$ km Realtime configuration against experimental HIRES ensemble forecasts that use a $\Delta x = 1.5$ km nest downscaled from the Realtime analyses at the model initialization. We validated 3-h Realtime and HIRES ensemble forecasts against observations using (i) subjective SFE participant impressions; (ii) deterministic matching of forecast reflectivity and rotation track objects to MRMS observed objects; and (iii) matching mesocyclone probability swath objects to observed rotation track objects to assess the probabilistic forecast skill and reliability. Our major findings are as follows:

- Compared to the Realtime WoFS, HIRES deterministic forecasts of reflectivity object occurrence have similar skill, as measured by CSI_{DET}, for most forecast times. However, although HIRES tends to over-forecast reflectivity objects during the first 30 forecast minutes, the HIRES ensemble

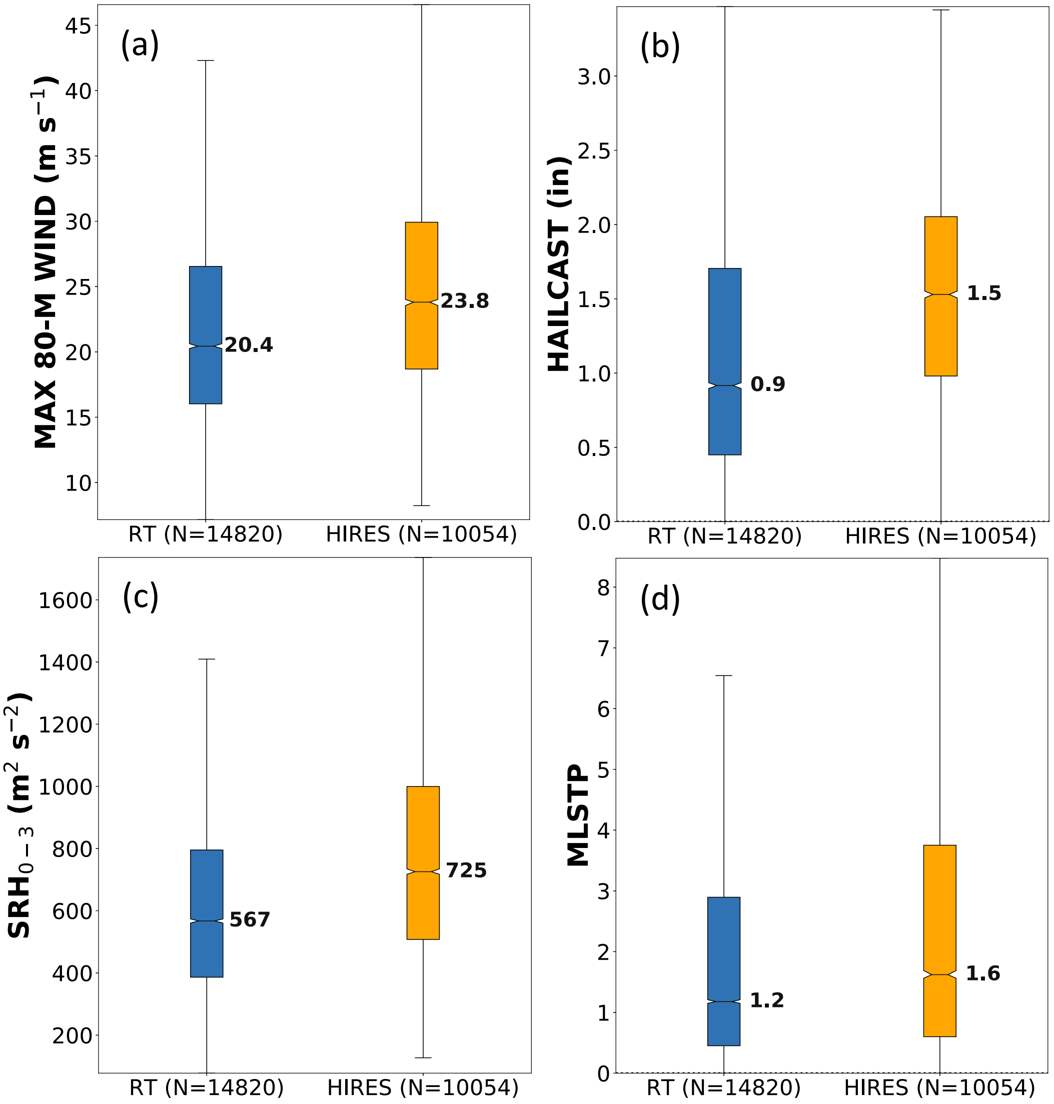

FIG. 15. Boxplots of (a) maximum 80-m wind speed (m s^{-1}), (b) WRF-HAILCAST maximum hail diameter (in), (c) 0–3-km storm-relative helicity (SRH$_{0\text{–}3}$; m^2 s^{-2}), and (d) mixed-layer significant tornado parameter (MLSTP). Each data point in the distribution represented by a boxplot is the maximum value taken from each WoFS-forecast reflectivity object and its surrounding NSE at forecast times t = 60, 120, and 180 min. Labeled notches show the distribution median