The Vice and Virtue of Increased Horizontal Resolution in Ensemble Forecasts of Tornadic Thunderstorms in Low-CAPE, High-Shear Environments

JOHN R. LAWSON,[a,b] COREY K. POTVIN,[a,b] PATRICK S. SKINNER,[a,b] AND ANTHONY E. REINHART[a,b]

[a] *Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma*
[b] *NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma*

(Manuscript received 27 August 2020, in final form 7 December 2020)

ABSTRACT: Tornadoes have Lorenzian predictability horizons $O(10)$ min, and convection-allowing ensemble prediction systems (EPSs) often provide probabilistic guidance of such events to forecasters. Given the $O(0.1)$-km length scale of tornadoes and $O(1)$-km scale of mesocyclones, operational models running at horizontal grid spacings (Δx) of 3 km may not capture narrower mesocyclones (typical of the southeastern United States) and certainly do not resolve most tornadoes per se. In any case, it requires $O(50)$ times more computer power to reduce Δx by a factor of 3. Herein, to determine value in such an investment, we compare two EPSs, differing only in Δx (3 vs 1 km), for four low-CAPE, high-shear cases. Verification was grouped as 1) deterministic, traditional methods using pointwise evaluation, 2) a scale-aware probabilistic metric, and 3) a novel method via object identification and information theory. Results suggest 1-km forecasts better detect storms and any associated rapid low- and midlevel rotation, but at the cost of weak–moderate reflectivity forecast skill. The nature of improvement was sensitive to the case, variable, forecast lead time, and magnitude, precluding a straightforward aggregation of results. However, the distribution of object-specific information gain over all cases consistently shows greater average benefit from the 1-km EPS. We also reiterate the importance of verification methodology appropriate for the hazard of interest.

KEYWORDS: Tornadogenesis; Thunderstorms; Ensembles; Forecast verification/skill

1. Introduction

Tornadoes are dangerous in both their severity and short predictability horizons. Significant tornadoes (rated EF2 or higher) typically occur within regions of climatologically large convective available potential energy (CAPE) and vertical wind shear (herein shear). In the United States, significant tornadoes occur most often in ''alleys'' across the southeastern and central Great Plains of the continental United States (CONUS; Brooks et al. 2003; Markowski and Richardson 2010; Dixon et al. 2011; Anderson-Frey et al. 2019); however, tornadoes of all strengths occur in atmospheric states with more modest CAPE. Sherburn and Parker (2014, their Fig. 1c) found this subset of tornadoes are relatively more common in the U.S. Southeast. As in Sherburn and Parker (2014), we term these high-shear, low-CAPE environments as HSLC regimes. Multiple fatalities associated with tornadoes within this low-predictability regime have motivated studies of tornadic thunderstorms in this environment. This effort aims to augment understanding in both social and meteorological fields, with the goal of improving the accuracy and communication of tornado forecasts and warnings in the U.S. Southeast. The fruits of this research will benefit not only that region: the present study focuses on improving numerical weather prediction (NWP) model skill in HSLC regimes across the United States. Specifically, the goal is to improve NWP guidance, yielding longer lead times, reduction of false-alarm rates, and increased hit rates for warnings issued by the National Weather Service. This aligns with research priorities of Warn-on-Forecast (WoF; Stensrud et al. 2009) and the Verification of the Origins of Rotation in Tornadoes Experiment-Southeast (VORTEX-SE).

A disproportionately large number of deadly tornadoes in the United States are associated with supercell thunderstorms, with an order of magnitude fewer occurring within quasi-linear convective systems (QLCSs) or with disorganized convective cells (Schoen and Ashley 2011). Despite the lower frequency of nonsupercellular tornadoes (Smith et al. 2012), their rapid formation poses a different forecast challenge to supercellular tornadoes (Kis and Straka 2010) and warning-verification skill is poorer (Brotzge and Erickson 2010); hence, we balance our focus herein between QLCS and supercellular tornado detection.

Forecasts of severe weather within low-shear flow (including the HSLC regime) have the lowest skill (Herman et al. 2018; Anderson-Frey et al. 2019). Supercells in the HSLC regime are associated with narrower updrafts than those storms occurring with higher CAPE ($\gtrsim$1000 J kg^{-1}; e.g., Markowski and Straka 2000) and are therefore more susceptible to the deleterious effect of dry-air entrainment (James and Markowski 2010; Kirkpatrick et al. 2011). To compound this lower predictability, supercell development is sensitive to nuances captured by neither bulk measures of the atmospheric regime (Lawson 2019) nor the undersampled observed atmosphere. Further, the location, timing, and magnitude of severe hazards are sensitive to NWP horizontal grid spacing (Δx; Potvin and Flora 2015).

Potvin's and Reinhart's current affiliation: NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma.

Potvin's and Skinner's current affiliation: School of Meteorology, University of Oklahoma, Norman, Oklahoma.

Corresponding author: John R. Lawson, john@jrl.ac

DOI: 10.1175/MWR-D-20-0281.1

© 2021 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 09/02/21 02:14 PM UTC

TABLE 1. Latitude–longitude domain dimensions for each case (rows) and EPS Δx (columns).

Case	Code	3 km	1 km
31 Mar 2016	A-20160331	250 × 250	322 × 322
1 May 2017	B-20170501	221 × 221	322 × 322
2 May 2017	C-20170502	251 × 251	322 × 322
4 May 2017	D-20170504	251 × 251	322 × 322

As with any partly chaotic system (Lorenz 1963; Zeeman 1979; Williams 1997), tipping points occur in phase space and result in low predictability of the severe–subsevere discrimination (Coffer and Parker 2018; Lawson 2019). This low predictability of thunderstorm evolution is not only an obstacle to predicting the occurrence, timing, and location of supercells, but also an upper bound on already-short practical predictability of associated damaging phenomena such as tornadoes.

Storm size is a function of CAPE and 0–6-km shear, among other bulk atmospheric diagnostics (Lawson 2019). A smaller storm will require the NWP model to be run at a smaller Δx for it to be detected sufficiently, and we might assume a useful forecast of tornadogenesis is more likely when its parent supercell is well resolved. While we cannot explictly forecast tornadoes at $\Delta x \geq 1$ km—as the effective resolution is $\sim 6\Delta x$ (Skamarock 2004)—we discuss tornadogenesis in terms of products derived from explicit prediction of the parent thunderstorm. Models operational at the time of writing typically run at 3-km Δx: some as convection-allowing ensemble prediction systems (EPSs); others as deterministic tools [e.g., the High-Resolution Rapid Refresh model (HRRR); Benjamin et al. (2016)]. The choice of $\Delta x = 3$ km is a trade-off between

TABLE 2. The WRF mixed-physics configuration, chosen to mimic the parent NEWSe/WoFS configuration as closely as possible. The cumulus scheme was only used during WoFS data assimilation to generate the ICs and LBCs for the present study. The GEFS member refers to the IC and LBC dataset used by each member in the parent HRRRe. (Members m19–m36 were created by applying the configuration in reverse order to the first 18 sets of ICs and LBCs.) The microphysics was Thompson for 2016 cases and NSSL 2-moment for 2017 cases.

Member	PBL	SW	LW	Surface layer	Cu scheme	GEFS member
m01	YSU	Dudhia	RRTM	MM5	Kain–Fritsch	p01
m02	YSU	RRTMG	RRTMG	MM5	Kain–Fritsch	p02
m03	MYJ	Dudhia	RRTM	Eta	Kain–Fritsch	p03
m04	MYJ	RRTMG	RRTMG	Eta	Kain–Fritsch	p04
m05	MYNN	Dudhia	RRTM	MYNN	Kain–Fritsch	p05
m06	MYNN	RRTMG	RRTMG	MYNN	Kain–Fritsch	p06
m07	YSU	Dudhia	RRTM	MM5	Grell	p07
m08	YSU	RRTMG	RRTMG	MM5	Grell	p08
m09	MYJ	Dudhia	RRTM	Eta	Grell	p09
m10	MYJ	RRTMG	RRTMG	Eta	Grell	p10
m11	MYNN	Dudhia	RRTM	MYNN	Grell	p11
m12	MYNN	RRTMG	RRTMG	MYNN	Grell	p12
m13	YSU	Dudhia	RRTM	MM5	Tiedtke	p13
m14	YSU	RRTMG	RRTMG	MM5	Tiedtke	p14
m15	MYJ	Dudhia	RRTM	Eta	Tiedtke	p15
m16	MYJ	RRTMG	RRTMG	Eta	Tiedtke	p16
m17	MYNN	Dudhia	RRTM	MYNN	Tiedtke	p17
m18	MYNN	RRTMG	RRTMG	MYNN	Tiedtke	p18
m19	YSU	Dudhia	RRTM	MM5	Kain–Fritsch	p18
m20	YSU	RRTMG	RRTMG	MM5	Kain–Fritsch	p17
m21	MYJ	Dudhia	RRTM	Eta	Kain–Fritsch	p16
m22	MYJ	RRTMG	RRTMG	Eta	Kain–Fritsch	p15
m23	MYNN	Dudhia	RRTM	MYNN	Kain–Fritsch	p14
m24	MYNN	RRTMG	RRTMG	MYNN	Kain–Fritsch	p13
m25	YSU	Dudhia	RRTM	MM5	Grell	p12
m26	YSU	RRTMG	RRTMG	MM5	Grell	p11
m27	MYJ	Dudhia	RRTM	Eta	Grell	p10
m28	MYJ	RRTMG	RRTMG	Eta	Grell	p09
m29	MYNN	Dudhia	RRTM	MYNN	Grell	p08
m30	MYNN	RRTMG	RRTMG	MYNN	Grell	p07
m31	YSU	Dudhia	RRTM	MM5	Tiedtke	p06
m32	YSU	RRTMG	RRTMG	MM5	Tiedtke	p05
m33	MYJ	Dudhia	RRTM	Eta	Tiedtke	p04
m34	MYJ	RRTMG	RRTMG	Eta	Tiedtke	p03
m35	MYNN	Dudhia	RRTM	MYNN	Tiedtke	p02
m36	MYNN	RRTMG	RRTMG	MYNN	Tiedtke	p01

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 09/02/21 02:14 PM UTC

TABLE 3. Sources of observational and forecast datasets, and their interpolated grids to enable forecast verification. Letter codes are N (native grid), N-I (close to native; grids misaligned, but interpolated), I (interpolated to a coarser grid), and C-N (native, but cut to the 1-km domain). Note MRMS azimuthal shear is valid within the 0–2- and 2–5-km layers (on identical grids). All grids were trimmed so their geographical extents were as close to coherence (with the 1-km domain) as possible.

Dataset	Variable(s)	Native Δx	1 km	3 km
ISU NEXRAD	Composite reflectivity	~1 km	N-I	I
MRMS	Azimuthal shear	~1 km (2016); ~0.5 km (2017)	N-I; I	I
3-km EPS	Composite reflectivity, UH_{25}, UH_{02}	3 km		C-N
1-km EPS	Composite reflectivity, UH_{25}, UH_{02}	1 km	N	

Δx, ensemble membership, and lead time, with Δx = 3 km a reasonable upper bound for representing deep-convection-related hazards (Potvin and Flora 2015).

A decrease in Δx may also benefit the forecast of vortices within QLCSs: so-called mesovortices (Weisman and Trapp 2003; Trapp and Weisman 2003; Smart and Browning 2009). These mesovortices are typically brief and may yield tornadoes less significant in their damage characteristics (i.e., EF0 or EF1). However, some rotation tracks may persist for much longer and/or cause more significant damage. While updraft width is positively correlated with tornado strength (Trapp et al. 2017), this does not preclude strong tornadoes occurring in tandem with narrow updrafts.

This discussion of Δx is set against the backdrop of practical limits in high-performance-computing resources. The increase of ensemble membership by one linearly increases the computing resources required to run the new system in the same wall-clock time as the old system (neglecting bottlenecks or latency of data transmission). Conversely, reducing Δx from 3 to 1 km requires $O(50)$ times the computing resources due to 1) a threefold increase in calculations performed in the latitude and longitude dimensions and 2) associated reduction in time-step length to avoid a fatal Courant–Friedrichs–Lewy condition violation. (We neglect an increase in vertical resolution that may accompany a decrease in Δx). This competition for resources at the convection-allowing scale may be better spent on appropriate ensemble reliability instead of Δx reduction (Schwartz et al. 2017; Loken et al. 2017).

Herein, we test the hypothesis that decreasing Δx from 3 to 1 km improves the detection of supercells and potentially tornadogenesis in HSLC regimes as found by Schwartz and Sobash (2019) and Sobash et al. (2019) in a multiyear deterministic-forecast study. The methodology herein differs from those studies by evaluating forecast skill in the first 3 h (rather than the first 36 h) with an object-based probabilistic technique grounded in information theory that is more suitable for rewarding the detection of infrequent strong rotation. Using two EPSs, we 1) detect the faint signal of tornadogenesis in forecast and observational data, 2) amplify this signal above the noise (i.e., uncertainty) by filtering EPS output with one of three methods, before 3) evaluating and analyzing the relative performances of both EPSs in detecting mesocyclone occurrence

TABLE 4. Tornado reports within the 1-km spatial and temporal domain for each case. The event number is given by NWS (see text for source). Time is in UTC. Warning lead time is in minutes. Latitude and longitude are in degrees. A nearby location is provided for the reader's reference. The state abbreviations are Mississippi (MS), Alabama (AL), Pennsylvania (PA), and Georgia (GA). The NWS weather forecast office (WFO) abbreviations are Jackson (JAN), Birmingham (BMX), Pittsburgh (PBZ), State College (CTP), and Peachtree City (FFC).