values; shaded "box" bottoms and tops show the distribution 25th and 75th percentiles, respectively; and the vertical "whisker" lines extend to the distribution minimum and maximum values. Realtime and HIRES data distributions are constructed separately and shown in blue and orange colors, respectively.

mean BIAS$_{\mathrm{DET}}$ is ~15% lower (improved) compared to that of Realtime after t = 60 min.

- HIRES deterministic forecasts of midlevel mesocyclone occurrence are more skillful than those of the Realtime WoFS, as evidenced by their statistically significant ~0.05 higher ensemble mean CSI$_{\mathrm{DET}}$ throughout the forecast verification period, driven primarily by their higher POD$_{\mathrm{DET}}$. However, the HIRES ensemble also has a statistically significant larger over-forecasting bias for midlevel mesocylones.
- Differences between HIRES and Realtime low-level mesocyclone deterministic forecasts are qualitatively similar to those for midlevel mesocyclones, but the improvement in HIRES ensemble mean CSI$_{\mathrm{DET}}$ is statistically significant only through the first two hours of the forecast period.
- Reducing WoFS Δx improves midlevel mesocyclone probabilistic forecast skill across most ensemble probability thresholds, as measured by a statistically significant higher HIRES NAUPDC and NCSI compared to Realtime. This improvement is most notable prior to t = 90 min.

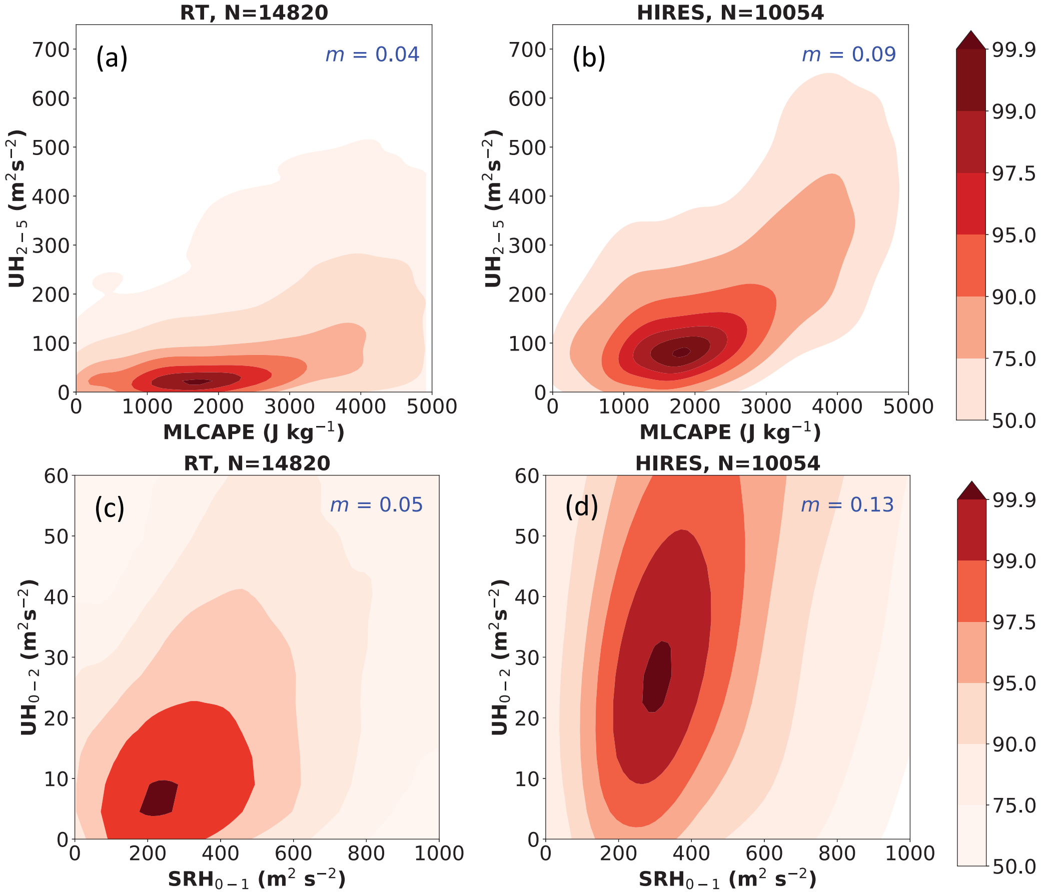

FIG. 16. (a),(b) Kernel density estimates of midlevel updraft helicity ($UH_{2–5}$) and mixed-layer CAPE, using maximum values computed from each reflectivity object and its surrounding NSE for (a) Realtime and (b) HIRES forecasts. (c),(d) As in (a) and (b), but for low-level updraft helicity ($UH_{0–2}$) and 0–1-km storm-relative helicity ($SRH_{0–1}$). Slopes of the distribution best-fit lines are annotated in the panels' upper-right-hand corners.

- Compared to the Realtime WoFS, HIRES low-level mesocyclone probabilistic forecast skill is degraded for the t = 0–90-min period and improved for the t = 90–180-min period; differences are statistically significant.
- Compared to Realtime reflectivity objects, HIRES reflectivity objects have a ~30% stronger mean w_{MAX}, 17% stronger median 80-m wind gust intensity, 67% larger median hail diameter, 28% higher median NSE $SRH_{0–3}$, and 33% higher median NSE MLSTP.
- Reducing Δx to 1.5 km results in a ~30%–40% higher median w_{MAX} for midlevel rotation track objects and a doubling in median ζ for low-level rotation track objects.
- SFE participants did not see large differences between the 18-member Realtime and 9-member HIRES ensembles at three different initialization times or added benefit from the HIRES.

Our results show that while reducing the WoFS grid spacing to 1.5 km generally improves deterministic skill (as measured by CSI_{DET}) in forecasting rotating storm occurrence, it has little positive impact on CSI_{DET} when considering a broader sample of rotating and nonrotating storms identified by their REFLCOMP exceeding the 99th climatological percentile (Figs. 3–5). Given that rotation track objects tend to have higher w_{MAX} compared to reflectivity objects for both the Realtime and HIRES WoFS (cf. 12, 13), our results are consistent with L21, who found that reducing WoFS Δx to 1 km only improved the detection of the most intense thunderstorms. Also like L21, we find that reduced WoFS grid spacing improves probabilistic mesocyclone forecasts, although L21 used an alternative probabilistic object verification metric focused on removal of forecaster prior uncertainty. Our finding a modest—but statistically significant—improvement in low-level mesocyclone detection in the HIRES WoFS for some forecast periods (Figs. 5c, 9b) also supports Potvin and Flora (2015), who showed that reducing Δx from 3 to 1 km can improve model representation of rapid low-level mesocyclone strengthening and weakening in idealized supercell simulations. Therefore, we find that reducing Δx to 1.5 km in downscaled WoFS forecasts initialized from Δx = 3 km analyses does, to some degree, benefit the WoF mission by improving detection of rotating storms at 0–3-h lead times.