Case	Event	Time	Lat	Lon	Warn time	Location reference	County	State	WFO	Rating
A-20160331	624183	2326	33.44	−88.32	9	Columbus	Lowndes	MS	JAN	EF1
	626734	2332	33.46	−88.28	12	Stafford	Pickens	AL	BMX	EF1
	626735	2343	33.47	−88.16	24	Ethelsville	Pickens	AL	BMX	EF1
	626738	0003	33.80	−87.85	7	Bluff	Fayette	AL	BMX	EF1
	626742	0027	33.87	−87.68	28	Bazemore	Fayette	AL	BMX	EF1
	626744	0104	33.66	−87.51	6	Alta	Fayette	AL	BMX	EF0
B-20170501	702457	1831	40.91	−80.07	2	Prospect	Butler	PA	PBZ	EF0
	702458	1856	41.09	−79.73	6	Eldorado	Butler	PA	PBZ	EF0
	702459	1909	41.19	−79.55	19	Beaver	Clarion	PA	PBZ	EF0
	702460	1926	41.33	−79.25	7	Turkey Ridge	Clarion	PA	PBZ	EF0
	702463	1927	41.34	−79.22	8	Cooksburg	Clarion	PA	PBZ	EF0
	702465	1927	41.41	−79.36	8	Frills Corner	Clarion	PA	PBZ	EF1
	741697	1930	41.43	−79.34	11	Gultonville	Clarion	PA	PBZ	EF1
	683366	2008	41.6	−78.72	21	Dohoga	Elk	PA	CTP	EF1
	683281	2244	40.94	−77.49	0	Rebersburg	Centre	PA	CTP	EF1
C-20170502	—									
D-20170504	701358	2012	33.65	−84.41	0	Atlanta (KATL)	Fulton	GA	FFC	EF0
	701370	0142	33.79	−84.24	2	Avondale Estates	De Kalb	GA	FFC	EF0

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 09/02/21 02:14 PM UTC

TABLE 5. Cases in the present study. Columns describe the nominal date (the universal day of first initialization), initialization times, U.S. states in the domain, the case label, and a brief description of the primary convective modes (cf. Fig. 2).

Case	Date	Initialization times (UTC)	States in 1-km domain	Primary modes
A-20160331	31 Mar 2016	1900, 2000, 2100, 2200, 2300	Arkansas, Louisiana, Mississippi, Alabama	Complex
B-20170501	1 May 2017	1900, 2000, 2100, 2200, 2300	West Virginia, Pennsylvania, Maryland	Linear
C-20170502	2 May 2017	2300, 0000, 0100, 0200, 0300	Oklahoma, Texas, Kansas	(Weak) cellular
D-20170504	4 May 2017	2200, 2300, 0000, 0100, 0200	Alabama, Georgia, South Carolina	Linear and cellular

and hence potentially tornadogenesis. We also perform the above process for identifying and verifying thunderstorms in composite reflectivity with a focus on detection of pertinent storm characteristics.

2. Methodology of signal detection

Circulations $\gtrsim$15 km in diameter can be resolved to a degree by Δx = 3 km (Potvin and Flora 2015): i.e., not tornadoes. As such, we employ proxies in observational and NWP data for capturing the typical tornadogenetic atmospheric process by identifying its necessary precursor: mesocyclogenesis (e.g., Markowski and Richardson 2010). For forecasts, we compute updraft helicity (UH; derivation and discussion in Kain et al. 2008) as a proxy for azimuthal wind shear. Trapp et al. (2005) found azimuthal shear in the 0–2-km layer could serve as a proxy for mesocyclone detection and potentially tornadogenesis. The connection between tornadogenesis and mesocyclone detection is weaker at 2–5 km, but farther from the ground, this layer is less noisy (Sobash et al. 2019). Herein, we correspondingly present UH in the 0–2-km (UH_{02}) and 2–5-km (UH_{25}) layers.

a. Ensemble design

We used the Weather Research and Forecasting system (WRF; Powers et al. 2017), version 3.9.1, for all numerical simulations. Each simulation's Δx = 3-km parent domain is geographically identical to the Warn-on-Forecast System (WoFS; formerly NEWSe; Wheatley et al. 2015), relocated once per universal day between cases. Unavoidably, this domain's latitude–longitude dimensions also changed between cases (see Table 1). Within the parent domain, we embedded a 1-km 322 × 322 nest with one-way feedback, inheriting the interpolated initial conditions (ICs) from the parent domain. The grid locations were prescribed manually to balance factors of 1) distance of convective initiation from the lateral boundary, 2) the ability to capture the majority of moist convection in the parent-domain area, and 3) avoidance of steep terrain. Lateral boundary conditions (LBCs) for both domains were introduced hourly, taken from the original WoFS member's dataset. We use 36 ensemble members, each initialized and forced by the analogous WoFS ICs and LBCs, respectively. These WoFS members are described by Table 2. For all runs, 51 vertical levels were used and stacked more tightly closer to the ground to improve the resolution of the planetary boundary layer. Model top was specified as 20 hPa.

The mixed-parameterization configuration for both ensembles was chosen to mimic the quasi-operational WoFS (Wheatley et al. 2015). This diversity increases the ensemble spread to address some model-based uncertainty. The drawback of ensemble diversity of parameterizations is the unique model climates for each member, which reduces the likelihood of capturing the (notional) true probability distribution. However, it is often a straightforward and pragmatic way to improve the reliability and/or discrimination of EPSs (Berner et al. 2015, 2017; Clark et al. 2018).

The WoFS is quasi-operational during the Spring Forecast Experiment in Norman, Oklahoma; more information can

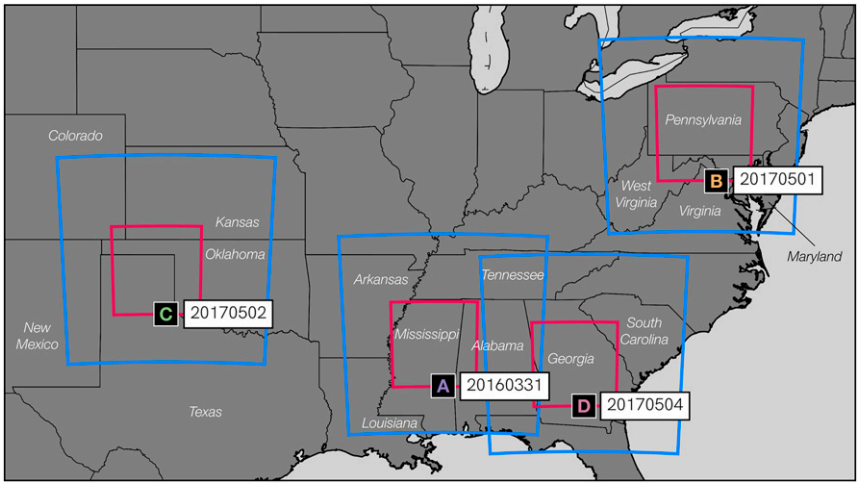

FIG. 1. The forecast domains for 3- (blue) and 1-km (red) EPSs. Each case is labeled as discussed in text. Some states are labeled for context.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 09/02/21 02:14 PM UTC

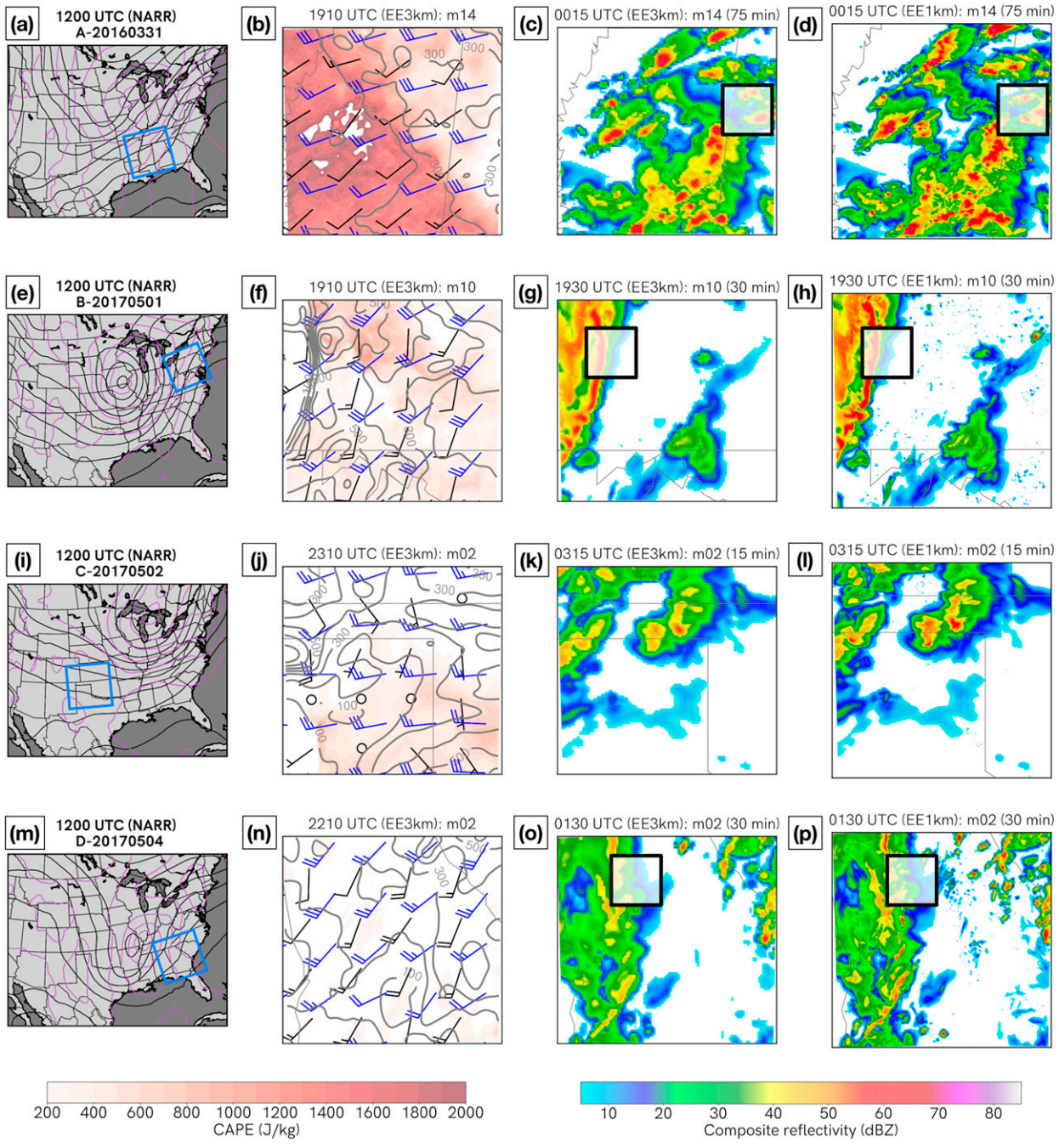

FIG. 2. Representative model analyses/forecasts for the first initialization time for each case. The rows indicate each case. (a),(e),(i),(m) The 1200 UTC NARR (3-km domain in blue shown for reference). (b),(f),(j),(n) The 3-km 10-min forecasts (quasi-analyses) of CAPE (red contour-fill, denoted by scale at bottom), 0–1-km shear (black barbs), and 0–6-km shear (blue barbs), taken from the member closest to the mean. Gray lines mark 3-km 0–3-km storm-relative helicity, contoured every 100 $m^2 s^{-2}$ and labeled every 200 $m^2 s^{-2}$. Also shown are (c),(g),(k),(o) 3- and (d),(h),(l),(p) 1-km composite reflectivity, respectively, valid at representative times from the same member as in (b), (f), (j), and (n). This is plotted as pixels colored continuously (i.e., not contour-filled with a 5-dBZ granularity, as is typical) to appropriately assess differences. The black boxes indicate regions of observed tornadic activity ±30 min of the valid time.

be found at https://hwt.nssl.noaa.gov/spring_experiment (accessed 1 July 2019). For each day, initial (1800 UTC) analyses are generated with ICs and LBCs from the ensemble HRRR (HRRRe; Dowell et al. 2016). The geographical location of the WoFS domain was chosen by consensus of experiment participants to maximize the likelihood of capturing that day's most severe (convective) weather. Further information on the advanced assimilation scheme can be found in Wheatley et al. (2015)

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 09/02/21 02:14 PM UTC

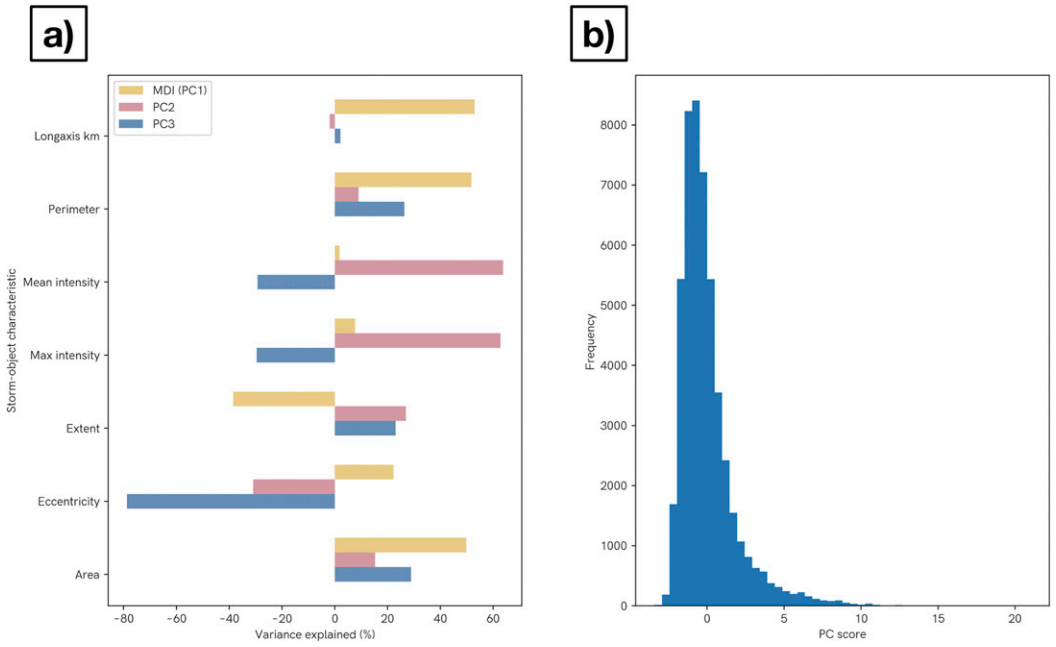

FIG. 3. Breakdown of morphology discrimination for the 3-km domain, as an example. (a) An explanation of variance, after principal component analysis of all objects, by the leading three principal components (of which the first is MDI), and (b) the distribution of MDI values, for all objects identified in the same domain.

and Wang et al. (2019), while WoFS's application and performance is assessed in, e.g., Skinner et al. (2018), Lawson et al. (2018), Jones et al. (2018), Flora et al. (2019), and Potvin et al. (2020). Datasets were interpolated as in Table 3.

b. *Observational datasets*

Throughout, we assume observational error is unbiased, normally distributed, and spatiotemporally uncorrelated [as in Snyder and Zhang (2003)]. This is necessitated by the lack of suitable uncertainty estimates.