The more challenging question to address going forward is whether the increased computational cost associated with halving the WoFS Δx could be more beneficial to the WoFS mission if put toward other purposes, such as increasing the ensemble size. The subjective evaluations comparing an 18-member Realtime and 9-member HIRES ensemble showed similar subjective ratings, and similar studies could be carried out in the future between an 18-member Realtime ensemble and a Realtime ensemble with larger membership to explore whether more ensemble members would create a larger difference in subjective rating. A future study examining WoFS Δx sensitivity in a larger number of cases to include more "classic" Great Plains supercells would also be beneficial. Our deterministic and probabilistic object-based analysis methods could also be further improved by developing an automated method for distinguishing between discrete mesocyclones and rotating storms embedded within squall lines, as it might be reasonable to expect that a smaller Δx could have a more beneficial impact on the latter, given their tendency to be associated with narrower, shallower, weaker, and more short-lived updrafts. Finally, we should keep in mind that this study only investigated the impact of reduced Δx in *downscaled* WoFS forecasts. It is possible that reducing Δx to 1.5 km within the 15-min WoFS data assimilation cycling—thus giving the background forecast fields a more finescale structure—could provide more significant benefit, particularly for later-cycle WoFS forecasts, although this would incur a significant additional computational expense. Nevertheless, we find that reducing WoFS free forecast Δx by a factor of ½ results in statistically significant improvement for most short-range deterministic and probabilistic mesocyclone forecast skill metrics evaluated; therefore, this WoFS configuration change should be strongly considered for future implementations.

Acknowledgments. The authors thank Thomas Galarneau for kindly providing an informal review of this manuscript with helpful comments. Gerry Creager's efforts in managing the NSSL high-performance computing system maintenance and operations were indispensable in ensuring smooth execution of the Realtime and HIRES WoFS forecast runs throughout the 2020 SFE period. This work was prepared by the authors with funding from a Presidential Early Career Award for Scientists and Engineers, the NSSL Forecast Research and Development Division, and the NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce.

Data availability statement. All datasets used for this study are stored on the NSSL high-performance computing server and the data can be made available upon request.

REFERENCES

Adams-Selin, R. D., and C. L. 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.

——, A. J. Clark, C. J. Melick, S. R. Dembek, I. L. Jirak, and C. L. Ziegler, 2019: Evolution of WRF-HAILCAST during the 2014–16 NOAA/Hazardous Weather Testbed Spring Forecasting Experiments. *Wea. Forecasting*, **34**, 61–79, https://doi.org/10.1175/WAF-D-18-0024.1.

Adlerman, E. J., and K. K. Droegemeier, 2002: The sensitivity of numerically simulated cyclic mesocyclogenesis to variations in model physical and computational parameters. *Mon. Wea. Rev.*, **130**, 2671–2691, https://doi.org/10.1175/1520-0493(2002)130<2671:TSONSC>2.0.CO;2.

Boyd, K., V. S. Costa, J. Davis, and C. D. Page, 2012: Unachievable region in precision-recall space and its effect on empirical evaluation. *Proc. 29th Int. Conf. on Machine Learning (ICML'12)*, Edinburgh Scotland, Pattern Analysis, Statistical Modelling and Computational Learning (PASCAL), IBMR, NSF, Microsoft Research, Facebook, 1619–1626.

Bröcker, J., and L. A. Smith, 2007: Increasing the reliability of reliability diagrams. *Wea. Forecasting*, **22**, 651–661, https://doi.org/10.1175/WAF993.1.

Brooks, H. E., and J. Correia Jr., 2018: Long-term performance metrics for National Weather Service tornado warnings. *Wea. Forecasting*, **33**, 1501–1511, https://doi.org/10.1175/WAF-D-18-0120.1.

Bryan, G. H., and H. Morrison, 2012: Sensitivity of a simulated squall line to horizontal resolution and parameterization of microphysics. *Mon. Wea. Rev.*, **140**, 202–225, https://doi.org/10.1175/MWR-D-11-00046.1.

——, J. C. Wyngaard, and J. M. Fritsch, 2003: Resolution requirements for the simulation of deep moist convection. *Mon. Wea. Rev.*, **131**, 2394–2416, https://doi.org/10.1175/1520-0493(2003)131<2394:RRFTSO>2.0.CO;2.

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

——, and Coauthors, 2021: A real-time, virtual spring forecasting experiment to advance severe weather prediction. *Bull. Amer. Meteor. Soc.*, **102**, E814–E816, https://doi.org/10.1175/BAMS-D-20-0268.1.

Davis, C. A., B. G. Brown, and R. G. Bullock, 2006a: Object-based verification of precipitation forecasts. Part I: Methodology 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.

Davis, J., and M. Goadrich, 2006: The relationship between Precision-Recall and ROC curves. *Proc. 23rd Int. Conf. on Machine Learning (ICML'06)*, Pittsburgh, PA, ICML, 233–240, https://doi.org/10.1145/1143844.1143874.

Done, J., C. A. Davis, and M. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the Weather Research and Forecasting (WRF) model. *Atmos. Sci. Lett.*, **5**, 110–117, https://doi.org/10.1002/asl.72.

Dowell, D., and Coauthors, 2016: Development of a High-Resolution Rapid Refresh Ensemble (HRRRE) for severe weather forecasting. *28th Conf. on Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., 8B.2, https://ams.confex.com/ams/28SLS/webprogram/Paper301555.html.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480, https://doi.org/10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

Flora, M. L., P. S. Skinner, C. K. Potvin, A. E. Reinhart, T. A. Jones, N. Yussouf, and K. H. Knopfmeier, 2019: Object-based verification of short-term, storm-scale probabilistic mesocyclone guidance from an experimental Warn-on-Forecast system. *Wea. Forecasting*, **34**, 1721–1739, https://doi.org/10.1175/WAF-D-19-0094.1.

——, 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.*, **49**, 1535–1557, https://doi.org/10.1175/MWR-D-20-0194.1.

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.

Hong, S., Y. Noh, and J. Dudhia, 2006: A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon. Wea. Rev.*, **134**, 2318–2341, https://doi.org/10.1175/MWR3199.1.

Houtekamer, P. L., and F. Zhang, 2016: Review of the ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **144**, 4489–4532, https://doi.org/10.1175/MWR-D-15-0440.1.

Janjić, Z. I., 2002: Nonsingular implementation of the Mellor–Yamada level 2.5 scheme in the NCEP Meso model. NCEP Office Note 437, 61 pp., http://www.emc.ncep.noaa.gov/officenotes/newernotes/on437.pdf.

Jones, T. A., K. Knopfmeier, D. Wheatley, G. Creager, P. Minnis, and R. Palikonda, 2016: Storm-scale data assimilation and ensemble forecasting with the NSSL experimental Warn-on-Forecast System. Part II: Combined radar and satellite data experiments. *Wea. Forecasting*, **31**, 297–327, https://doi.org/10.1175/WAF-D-15-0107.1.

——, P. S. Skinner, N. Yussouf, K. Knopfmeier, A. Reinhart, and D. Dowell, 2019: Forecasting high-impact weather in landfalling tropical cyclones using a Warn-on-Forecast system. *Bull. Amer. Meteor. Soc.*, **100**, 1405–1417, https://doi.org/10.1175/BAMS-D-18-0203.1.

——, and Coauthors, 2020: Assimilation of the *GOES-16* radiances and retrievals into the Warn-on-Forecast System. *Mon. Wea. Rev.*, **148**, 1829–1859, https://doi.org/10.1175/MWR-D-19-0379.1.

Kain, J. S., and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931–952, https://doi.org/10.1175/WAF2007106.1.

Lampert, T. A., and P. Gançarski, 2014: The bane of skew. *Mach. Learn.*, **97**, 5–32, https://doi.org/10.1007/s10994-013-5432-x.