We verify rotation with the Multi-Radar Multi-Sensor (MRMS) development system at CIMMS/NSSL. MRMS blends radars across the CONUS into a seamless gridded dataset (Smith et al. 2016). This verification dataset was processed to match the (original WoFS) 3-km domain using only the assimilated radars. The azimuthal wind shear (AzShear)

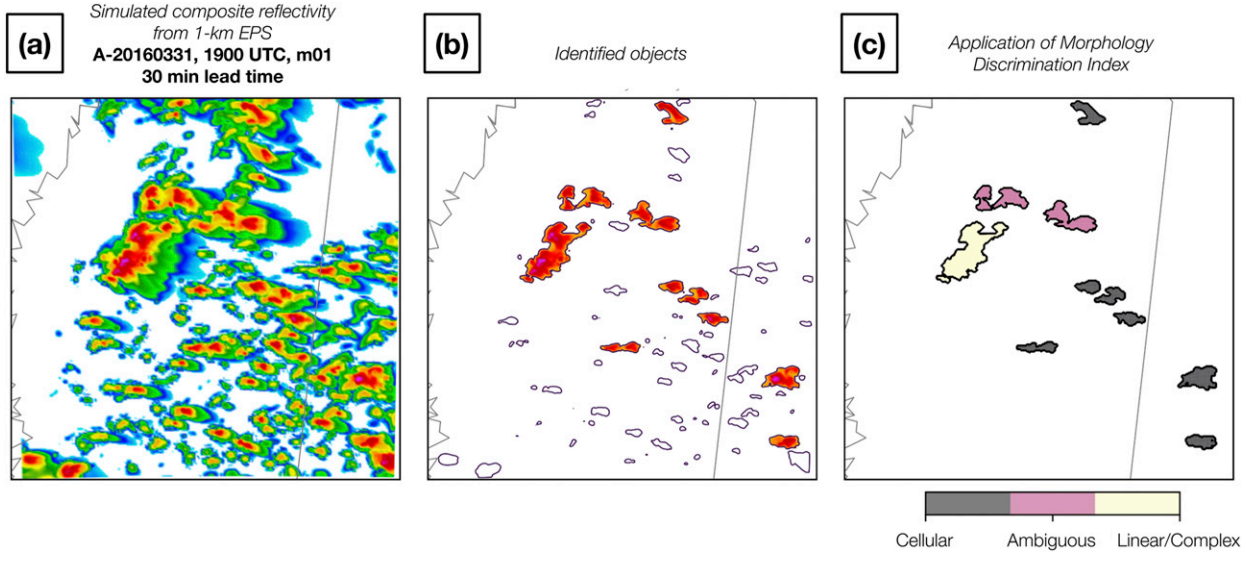

FIG. 4. An illustration of the object-identification and MDI methodologies. Taken from the first member of the 1-km EPS. Valid at 1930 UTC 31 Mar 2016 (a 30-min forecast). (a) The simulated composite reflectivity at this time, (b) all objects (black contours) with criteria-meeting identified objects filled with simulated dBZ, while (c) each object after MDI computation, where the morphology group for each object is denoted by the legend below.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 09/02/21 02:14 PM UTC

TABLE 6. Percentile–magnitude lookup table for the analysis herein. Values of UH_{02} and UH_{25} are in $m^2 s^{-2}$; $AzShear_{02}$ and $AzShear_{25}$ are in s^{-1}; and composite reflectivity (simulated and observed) is in dBZ. Small differences between observation magnitudes, at different Δx, are artifacts of the interpolation scheme.

Percentile	Variable	3-km EPS	1-km EPS	Variable	Obs (3 km)	Obs (1 km)
99	UH_{02}	5.67	15.33	$AzShear_{02}$	3.6×10^{-3}	3.7×10^{-3}
99.5		9.52	24.01		4.4×10^{-3}	4.5×10^{-3}
99.9		25.97	48.65		6.7×10^{-3}	6.9×10^{-3}
99.95		35.63	57.05		8.2×10^{-3}	8.3×10^{-3}
99	UH_{25}	9.6	26.8	$AzShear_{25}$	4.2×10^{-3}	4.2×10^{-3}
99.5		17.2	50.4		5.1×10^{-3}	5.2×10^{-3}
99.9		52.4	153.2		7.6×10^{-3}	7.8×10^{-3}
99.95		75.2	212.4		9.0×10^{-3}	9.2×10^{-3}
70	Composite reflectivity	15.0	15.8	NEXRAD	14.9	14.5
80		23.5	24.8		23.5	23.4
90		34.7	36.2		32.9	32.8
95		41.9	43.4		38.7	38.6
96	(Object threshold)	44.3	45.7	(Object threshold)	40.5	40.5
99		52.0	53.7		49.0	49.1
99.9		63.5	65.7		57.3	57.5

product uses a linear least squares derivative approach that calculates the maximum range-corrected cyclonic azimuthal component of the horizontal shear: a proxy of rotation in Doppler radars (Smith and Elmore 2004; Miller et al. 2013; Mahalik et al. 2019). AzShear is produced in two layers (0–2 and 2–5 km AGL; $AzShear_{02}$ and $AzShear_{25}$, respectively) and output every 5 min on a WoFS-matched 0.01° (2016; ~1 km) and 0.005° (2017; ~0.5 km) domain. No further postprocessing to AzShear was performed: the verification methodologies were chosen to filter noise, as discussed in forthcoming sections.

To verify forecasts of composite reflectivity, we use NEXRAD Level III archives stored at Iowa State University (ISU; https://mesonet.agron.iastate.edu/docs/nexrad_composites/, accessed 1 January 2019). For creation of the dataset, base reflectivity data are composited before suspected false echoes are removed through comparison with the Net Echo Top product.

We gathered filtered (i.e., quality-controlled) reports of tornado and >2.54-cm (>1-in.) hail observations from archives at https://verification.nws.noaa.gov/services/public/index.aspx (accessed 1 January 2019). This dataset is generated by removing potential duplicates or erroneous reports. Reports for each case are detailed in Table 4.

c. *Analyses and reanalyses*

For each case's context on the synoptic (CONUS) scale, we plot North American Regional Reanalysis (NARR) data for geopotential-height (Z) "truth." On the meso-γ scale, we employ $\Delta x = 3$ km model output at a forecast time of 10 min as a quasi-analysis. This is assumed close to the observed state due to its rapid and advanced ensemble data assimilation (Wheatley et al. 2015; Jones et al. 2016), while allowing for mild spinup inertia (e.g., some fields are zero at initialization).

d. *Cases*

Four cases in 2016 and 2017 were chosen for their demonstration of supercellular and/or QLCS activity within (generally) HSLC environments (Table 5; Fig. 1). Three cases include reported tornadoes, with the fourth acting as a null case. For each case, five initialization times were chosen during the most active periods, each an hour apart such that 7 h of the case were simulated. In general, statistical independence is not assumed during analysis. The authors will focus on a relatively small case sample size due to 1) the wish to compare subjective interpretations with various traditional and recent verification techniques, 2) the relatively small number of suitable cases in the WoFS archive, and 3) the computational expense of object identification and associated verification. A rigorous and comprehensive climatological study is outside the scope of the present manuscript, requiring a larger dataset (e.g., Schwartz et al. 2019; Schwartz and Sobash 2019; Sobash et al. 2019).

The four cases are abbreviated with a letter mnemonic (YYYYMMDD) and briefly described in the following subsections. Figure 2 presents diagnostic and forecast variable fields for each case. CAPE is computed with the lowest-100-hPa method (Blumberg et al. 2017). Storm-relative helicity [SRH; Kerr and Darkow 1996; Markowski et al. 1998] is shown between ground and 3 km. Shear and helicity in the 0–1-km layer may be critical for tornadogenesis (e.g., Rasmussen 2003; Aran et al. 2009; Geerts et al. 2009), hence we plot the bulk shear between 0–1 and 0–6 km alike. Further information on each tornado report is found in Table 4.

1) 31 MARCH 2016 (A-20160331)

At 1200 UTC, a strong Z trough was analyzed throughout the troposphere over the Great Plains (Fig. 2a), while a surface cold front is evident in the 925-hPa Z minimum running from northern Wisconsin toward Texas (frontal analysis by the Weather Prediction Center; https://www.wpc.ncep.noaa.gov/archives/web_pages/sfc/sfc_archive.php; not shown). Convection originates from a point within locally high CAPE values in southern Arkansas and northern Louisiana, around 1900 UTC (1400 CDT; Fig. 2b). (Simulations for this case begin at this time.) As the complex strengthens, tornado watches were issued by the SPC in advance of the eastward-moving QLCS in both Mississippi

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 09/02/21 02:14 PM UTC

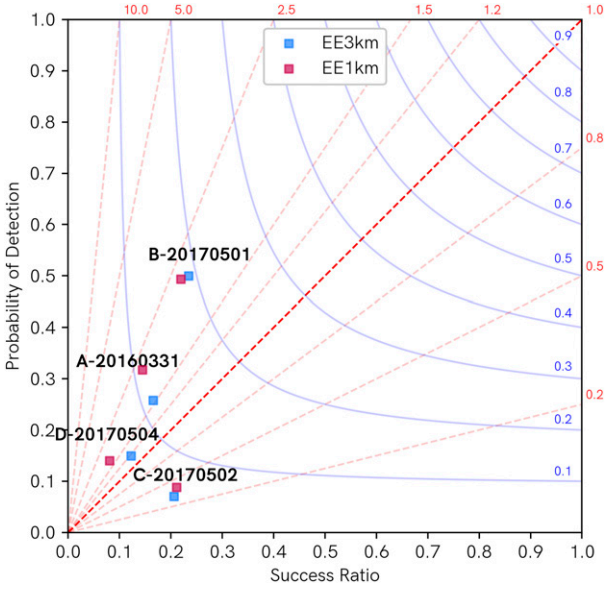

FIG. 5. Gridpointwise performance diagram of forecasted composite reflectivity, for each case (labeled) at 60-min lead time, aggregated over all members and initialization times. The diagram is formed from a contingency table detecting exceedance of 40-dBZ reflectivity. Points from the 3- and 1-km EPSs are colored blue and red, respectively. The black squares in Fig. 6 correspond to the four pairs of scatter points, for reference. Red dashed lines represent lines of constant bias; blue lines represent lines of constant critical success index.

and Alabama. By 0200 UTC (2100 CDT), tornado watches were in effect from Mississippi east to Georgia as the complex exited the area of interest. Tornadoes were reported in Mississippi and Alabama, primarily after 2300 UTC (1800 CDT), within a thunderstorm complex (EPS output near this time is shown in Figs. 2c and 2d). Tornadic supercells were observed and detected by radar near the Mississippi–Alabama border, marked by a black box in Figs. 2c and 2d.

2) 1 MAY 2017 (B-20170501)

Six hours before the first tornado reports in western Pennsylvania, the area of interest lay east of a cutoff Z minimum over Iowa at 500 hPa; there is little directional shear due to vertical stacking of Z minima (Fig. 2e). While tornadoes were observed (~1830 UTC onward; Fig. 2f), an eastward-moving QLCS marks a tight gradient in SRH. The SRH values maximize along the QLCS in the northwestern domain quadrant (Fig. 2f) concurrent with 15 m s^{-1} 0–1-km shear and 30 m s^{-1} 0–6-km shear. This region is associated with weaker (EF0, EF1) tornadoes near the black box in Figs. 2g and 2h. All observed reports in the state occurred under a SPC tornado watch. Later, another tornado was observed farther east in Rebersburg (near State College) with the same eastward-moving QLCS.

3) 2 MAY 2017 (C-20170502)

This case occurred ahead of a weak 500-hPa Z trough (Fig. 2i), while 925-hPa Z gradients were slack. No tornadoes were observed this day, but multiple hail and wind reports were reported throughout a swath of the Texas and Oklahoma Panhandles associated with supercells (not shown). This case is included as a null case to provide a balanced sampling of convective episodes.

4) 4 MAY 2017 (D-20170504)

At 1200 UTC, a negatively tilted Z trough is evident at low- and midlevels (Fig. 2m). There are two tornadoes reported within our domain of interest at 2012 and 0142 UTC (both EF0 with <3-min warning time). Between these reports, we estimate 10–20 m s^{-1} 0–1-km shear and 20–25 m s^{-1} 0–6-km shear (Fig. 2n). The second tornado was reported within a QLCS in central Georgia (Figs. 2o,p).

3. Method: Gridpointwise verification

We can form a 2 × 2 contingency table (Green and Swets 1966; Jolliffe and Stephenson 2003) as follows for verifying forecasts of composite reflectivity: 1) For each grid point in each field of simulated composite reflectivity, check the forecast value exceeds a given magnitude; 2) repeat for observations (on their corresponding grids); 3) for $\Delta x = 3$ km and $\Delta x = 1$ km we then represent verified event forecasts A as hits, unverified forecasts B as false alarms, and not-forecast observed events C as misses. We leave the final box (D) undefined herein, deferring the issue of defining not-observed, not-forecast frequencies.