Lawson, J. R., C. K. Potvin, P. S. Skinner, and A. E. Reinhart, 2021: The vice and virtue of increased horizontal resolution in ensemble forecasts of tornadic thunderstorms in low-CAPE, high-shear environments. *Mon. Wea. Rev.*, **149**, 921–944, https://doi.org/10.1175/MWR-D-20-0281.1.

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**, 171–194, https://doi.org/10.1175/2009JAS2965.1.

Markowski, P., and Y. Richardson, 2010: *Mesoscale Meteorology in Midlatitudes*. Wiley-Blackwell, 430 pp.

Mashiko, W., 2016a: A numerical study of the 6 May 2012 Tsukuba City supercell tornado. Part I: Vorticity sources of low-level and midlevel mesocyclones. *Mon. Wea. Rev.*, **144**, 1069–1092, https://doi.org/10.1175/MWR-D-15-0123.1.

——, 2016b: A numerical study of the 6 May 2012 Tsukuba City supercell tornado. Part II: Mechanisms of tornadogenesis. *Mon. Wea. Rev.*, **144**, 3077–3098, https://doi.org/10.1175/MWR-D-15-0122.1.

Mellor, G. L., and T. Yamada, 1982: Development of a turbulence closure model for geophysical fluid problems. *Rev. Geophys.*, **20**, 851–875, https://doi.org/10.1029/RG020i004p00851.

Miller, M. L., V. Lakshmanan, and T. M. Smith, 2013: An automated method for depicting mesocyclone paths and intensities. *Wea. Forecasting*, **28**, 570–585, https://doi.org/10.1175/WAF-D-12-00065.1.

Nakanishi, M., and H. Niino, 2004: An improved Mellor–Yamada level-3 model with condensation physics: Its design and verification. *Bound.-Layer Meteor.*, **112**, 1–31, https://doi.org/10.1023/B:BOUN.0000020164.04146.98.

——, and ——, 2006: An improved Mellor–Yamada level-3 model: Its numerical stability and application to a regional prediction of advection fog. *Bound.-Layer Meteor.*, **119**, 397–407, https://doi.org/10.1007/s10546-005-9030-8.

NOAA/NCEI 2021: NOAA/National Centers for Environmental Information database. NOAA/NCEI, accessed 15 May 2021, https://www.ncdc.noaa.gov/stormevents/.

Noda, A., and H. Niino, 2003: Critical grid size for simulating convective storms: A case study of the Del City supercell storm. *Geophys. Res. Lett.*, **30**, 1844, https://doi.org/10.1029/2003GL017498.

Peters, J. M., C. J. Nowotarski, and H. Morrison, 2019: The role of vertical wind shear in modulating maximum supercell updraft velocities. *J. Atmos. Sci.*, **76**, 3169–3189, https://doi.org/10.1175/JAS-D-19-0096.1.

——, ——, and G. L. Mullendore, 2020: Are supercells resistant to entrainment because of their rotation. *J. Atmos. Sci.*, **77**, 1475–1495, https://doi.org/10.1175/JAS-D-19-0316.1.

Potvin, C. K., and M. L. Flora, 2015: Sensitivity of idealized supercell simulations to horizontal grid spacing: Implications for Warn-on-Forecast. *Mon. Wea. Rev.*, **143**, 2998–3024, https://doi.org/10.1175/MWR-D-14-00416.1.

——, E. M. Murillo, M. L. Flora, and D. M. Wheatley, 2017: Sensitivity of supercell simulations to initial-condition resolution. *J. Atmos. Sci.*, **74**, 5–26, https://doi.org/10.1175/JAS-D-16-0098.1.

——, 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.

Roberts, B., M. Xue, A. D. Schenkman, and D. T. Dawson II, 2016: The role of surface drag in tornadogenesis within an idealized supercell simulation. *J. Atmos. Sci.*, **73**, 3371–3395, https://doi.org/10.1175/JAS-D-15-0332.1.

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

Saito, T., and M. Rehmsmeier, 2015: The Precision-Recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, **10**, e0118432, https://doi.org/10.1371/journal.pone.0118432.