These contingency tables are created from grids, not events, in contrast to scores presented in section 5. Regardless of how the contingency table is created, however, we can compute traditional metrics (e.g., Jolliffe and Stephenson 2003):

$$\mathrm{POD} = \frac{A}{A + C}, \tag{1}$$

$$\mathrm{FAR} = \frac{B}{A + B}, \tag{2}$$

$$\mathrm{bias} = \frac{A + B}{A + C}, \tag{3}$$

$$\mathrm{CSI} = \frac{A}{A + B + C}, \tag{4}$$

$$\mathrm{SR} = 1 - \mathrm{FAR}. \tag{5}$$

These metrics are probability of detection (POD), false-alarm ratio (FAR), frequency bias (hereafter bias), critical success index (CSI), and success ratio (SR). These traditional gridpoint-comparison metrics are only valid at the truncated (Δx) scale; hence, these scores provide a lower bound on the potential benefit of decreasing Δx in EPSs. It is inappropriate to rely exclusively on a pointwise comparison when there is operational tolerance to temporal and spatial error. This is evident to the human forecaster using EPS output running at $O(1)$ km: Lorenzian saturation has been reached at time and length scales close to those of interest but there is still useful information present regarding convective mode, storm characteristics, etc. This discrepancy between practical predictability and skill at the truncated scale—versus that of the object itself—is discussed in Potvin et al. (2017), Flora et al. (2018), and Lawson (2019), among others.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 09/02/21 02:14 PM UTC

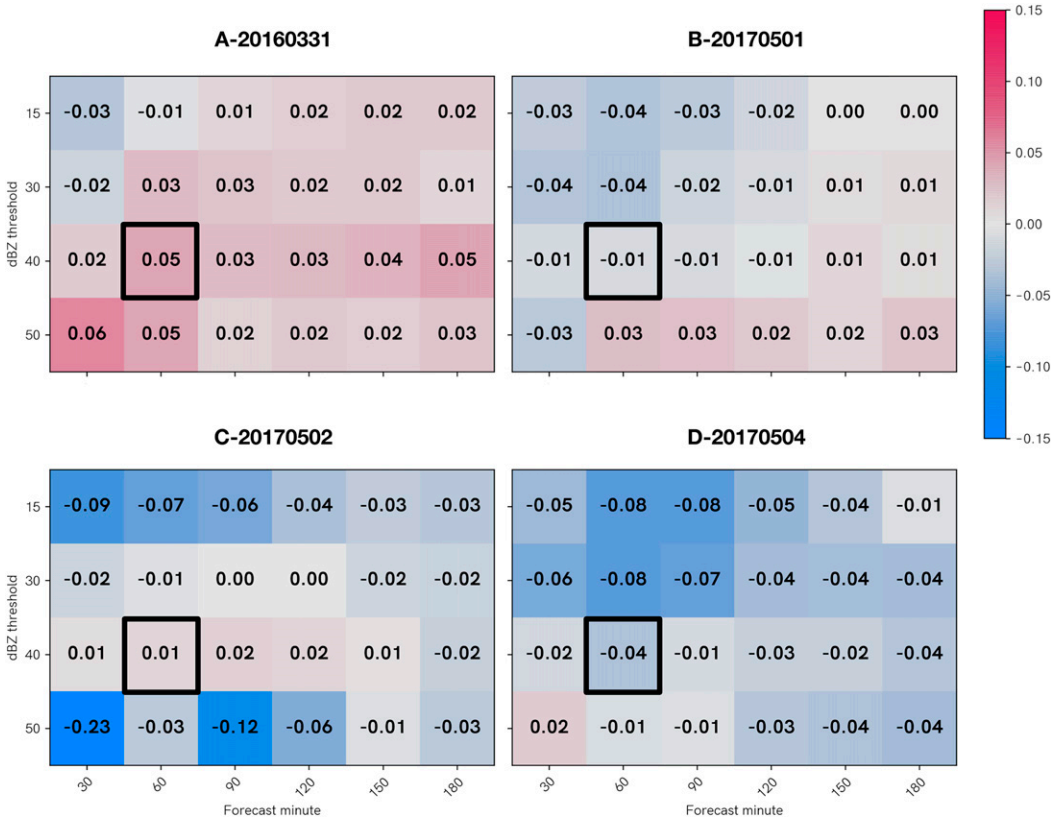

FIG. 6. EPS differences in gridpointwise composite-reflectivity verification, calculated as the vector in POD–FAR space. Color represents the model with a performance closer to optimal (i.e., the top-right corner of Fig. 5) for that threshold and forecast time, with blue indicating a better performance for the 1-km EPS, and red vice versa. The black squares denote the four scatter-point pairs shown in Fig. 5.

4. Method: Scale-aware verification

To avoid double penalties in time and space, and reduce Δx-scale noise, we employ the extended fractions skill score (Roberts and Lean 2008; Schwartz et al. 2010; Duc et al. 2013; Schwartz and Sobash 2017):

$$\mathrm{eFSS} = 1 - \frac{\mathrm{FBS}}{\mathrm{FBS}_{\mathrm{ref}}}, \tag{6}$$

where FBS is the fractions Brier score, and $\mathrm{FBS}_{\mathrm{ref}}$ is its reference forecast:

$$\mathrm{FBS} = \frac{1}{N_x N_y N_t} \sum^{N_x} \sum^{N_y} \sum^{N_t} (M - O)^2, \tag{7}$$

$$\mathrm{FBS}_{\mathrm{ref}} = \frac{1}{N_x N_y N_t} \sum^{N_x} \sum^{N_y} \sum^{N_t} (M^2 + O^2). \tag{8}$$

Above, M and O represent the four-dimensional (time, ensemble member, latitude, longitude) windows that specify the fraction of grid points within the window that exceed a given threshold (e.g., if composite reflectivity is being evaluated, the threshold is set in dBZ). The symbols N_x, N_y, and N_t are the lengths (in voxels) of the windows in the longitudinal, latitudinal, and temporal dimensions, respectively. The ensemble dimension is implicit in the above definitions; further derivations can be found in Duc et al. (2013). Herein, we compute eFSS with a fast Fourier Transform (FFT) method derived from Faggian et al. (2015), using a square kernel to reduce computational expenses [as in Roberts and Lean (2008)].

The model climatologies of both EPSs are so different that the use of magnitude thresholding would yield biased results; further, comparison of UH and AzShear is one of nonequivalent quantities. Therefore, for scale-aware and object-based results in the present paper, we show only scores related to exceeding percentiles of a given variable's distribution. Here, we are most interested in detection of rotation near the top decile of occurrence.

5. Method: Object identification and classification

As the focus of the present study is on thunderstorms—discrete phenomena in fields such as composite reflectivity—we deploy the following algorithms to reveal and categorize storm objects in forecast and observational data. The identification and matching methodologies were formed to meet the requirements of the present study, but inspired by similar object-based methods such as the use of object-event contingency tables (Skinner et al. 2018), the Structure–Amplitude–Location method (SAL; Wernli et al. 2008) and its probabilistic extension (eSAL;

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 09/02/21 02:14 PM UTC

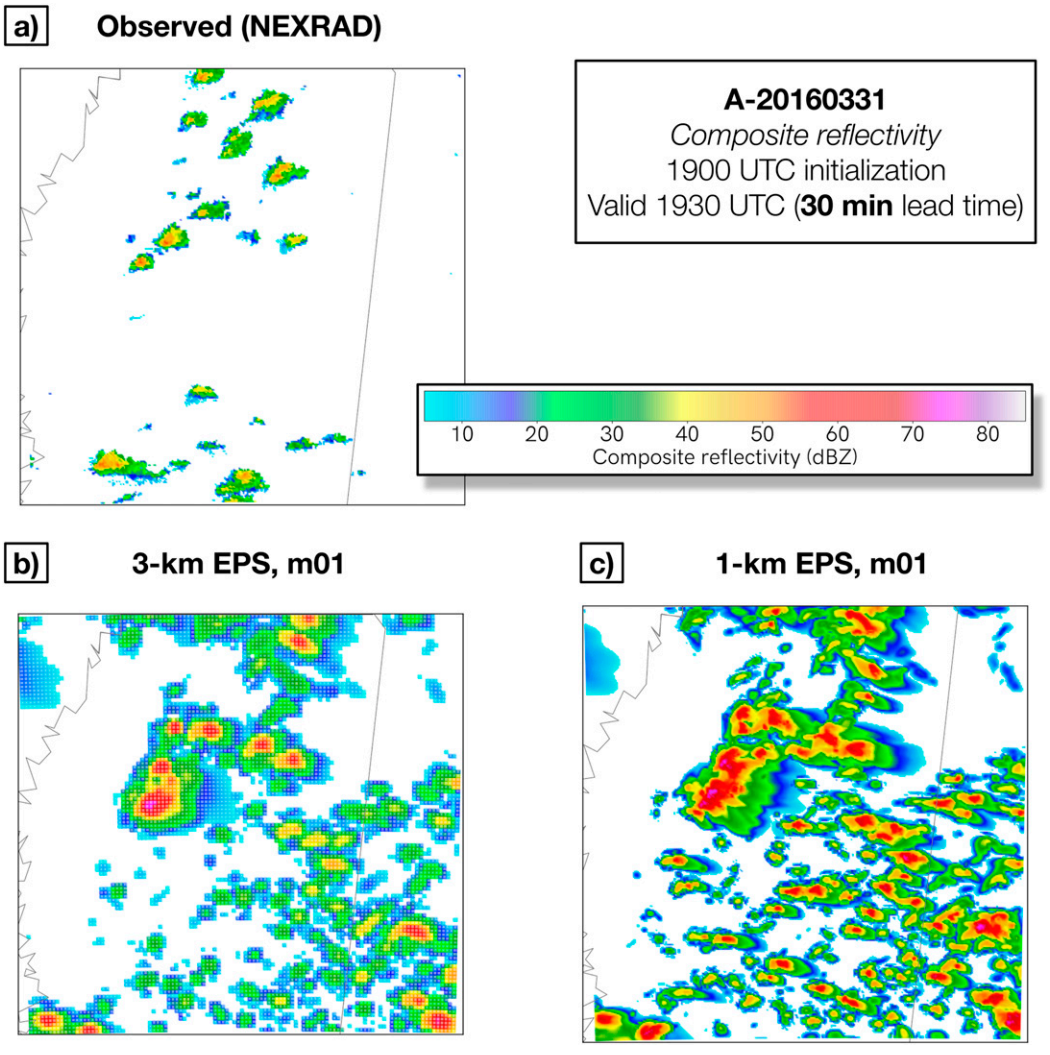

FIG. 7. An example of stronger, spurious convection seen in the 1-km domain at early lead times: (a) observed data (interpolated to the 3-km grid) and (b),(c) the 30-min 3- and 1-km forecasts, respectively, from the m01 members of the 1900 UTC runs for A-20160331.

Radanovics et al. 2018), and Method for Object-Based Diagnostic Evaluation (MODE; Davis et al. 2006).

a. *Object identification*

First, we use the Python library *scikit-image* (van der Walt et al. 2014) to identify features in the composite reflectivity field. Identified objects are dropped from the catalog if they comprise fewer than 144 ($\Delta x = 1$ km) or 48 ($\Delta x = 3$ km) pixels exceeding the 96th percentile of reflectivity for that dataset (this percentile minimizes the difference between the percentile–dBZ trends for each domain near 45 dBZ), or if the object touches the domain edge. The removal of at-edge objects is crude, but we find that most storms move through the domain during the 7 h of simulation and are identified at some point. Ramifications of at-edge object removal include risk that largest (and likely most-predictable) objects are more likely to be ignored, given their increased likelihood of touching the domain edge. Further, this is more likely to occur in the finer 1-km EPS. The area-footprint and thresholds were chosen after extensive trial-and-error, based on subjectively capturing most objects pertinent to a forecaster.

After object identification, further characteristics were computed for each object. Most of these characteristics were object properties taken directly from *scikit-image* output; others were computed manually. Some storm characteristics were computed in conjunction with other forecast fields (e.g., vertical motion W; discussed in section 9). This process was repeated over all times and all runs, with the 36 forecast members yielding ~208 000 objects, while ~2600 observed objects were identified (a number of forecast objects feature in overlapping forecast times).

b. *Object classification and principal component analysis*

After an object database was created, preliminary work revealed that most objects lay on a spectrum from cellular to

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 09/02/21 02:14 PM UTC

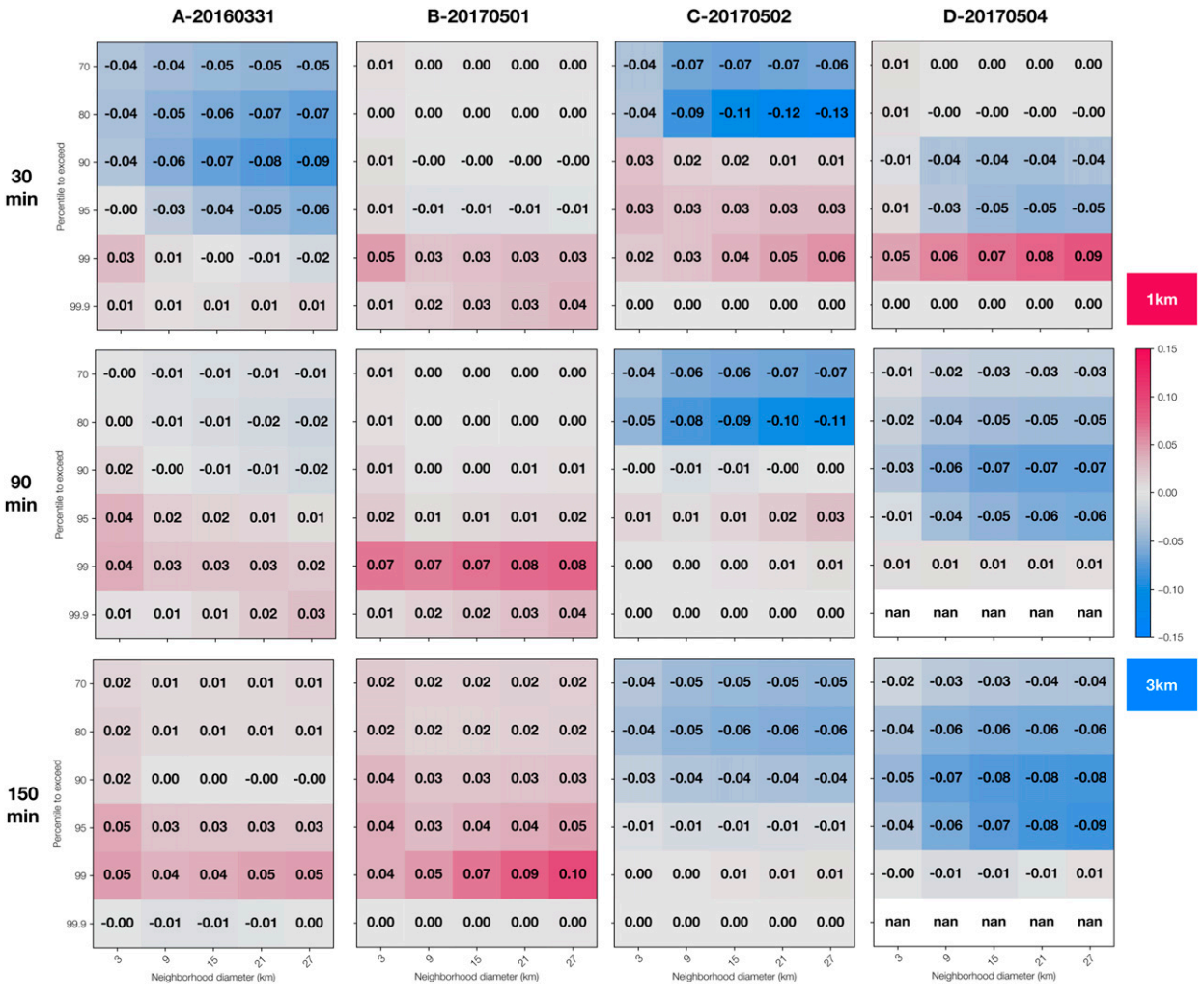

FIG. 8. EPS differences in forecast skill of composite reflectivity, as measured with eFSS, for (left to right) all four cases and (top to bottom) three representative forecast lead times. Cases are aggregated over initialization times. All eFSS computations are done with a ±10-min temporal tolerance; spatial tolerance is indicated on the *x* axes of each panel. Threshold, in terms of percentiles, is shown on the *y* axes. Not-a-number (nan) shown for partitions of insufficient sample size.

linear morphologies. To reveal this main axis of variation, we performed a principal component analysis (PCA) for each dataset. The PCA did not vary substantially between grid products (i.e., the leading principal component PC1 was a similar linear combination of factors in all four grids), suggesting the spectrum of morphologies was consistent across the four domains.

To begin the PCA, we subjectively chose seven characteristics that visually discriminated between cellular and linear convective modes. These were area, eccentricity, extent, maximum and mean composite reflectivity, perimeter, and major-axis length. While PC1 was not guaranteed to reveal an optimal discriminator between classes, in this case it represented variation in convective mode. Figure 3a shows the projection of the three leading PCs onto feature space. The interpretation of PC1 as the measure of an object's ''QLCS-ness'' fits with the strong positive correlation between object area, eccentricity, long-axis length, and perimeter length. Hence, we use this PCA model to fit and transform each object's characteristics into the leading principal component value, hereby termed Morphology Discrimination Index (MDI). Figure 3b shows the distribution of MDI values in the climatology. Given the success of PC1 in discriminating classes, we discard other PCs in further analysis.

Each identified storm was classified by its MDI value. Using trial-and-error during preliminary testing, we found that three categories were reasonably discriminated using two division thresholds. Objects with MDI values < -0.5 were typically discrete with a low (i.e., circle-like) eccentricity and were placed into a Cellular group. At >0.5, objects were often associated with a QLCS, and more occasionally, part of a thunderstorm complex. Objects in this range of MDI values were therefore denoted Linear/Complex. Finally, objects in the range $[-0.5, 0.5]$ were difficult to categorize, and as such we use a third remainder category (Ambiguous) to improve the

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 09/02/21 02:14 PM UTC

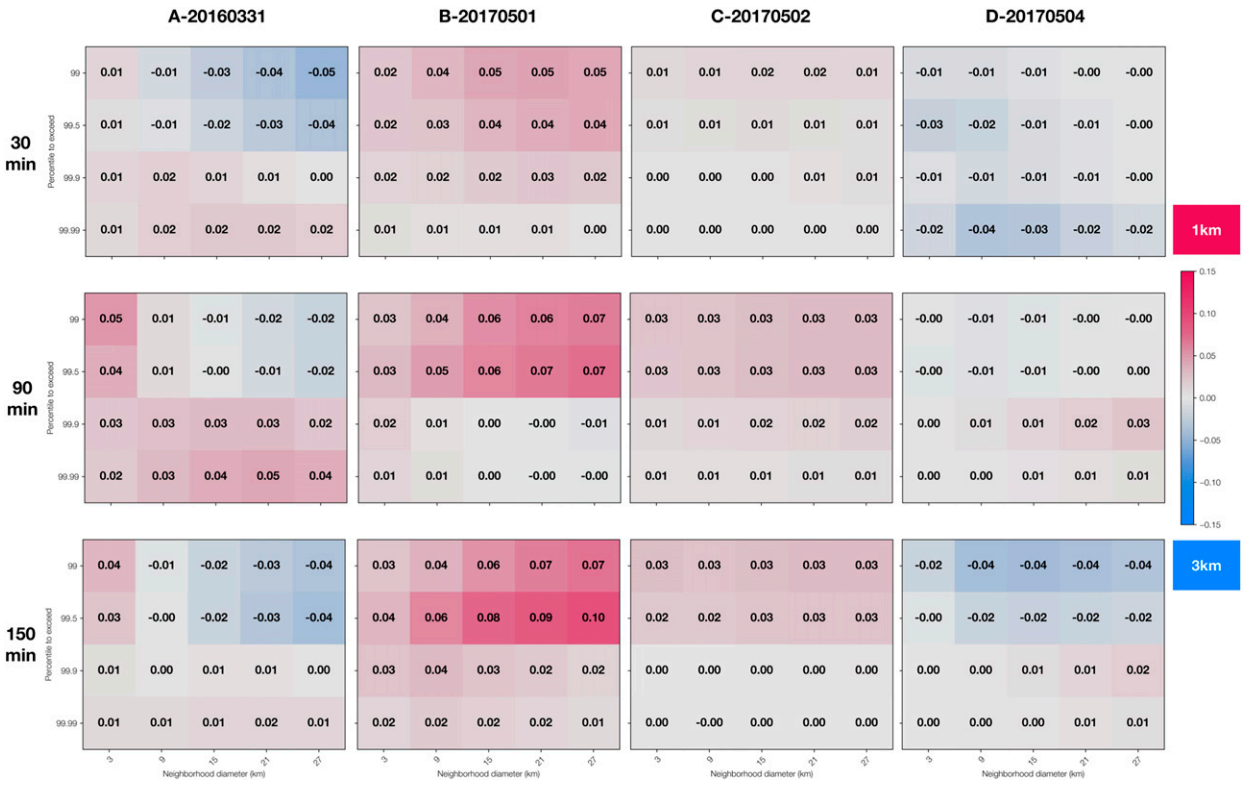

FIG. 9. As in Fig. 8, but for low-level rotation UH_{02}.

signal-to-noise ratio of our discrimination. An example of this process is shown in Fig. 4, where the overforecasting of cells is pronounced. Overforecasting dBZ in WoFS from Thompson microphysics (used in 2016 WoFS), particular in the first hour and using early HRRRe ICs, is discussed in Skinner et al. (2018). This overforecasting is reduced for 2017 cases, which used NSSL two-moment microphysics (Mansell et al. 2010; see also Table 2).

c. *Object matching*

Objects were matched using the total interest (TI) score (Davis et al. 2006) as in Skinner et al. (2018). Namely,

$$\mathrm{TI} = \frac{\Delta t_{\max} - \Delta t}{2\Delta t_{\max}}\left(\frac{\Delta c_{\max} - \Delta c}{\Delta c_{\max}} + \frac{\Delta m_{\max} - \Delta m}{\Delta m_{\max}}\right), \qquad (9)$$

where Δt represents the temporal difference between an object pair, Δc is the difference between object-pair centroids, and Δm is the smallest distance between any pixel in each of the pair's objects. The subscript *max* indicates a maximum threshold for matching two objects: we specify $\Delta t_{\max}$ (maximum permissible Δt) equal to 20 min (i.e., a 40-min window), while Δm and Δc are both set to 40 km. In the event of multiple matches between forecasted and observed objects, the maximum TI score over all relevant matched pairs dictates the chosen pair.

6. Object verification

Object-based verification methodologies have become more common (e.g., Ebert 2008; Gilleland et al. 2010) for appropriately verifying at the meso-γ scale. The conversion of gridded data to a set of objects better fits human intuition in terms of forecast verification. The caveats to object-based methodologies include an undefined score in the absence of objects and subjectivity of parameter prescription.

As an analog to the traditional performance diagrams in section 3, we form contingency tables of object occurrence rather than exceedance of a threshold. The resulting object-based performance diagrams (shown later) are a useful extension to traditional, gridpointwise performance diagrams. However, to preserve more information about the ensemble distribution, we evaluated the probabilistic forecast of rotation as follows: 1) forecast probabilities for a given observed object were generated by matching the given object to one in each forecast member, where feasible; 2) a lack of match indicated a *miss* for the (unidentified) object's event in question (e.g., its existence); 3) this yields the probability of exceeding a given rotation threshold or the probability of an object (meeting given criteria) occurring within tolerances. For rotation verification, we identified the maximum UH or AzShear value within each object's bounds, then assessed whether the forecast object exceeded four observed rotation thresholds (Table 6).

Object-specific information gain

We now use the concept of NWP models removing prior uncertainty of a situation (Roulston and Smith 2002), or conversely, a way of *gaining information* (Peirolo 2011) over a prior baseline. As the rarity and extremity of weather

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 09/02/21 02:14 PM UTC

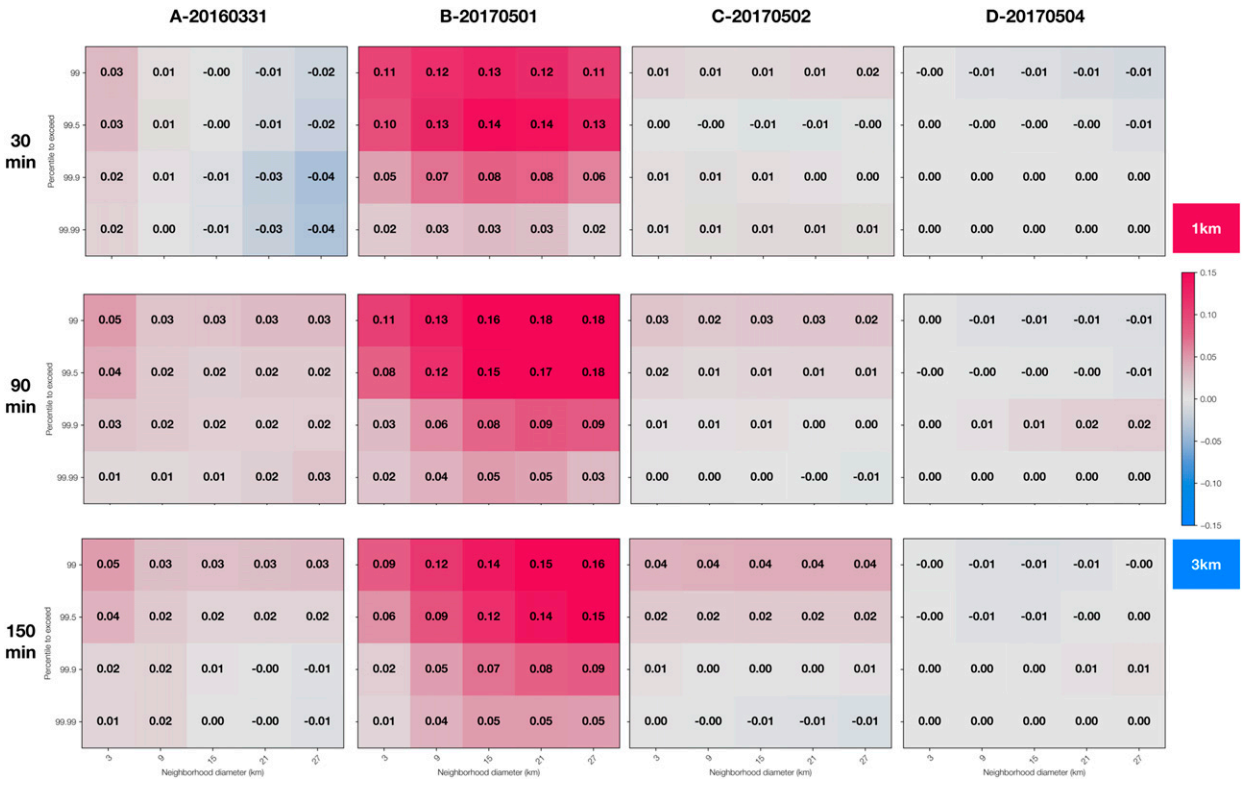

FIG. 10. As in Fig. 8, but for midlevel rotation UH_{25}.

phenomena are often correlated (Sterk et al. 2016), we employ a more appropriate way of rewarding the ensemble forecasts herein, using information theory (summarized in Pierce 1980; Cover and Thomas 2012).

In abstract terms, information theory can be interpreted as the discretization of a probabilistic system to a collection of binary choices (Shannon 1948; Gleick 2012). In a meteorological context, we evaluate the ability of both EPSs to transmit *information* from model to forecaster via (filtered) model output. This filtering is often subconsciously performed by the human forecaster—aware of model biases and typical orders of location and timing error—but this filtering is (forgivably) not always optimal (Wilson et al. 2019).[1] Herein, we mimic such filtering by identifying objects, matching them within tolerances, and evaluating the information gain for each ensemble over a naive measure of prior uncertainty.

Consider a forecaster receiving the output from a NWP model, prior to issuing a forecast for thunderstorm occurrence in the next hour. Shannon (1948) implies the forecaster's information deficit H_o of a thunderstorm object, in bits, is purely a function of a storm's rarity (representing the *surprise* of observing the event):

$$H_o = -\log_2 E(o), \tag{10}$$

where *expectation* probability $E(o)$ of the observed object is in the range [0, 1]. (This is distinct from the mathematical *expectation function*.) From this definition of *self-information*, many similar scores have been used in meteorology as the logarithm score (Good 1952), ignorance (IGN; Roulston and Smith 2002), and the Kullback–Leibler Divergence Score (Weijs et al. 2010; Ding et al. 2019). We then express the IGN of an event (i.e., object occurrence) as a function of the *forecast* probability $P(o)$ of the observed object, as an analogy of Eq. (10):

$$\mathrm{IGN}_o = -\log_2 P(o). \tag{11}$$

This is a fundamental measurement of information content or deficit, assuming optimal data compression, as discussed in Shannon (1948, p. 396). For context, an expectation of 1% yields an IGN_o of 6.64 bits (for comparison, before flipping a coin, IGN = 1 bit). The advantages of the information-theory framework include a unitless measure (as opposed to Brier-type scores) that, through its logarithm, encodes increased rewards for correct forecast of extreme and/or rare events. Indeed, when using IGN, the evaluator must bound forecast probabilities to avoid values diverging to infinity. At first glance to the meteorologist, this may appear undesirable; however, an infinite punishment matches the potentially dire consequences of decision-making predicated on absolute zero or unity certainty. Such binary (deterministic) forecasts are

[1] This suboptimality is analogous to an inefficient data compression algorithm (Shannon 1948), hence further motivating specific training as addressed in Wilson et al. (2019).