Schoen, J. M., and W. S. Ashley, 2011: A climatology of fatal convective wind events by storm type. *Wea. Forecasting*, **26**, 109–121, https://doi.org/10.1175/2010WAF2222428.1.

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.

Scott, D. W., 1992: *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley and Sons, 360 pp.

Shin, H. H., and S.-Y. Hong, 2015: Representation of the subgrid-scale turbulent transport in convective boundary layers at gray-zone resolutions. *Mon. Wea. Rev.*, **143**, 250–271, https://doi.org/10.1175/MWR-D-14-00116.1.

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

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, https://doi.org/10.1175/BAMS-D-14-00173.1.

Stensrud, D. J., and Coauthors, 2009: Convective-scale Warn-on-Forecast system: A vision for 2020. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1500, https://doi.org/10.1175/2009BAMS2795.1.

——, and Coauthors, 2013: Progress and challenges with Warn-on-Forecast. *Atmos. Res.*, **123**, 2–16, https://doi.org/10.1016/j.atmosres.2012.04.004.

Thompson, R. L., R. Edwards, J. A. Hart, K. L. Elmore, and P. Markowski, 2003: Close proximity soundings within supercell environments obtained from the Rapid Update Cycle. *Wea. Forecasting*, **18**, 1243–1261, https://doi.org/10.1175/1520-0434(2003)018<1243:CPSWSE>2.0.CO;2.

Trapp, R. J., G. J. Stumpf, and K. L. Manross, 2005: A reassessment of the percentage of tornadic mesocyclones. *Wea. Forecasting*, **20**, 680–687, https://doi.org/10.1175/WAF864.1.

Van der Walt, S., J. L. Schonberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, 2014: Scikit-image: Image processing in Python. *PeerJ*, **2**, e453, https://doi.org/10.7717/peerj.453.

Verrelle, A., D. Ricard, and C. Lac, 2015: Sensitivity of high-resolution idealized simulations of thunderstorms to horizontal resolution and turbulence parameterization. *Quart. J. Roy. Meteor. Soc.*, **141**, 433–448, https://doi.org/10.1002/qj.2363.

Weisman, M. L., W. C. Skamarock, and J. B. Klemp, 1997: The resolution dependence of explicitly modeled convective systems. *Mon. Wea. Rev.*, **125**, 527–548, https://doi.org/10.1175/1520-0493(1997)125<0527:TRDOEM>2.0.CO;2.

——, C. A. Davis, W. Wang, K. W. Manning, and J. B. Klemp, 2008: Experiences with 0-36-h convective forecasts with the WRF-ARW model. *Wea. Forecasting*, **23**, 407–437, https://doi.org/10.1175/2007WAF2007005.1.

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**, 1795–1817, https://doi.org/10.1175/WAF-D-15-0043.1.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences.* 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

Wyngaard, J. C., 2004: Toward numerical modeling in the "terra incognita." *J. Atmos. Sci.*, **61**, 1816–1826, https://doi.org/10.1175/1520-0469(2004)061<1816:TNMITT>2.0.CO;2.

Yokota, S., H. Niino, H. Seko, M. Kunii, and H. Yamauchi, 2018: Important factors for tornadogenesis as revealed by high-resolution ensemble forecasts of the Tsukuba supercell tornado of 6 May 2012 in Japan. *Mon. Wea. Rev.*, **146**, 1109–1132, https://doi.org/10.1175/MWR-D-17-0254.1.

Yussouf, N., and K. H. Knopfmeier, 2019: Application of the Warn-on-Forecast system for flash flood producing heavy convective rainfall events. *Quart. J. Roy. Meteor. Soc.*, **145**, 2385–2403, https://doi.org/10.1002/qj.3568.

——, T. A. Jones, and P. S. Skinner, 2020: Probabilistic high-impact rainfall forecasts from landfalling tropical cyclones using Warn-on-Forecast system. *Quart. J. Roy. Meteor. Soc.*, **146**, 2050–2065, https://doi.org/10.1002/qj.3779.