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 09/02/21 02:14 PM UTC

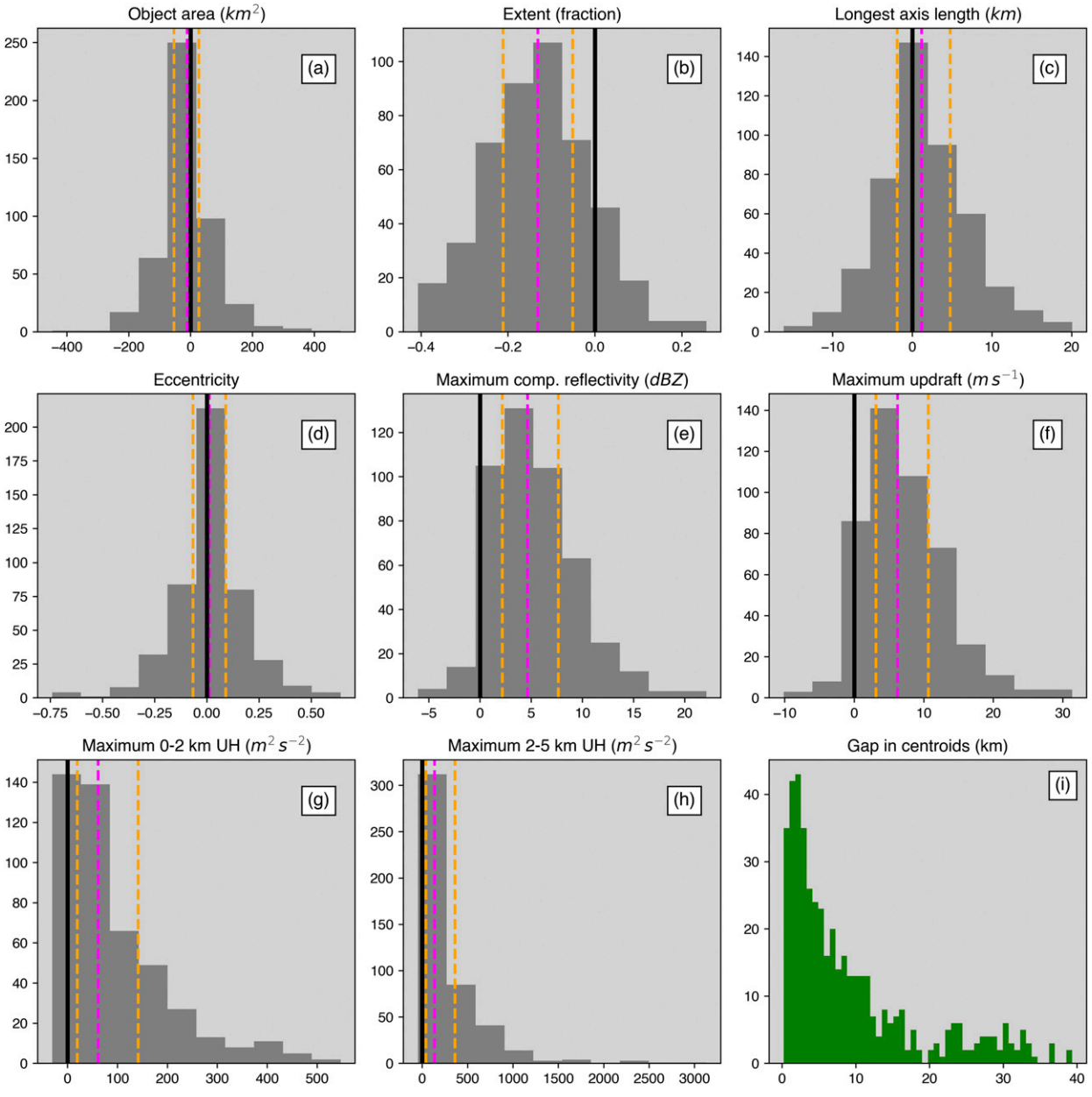

FIG. 11. Distribution of (a)–(h) EPS differences between domain-matched cellular objects (1-km minus 3-km), where the black line indicates the *x*-axis zero line, with the median marked with a magenta dashed line, and the interquartile range delineated by orange dashed lines. Hence, a positive median value may be interpreted as a typical 1-km object having a higher value of the given diagnostic. (i) The distribution of EPS object-centroid Euclidian differences (in km). The *x* axes denote magnitudes of each diagnostic, while *y* axes indicate the number of objects in a given bin.

especially egregious when ensemble undersampling is unavoidable. We finally note that the missing information represented by H and related quantities (e.g., IGN) was termed by Shannon as *information entropy*; we prefer to use the more intuitive *uncertainty* herein as argued by Ben-Naim (2008).

Because we are comparing uncertainty removal of two independent systems, and if we only consider objects that were observed (i.e., D is undefined), we can recast IGN_o as information gain (IG; e.g., Peirolo 2011) via identities as the logarithm of the ratio of forecast probability to the expectation of an object occurring:

$$\mathrm{IG}_o = \log_2\left[\frac{P(o)}{E(o)}\right]. \tag{12}$$

If we assume a blanket $E(o)$ to represent the average expectation of a cellular object occurring at a random point in the domain within the next three hours, we find an object-specific

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 09/02/21 02:14 PM UTC

information gain IG_o. A positive value represents information gained by the forecaster with a forecast in hand; negative values indicate information lost. If the distribution median of IG_o approaches zero, this represents a filtered form of Lorenzian saturation; i.e., we are evaluating the predictability of events rather than that at the truncated (Δx) scale (Flora et al. 2019). As $E(o)$ is decreased from unity to zero, there is an exponentially increasing reward of a "good" forecast (i.e., $P(o)$ is maximized). Changes to $E(o)$ do not change the overall IG_o distribution shape: only the sign and magnitude of information gained or lost. Considering the above, we set $E(o)$ as 0.1 for all cases, and bound all forecast probabilities within the range [0.01, 0.99] to avoid divergence to infinity. This caps maximum information gained or lost at 2.3 bits and −4.32 bits, respectively. This asymmetry is intuitive: the more expected an event, the less *usefully surprising* information a forecaster receives from the NWP output, and the more damaging (i.e., increasing negative IG_o) a randomly chosen object would be, on average. A preferred characteristic of IG is its *propriety*, and therefore it cannot be hedged: IG_o is maximized when observed and forecast distributions align (Peirolo 2011).

7. Results: Gridded data evaluation

In Fig. 5, we present a performance diagram (Roebber 2009) showing the SR, POD, bias, and CSI. The scatter points indicate results of exceeding 40 dBZ, at 60 min lead time for each case (this represents a potential use case for WoFS when assisting nowcasts of severe weather). A perfect forecast would be marked at (1, 1). The results are mixed and little can be said in terms of systematic advantage at each native Δx. We also show differences between the EPSs as POD–FAR vectors (Fig. 6), oriented such that red pixels represent thresholds, lead times, and cases where $\Delta x = 1$ km performance was closer to optimal. For context, each scatter pair in Fig. 5 is marked with a black box in Fig. 6. The case-to-case variability suggests improvements for $\Delta x = 1$ km may be associated with more frequent or stronger rotation, or a function of atmospheric regime, but there is an overforecasting bias of reflectivity (Fig. 7). Rather than test the previous for statistical significance at a noisy Δx scale, we defer further evaluation to the following sections that use tools more appropriate to the scales of interest. Analogous scores in the object-based framework will be shown in section 9.

8. Results: Scale-aware evaluation

The following results compare percentiles between experimental domains, rather than magnitude values, and the conversion can be found in Table 6. During preliminary testing, we tested eFSS with temporal window sizes of 1, 3, and 5 steps (i.e., ±0, ±5, and ±10 min, respectively). The addition of a temporal window did not greatly change the relative performance of the EPSs, but magnitudes of difference were reduced, and the larger sample of points for the ±10-min window allowed higher percentiles to be evaluated. The temporal window also mathematically excuses small errors in timing of storm location (in composite-reflectivity data) and mesocyclone location (in UH/AzShear data): a fair yardstick with which to measure operational or development EPSs. In the following, we present eFSS with a ±10-min (five forecast times) temporal window. We also introduce progressively larger tolerances in latitude–longitude space to avoid double-penalizing small phase errors in thunderstorm location. The eFSS scores calculated here (see section 4) are computed on spatial neighborhoods from native Δx to 27 km in diameter. This allows direct comparison at neighborhoods of 3 km and larger, measuring whether the change in Δx has affected skill at a given scale.

First, we combine all five initializations times for each case and compute eFSS for reflectivity forecasts (Fig. 8). We find 1) the $\Delta x = 1$-km EPS outperforms $\Delta x = 3$ km in the highest 1%–5% of reflectivity values (≳40 dBZ), consistent with the weak signal in Fig. 6; 2) the 1-km EPS progressively outperforms the 3-km EPS at larger Δx (scales of 5 km); and 3) conversely, overforecasting at $\Delta x = 1$ km (i.e., spurious convection) is evident as worse eFSS scores earlier in the forecasts. In general, results are mixed across all variables: case, length scale, lead time, and threshold. At this point, it is difficult to reject a null hypothesis that added value at $\Delta x = 1$ km cannot justify the additional computational expense.

In eFSS differences of UH_{02} forecasts (verified with their observational-proxy counterpart $AzShear_{02}$), there is a general trend for $\Delta x = 1$ km to perform better than $\Delta x = 3$ km at detecting low-level rotation (Fig. 9), particularly for the QLCS in B-20170501, but the gains are otherwise weak or inconsistent. For UH_{25}, presented in Fig. 10, we find eFSS gains at $\Delta x = 1$ km are larger in many pixels, but otherwise results are again mixed. In summary, a scale-aware evaluation suggests $\Delta x = 3$ km is superior at detecting weak-to-moderate reflectivity magnitudes, but $\Delta x = 1$ km is better at detecting high reflectivity and low- to midlevel rotation—with the caveat that results are inconsistent across cases.

9. Results: Storm-object attributes

Objects were first identified in composite-reflectivity fields using the methodology described in section 5. In this section, we consider either/both linear and cellular reflectivity objects.

a. Deterministic (member-to-member) framework

We compare attributes for EPS cellular objects only, due to their increased likelihood of producing tornadoes. The nine panels in Fig. 11 represent differences between EPS-matched objects for the following quantities: *area*, computed by counting each pixel that is part of the object; *extent*, which is the fraction of points in the bounding box that are part of the object (an object that resembles a Swiss cheese will have a low extent); *longest axis length* of the object; *eccentricity*, where unity indicates a circle; *maximum composite reflectivity* within the object; *maximum updraft speed* forecasted within the object; *maximum* UH_{02} and UH_{25} in the layer; and *distance between object centroids* in the horizontal plane.

On average, Fig. 11 suggests $\Delta x = 1$-km objects are smaller in area but longer in their major axis (Figs. 11c,d), and ~5 dBZ more intense (Fig. 11e), than their $\Delta x = 3$-km counterparts. Storm updrafts are typically stronger at $\Delta x = 1$ km; 25% of

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 09/02/21 02:14 PM UTC

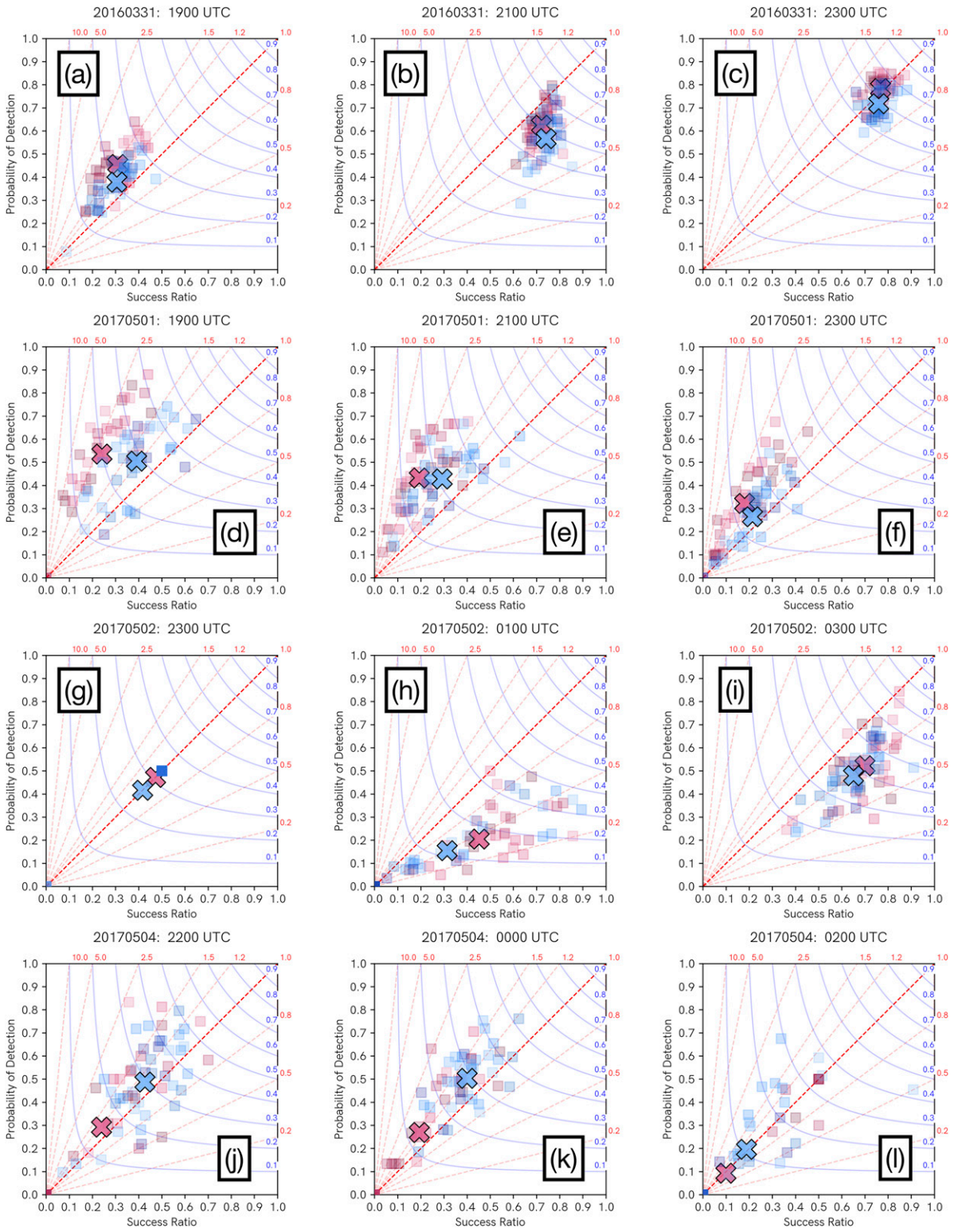

FIG. 12. Cellular-object performance diagrams of each EPS in detecting a thunderstorm, showing each ensemble member (squares; color denotes the EPS Δx) and the mean skill of the EPS (marked with a cross; color as above), for the (a),(d),(g),(j) first; (b),(e),(h),(k) third; and (c),(f),(i),(l) fifth initialization times, and for the (a)–(c) A-20160331, (d)–(f) B-20170501, (g)–(i) C-20170502, and (j)–(l) D-20170504 cases. Blue and red colors mark 3- and 1-km EPS results, respectively.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 09/02/21 02:14 PM UTC

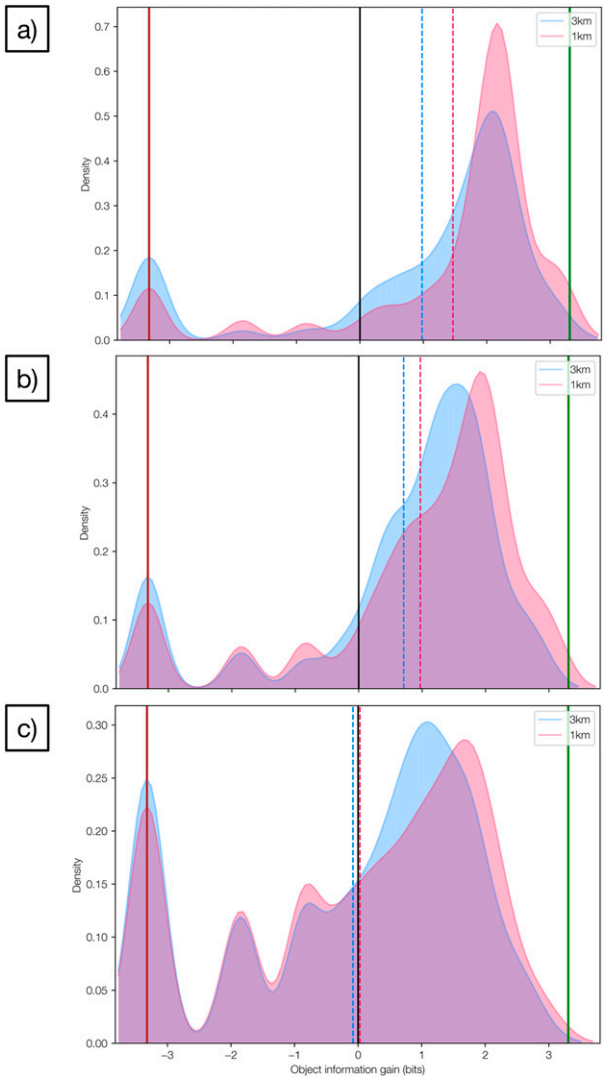

FIG. 13. Distribution of cellular-object information that is gained over a naive 10% expectation of object occurrence, in bits, for the (a) 0–1-, (b) 1–2-, and (c) 2–3-h lead-time windows of forecast-object occurrence. The *x*-axis zero-line (black line) also marks the boundary between a transfer of information that is useful vs detrimental to the end-user (assuming optimal posterior decision-making). Fill indicates the EPS Δx; dashed lines indicate median information gain/loss, colored likewise for each EPS. The red and green solid lines indicate maximum information loss and gain, respectively, which is solely a function of the choice of EPS-probability bounding (see text); the distribution extends past these limits due to the smoothing used. The *y* axis is normalized for each time.

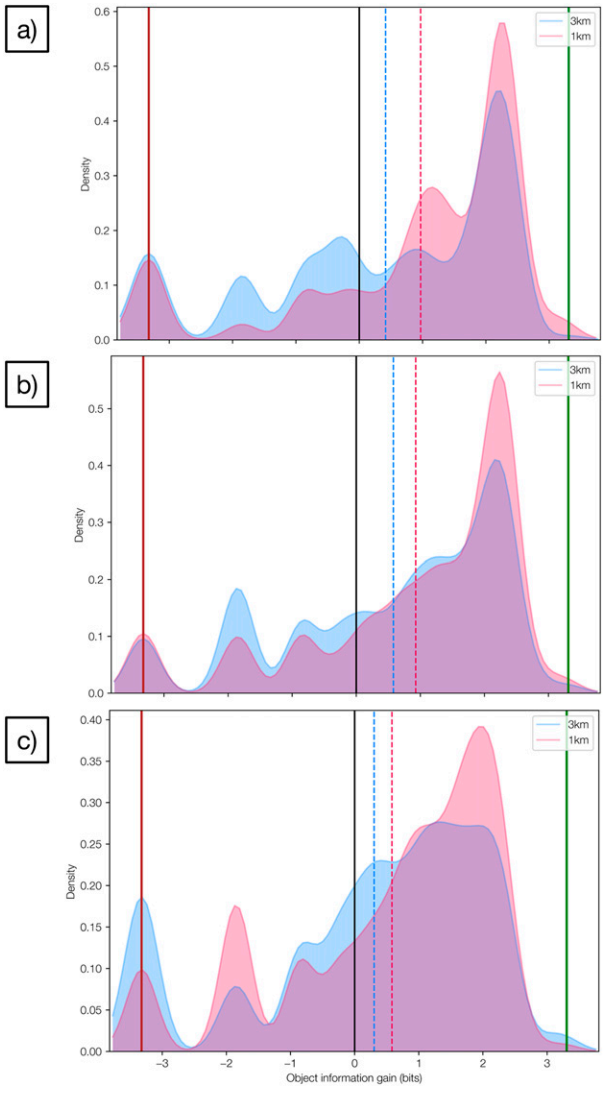

FIG. 14. As in Fig. 13, but for linear objects.

Δx = 1-km objects possess an updraft >10 m s^{-1} stronger. Rotation at Δx = 1 km was also stronger at low and midlevels (Figs. 11g,h). The typical location difference between matched objects was <20 km.

We show performance diagrams for the first, third, and fifth initialization times of all four cases in Fig. 12 for cellular-object-based contingency scores averaged over all forecast times. For brevity, we drop the second and fourth runs from discussion. When compared to gridpointwise performance diagrams (Fig. 5), we find more scatter points are closer to top right, despite including objects late in the forecast (i.e., low predictability), than in the traditional equivalents (not shown). This higher estimate of practical predictability stems from the filtering process of object identification, and the larger scale of interest associated with thunderstorm properties (rather than assessment of Lorenzian predictability at the Δx scale). Before proceeding, we remind the reader that object-based performance diagrams are a deterministic treatment of EPS output; object-based probabilistic skill gain is measured later in the section.

For the first initialization run (1900 UTC) for A-20160331 (Fig. 12a), we again see evidence of overforecasting of strong storms (cf. Fig. 7) as bias values >1 (i.e., left of the $x = y$ diagonal). The overforecasting bias is worse in Δx = 1 km. On average, members from either EPS perform approximately as well as each other; however, the POD ranges from 0.25 to 0.55

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 09/02/21 02:14 PM UTC

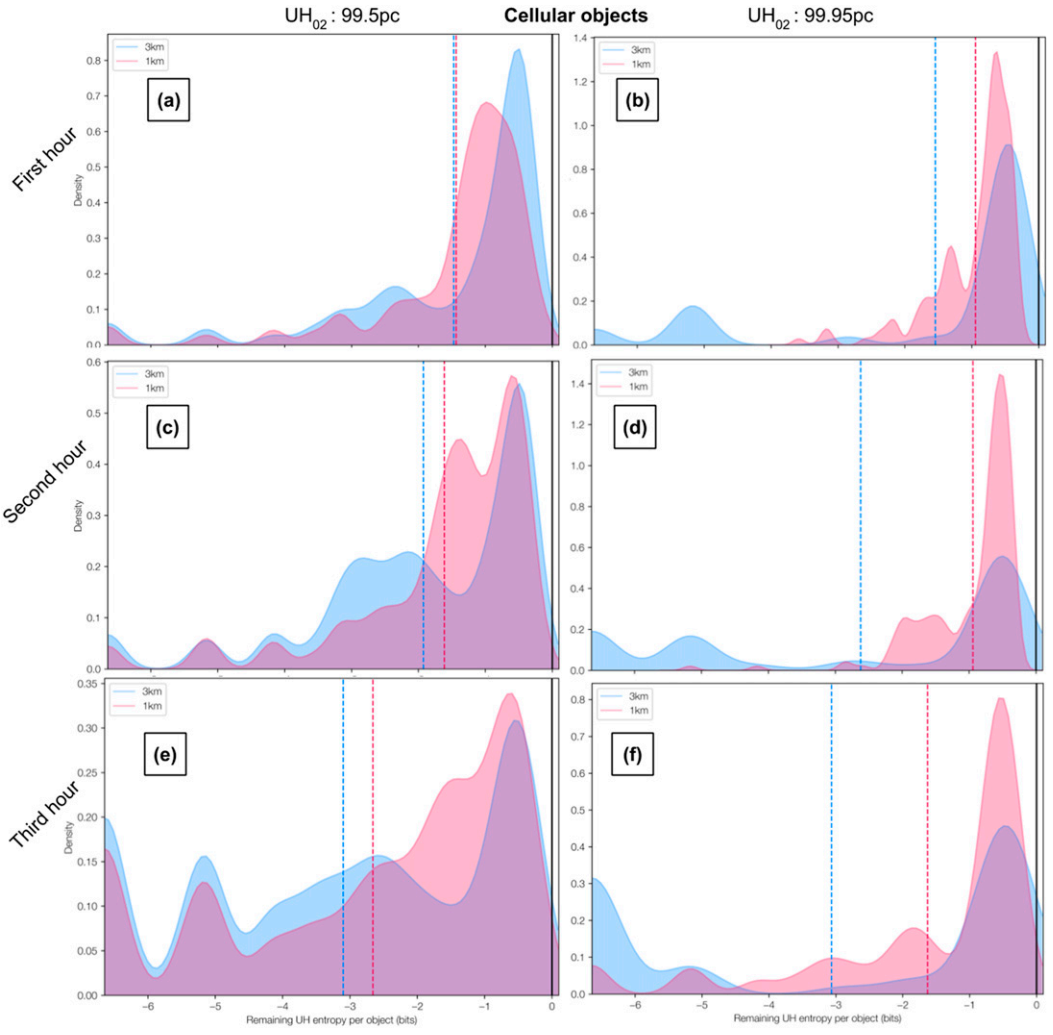

FIG. 15. Distribution of remaining uncertainty of low-level rotation UH_{02} per cellular object, in bits, for the (a),(b) 0–1-, (c),(d) 1–2-, and (e),(f) 2–3-h lead-time windows, exceeding the (a),(c),(e) 99.5th and (b),(d),(f) 99.95th percentiles. Forecasts lower than 20% were excluded from the figure to remove the excessive signal of correct negatives. Zero represents a perfect forecast. Median remaining uncertainty is shown by dotted lines; those in (a) are collocated.

for $\Delta x = 3$ km, and 0.25 to (a slightly higher) 0.65 for $\Delta x = 1$ km. Intermember differences span these ranges: there is often little correspondence between performance and member (not shown). On average, the 2100 UTC initialization performs better (Fig. 12b) than at 1900 UTC. The 1-km members still produce more objects than those at 3 km. There is also ~0.1 improvement in mean POD and CSI at $\Delta x = 1$ km, albeit at a slightly lower SR. For the run 2 h later (2300 UTC; Fig. 12c), the performance is even better than the 2100 UTC initialization. There is a tight cluster of points, suggesting relatively high predictability for this run. The 1-km EPS maintains a slight advantage for this time in all four variables.

In contrast, B-20170501 performs more poorly. The first initialization (1900 UTC; Fig. 12d) shows the finer grid substantially overforecasts storms; this bias eases for later cases (Figs. 12e,f), but is consistently larger than the coarser grid. In contrast to the later initializations of A-20160331 (above), we find a larger spread for all initializations of B-20170501: this indicates a relatively low casewide predictability.

As expected from the case with fewest storm objects, C-20170502 has mostly undefined performance scores for members in the first initialization (2300 UTC; Fig. 12g). Later, for the 0100 UTC run (Fig. 12h), the members that do contain sufficient objects for contigency-table calculation show slight improvement for $\Delta x = 1$ km, mainly in better SR, and there is an underforecasting bias for this case in both EPSs. The final 0300 UTC run (Fig. 12i) shows a slight improvement in mean 1-km performance, with a few finer-grid members demonstrating >0.7 POD.

Finally, D-20170504 is dominated by QLCS activity, and all three initializations time show a substantial improvement in

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 09/02/21 02:14 PM UTC

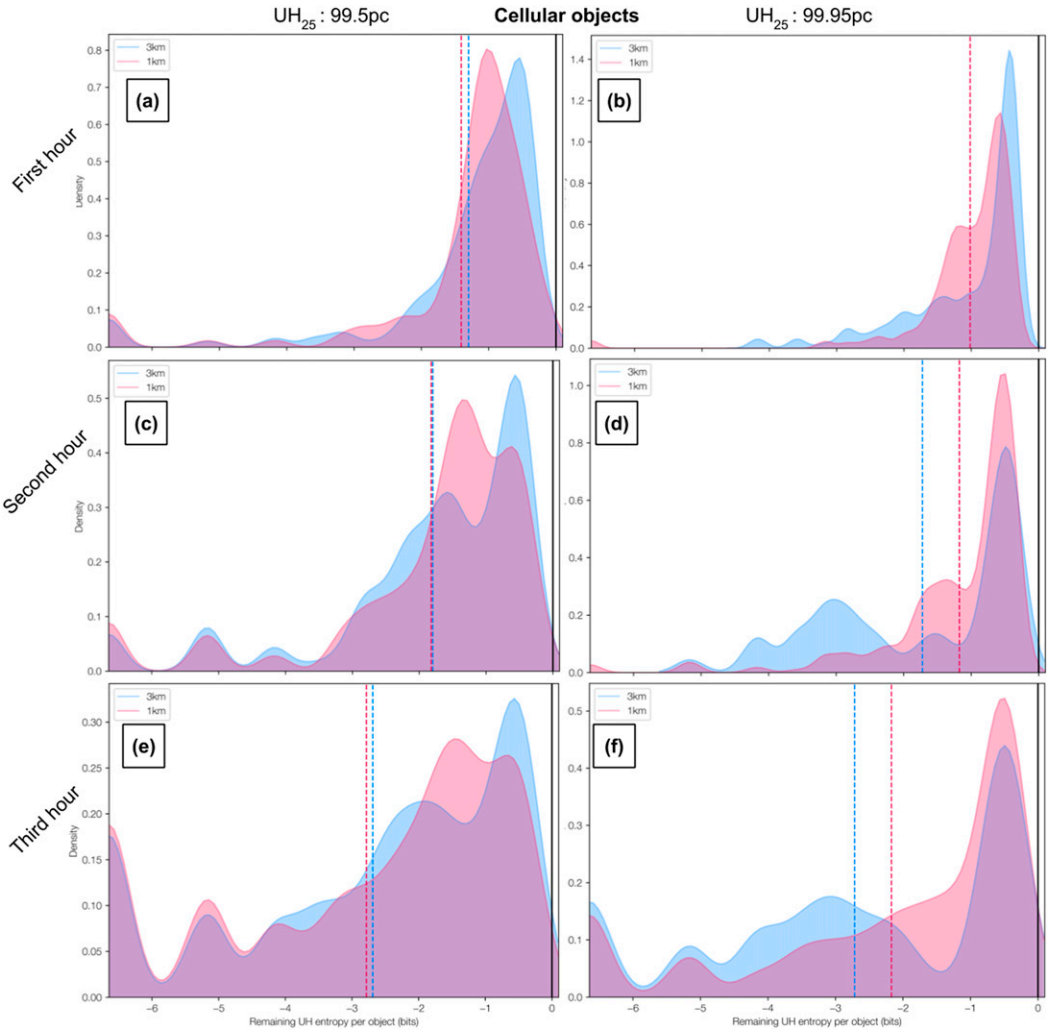

FIG. 16. As in Fig. 15, but for midlevel rotation UH_{25}. The median lines for each domain in (b) and (c) are collocated.

mean POD, SR, and CSI for the 3-km EPS. In this case, bias is close to unity (optimal) for both EPSs.

In the cases of high spread in POD–FAR space, an average is not sufficient to reconstruct the complexity of each forecast's probability distribution (i.e., we are losing information during the evaluation process). As such, we next perform a probabilistic evaluation of object forecasts.

b. *Probabilistic framework*

Before proceeding, we subset the Linear/Complex into a further subset of linear-only objects. This was done by including those Linear/Complex objects that exceeded 0.85 eccentricity, based on sensitivity testing.

1) OBJECT OCCURRENCE

To aggregate an estimate of information gained for each experiment about the objects' existence (within tolerances) over a prior expectation of 10%, we present the distributions of IG_o for all objects in the study, grouped by mode and forecast hour. A smoothing is applied; the sigma and kernel shape were tested to ensure a fair representation in Fig. 13, with the trade-off that values extend beyond the mathematical bounds.

We begin with cellular objects. During the first hour (Fig. 13a), we find a median information gain that is ~0.5 bits larger in the 1-km EPS versus at 3 km. For later forecast times (Figs. 13b,c), the medians of both EPSs approach zero as predictability is lost, and $\Delta x = 1$ km maintains a modest improvement in median information gain throughout. For linear objects, the finer grid consistently provides more value over $\Delta x = 3$ km for all three time periods (Figs. 14a–c).

2) ROTATION

In the following plots, we measure the remaining uncertainty in rotation for each EPS's identified storms. Hence, an optimal forecast would have zero bits of remaining uncertainty. We separate the following analyses into linear/cellular, low/midlevel

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 09/02/21 02:14 PM UTC

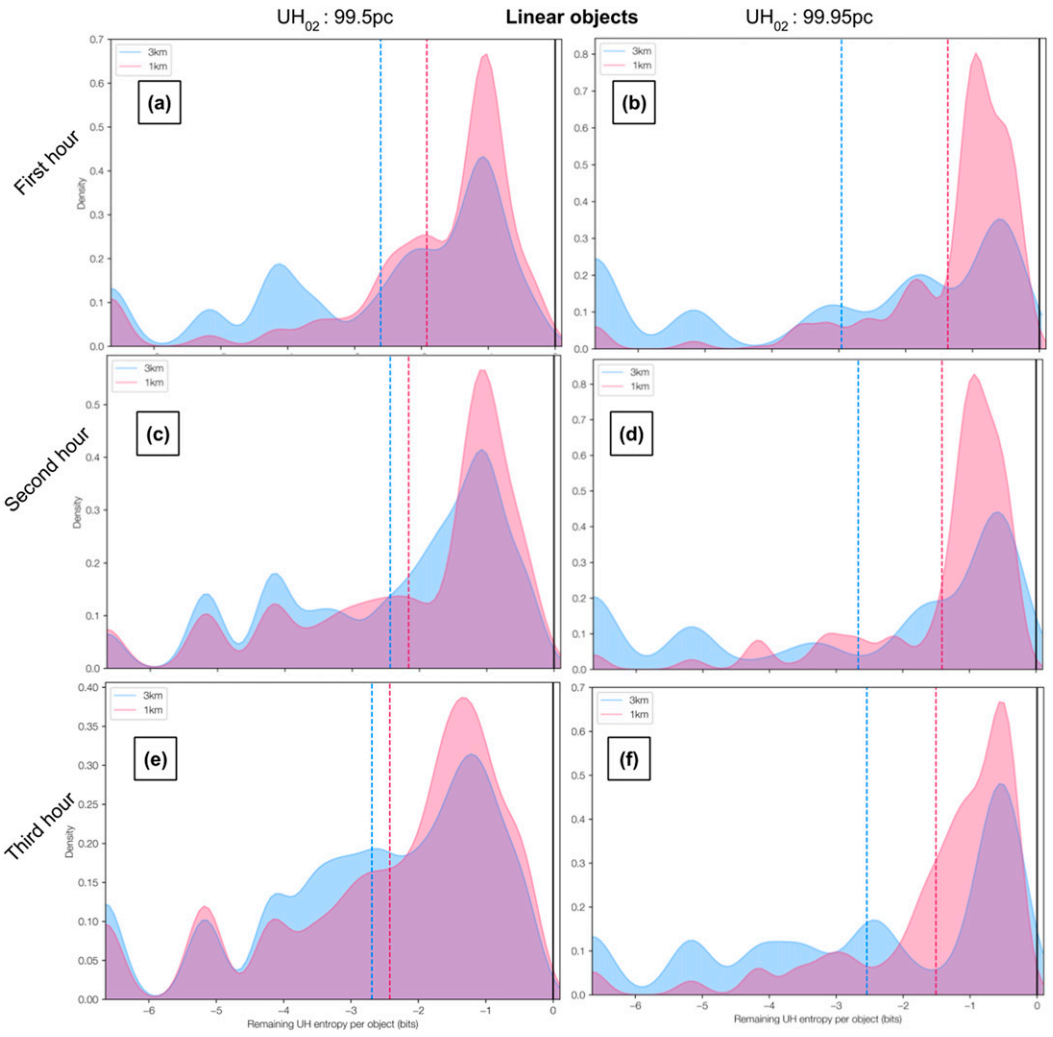

FIG. 17. As in Fig. 15, but for low-level rotation UH_{02} associated with linear objects.

rotation, exceedance of 99.5th and 99.95th percentiles of rotation at that level, and forecast hour.

First, cellular-object low-level rotation (Fig. 15) shows consistently larger uncertainty removal (lower negative values) in the 1-km EPS for both percentiles of strong low-level rotation, but substantially more value at the extreme percentile (99.95th). Remaining uncertainty increases with lead time due to predictability loss. The added value of $\Delta x =$ 1 km increases with lead time, likely due to first-hour overforecasting at 1 km. Moving to midlevel rotation for cellular objects, benefit of the finer grid is only evident at the more extreme percentile (Fig. 16). In summary, we may infer cellular-object rotation is better forecast at $\Delta x = 1$ km, especially at extremely high levels.

For linear objects detecting low-level rotation (Fig. 17), an advantage of using $\Delta x = 1$ km is seen consistently and conclusively across all lead times and both percentiles. The benefit of $\Delta x = 1$ km is likewise seen in midlevel-rotation detection (Fig. 18), albeit more modestly. In summary, the 1-km EPS is superior in gaining information about storm existence over a prior expectation of 10%, and consistently removes more uncertainty than $\Delta x = 3$ km for rotation at both levels (more so at lower levels) and at both top-percentile thresholds (more so at the 99.95th percentile).

10. Conclusions

Herein, we addressed whether benefit arises from ensemble-forecast resolution increases, particularly for thunderstorms in low-CAPE, high-shear environments. We evaluated rotation associated with cellular and linear features in composite reflectivity (i.e., supercells and quasi-linear convective systems). The forecast data were generated from two convection-allowing ensemble prediction systems, differing only in their domain size and horizontal grid resolution (3 vs 1 km), for five initialization times over 4 days. Performance of each ensemble experiment was assessed with three methods: 1) traditional pointwise metrics that contain no tolerance for spatial and/or timing error; 2) a scale-aware metric that tolerates space–time errors; and 3) a sequence of object-identification algorithms

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 09/02/21 02:14 PM UTC

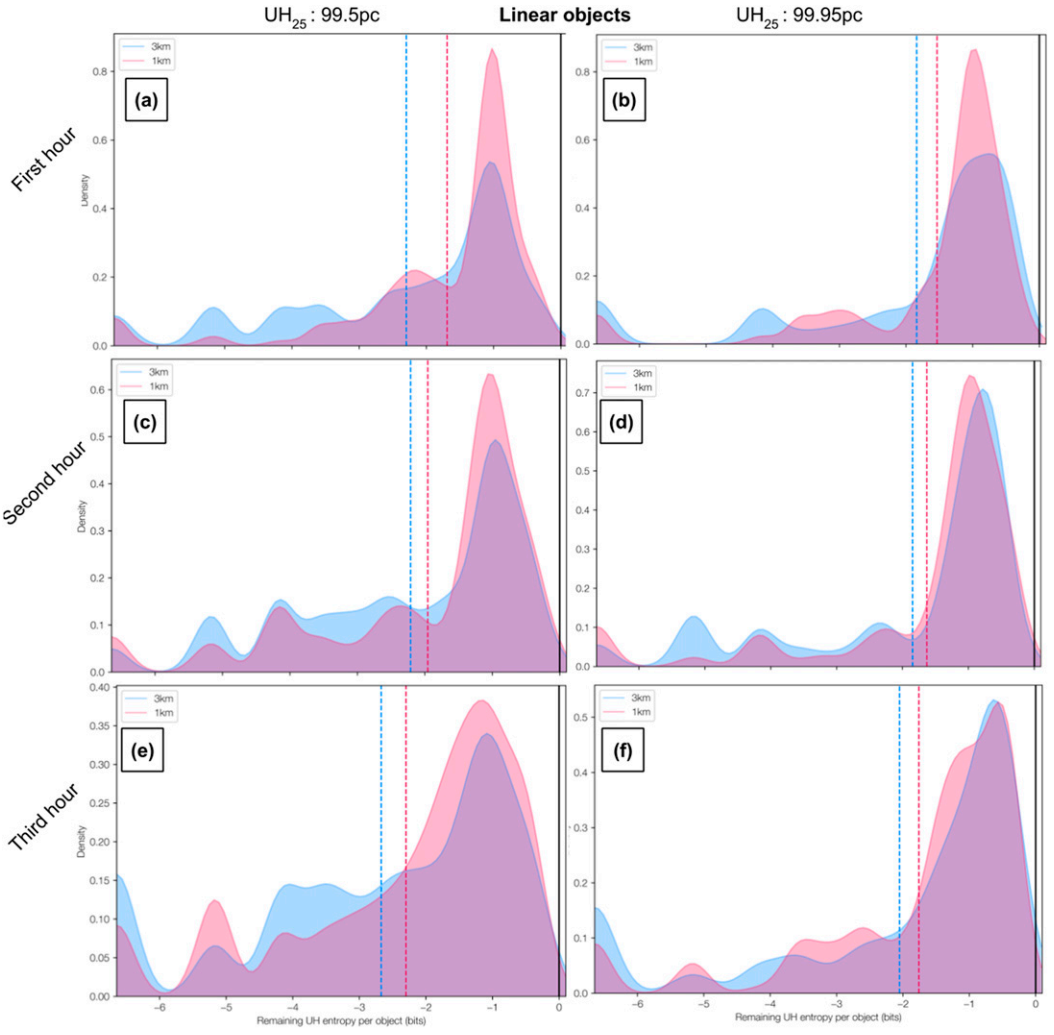

FIG. 18. As in Fig. 15, but for midlevel rotation UH_{25} associated with linear objects.

that yields an estimate of bulk information gain for each ensemble.

In conclusion, we find that scale-aware verification results are sensitive to the case, variable, lead time, and magnitude of the variable in question. The 1-km forecasts are typically better for detecting high reflectivity, and low- and midlevel strong rotation, but to the detriment of weak-to-moderate reflectivity forecast skill that might degrade the detection of parent thunderstorms. When viewed in an object-based uncertainty-removal paradigm, the distribution of object-specific information gain reveals that the NWP output of thunderstorm occurrence—a necessary precursor for tornadoes—is more valuable to the forecaster when a finer (1-km) grid is used. More uncertainty is also removed by the 1-km EPS regarding rotation detection within both linear and cellular objects, at both low- and midlevels (more so for low levels) and increasingly so at the extreme (99.95th percentile) rotation threshold. This corroborates the advantage of finer grids in (Potvin and Flora 2015; Sobash et al. 2019). However, given continuously increasing computer resources, it begs the question of whether these results justify decreasing Δx, rather than spending resources on increased ensemble membership, more frequent initialization times, and so on.

The authors encourage further use of information-theoretical frameworks to adequately reward ensemble forecasts of rare weather events, reducing probabilistic evaluation of any variable or diagnostic to the degree of information gain. Moreover, additional insight is gained after transformation of the gridded fields to thunderstorm objects for storm-characteristic verification. Work is ongoing regarding the effect of higher resolution on various thunderstorm modes within high-shear, high-CAPE flow; a larger dataset of cases is also required to confirm the findings herein.

Acknowledgments. Funding for this research was provided by the VORTEX-SE project and NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce. In addition to software packages cited above, we thank the developers of open-source software. We thank the editor

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 09/02/21 02:14 PM UTC

and anonymous reviewers, along with Chris Kerr and Derek Stratman for internal review and troubleshooting. Finally, the authors thank the following for inspiring discussion during the writing of this manuscript: Harold Brooks, Timothy DelSole, Montgomery Flora, Craig Schwartz, Louis Wicker, and all colleagues at CIMMS/NSSL.

REFERENCES

Anderson-Frey, A. K., Y. P. Richardson, A. R. Dean, R. L. Thompson, and B. T. Smith, 2019: Characteristics of tornado events and warnings in the southeastern United States. *Wea. Forecasting*, **34**, 1017–1034, https://doi.org/10.1175/WAF-D-18-0211.1.

Aran, M., J. Amaro, J. Arús, J. Bech, F. Figuerola, M. Gayà, and E. Vilaclara, 2009: Synoptic and mesoscale diagnosis of a tornado event in Castellcir, Catalonia, on 18th October 2006. *Atmos. Res.*, **93**, 147–160, https://doi.org/10.1016/j.atmosres.2008.09.031.

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

Ben-Naim, A., 2008: *A Farewell to Entropy, A: Statistical Thermodynamics Based On Information*. World Scientific, 412 pp.

Berner, J., K. R. Fossell, S.-Y. Ha, J. P. Hacker, and C. Snyder, 2015: Increasing the skill of probabilistic forecasts: Understanding performance improvements from model-error representations. *Mon. Wea. Rev.*, **143**, 1295–1320, https://doi.org/10.1175/MWR-D-14-00091.1.

——, and Coauthors, 2017: Stochastic parameterization: Toward a new view of weather and climate models. *Bull. Amer. Meteor. Soc.*, **98**, 565–588, https://doi.org/10.1175/BAMS-D-15-00268.1.

Blumberg, W. G., K. T. Halbert, T. A. Supinie, P. T. Marsh, R. L. Thompson, and J. A. Hart, 2017: SHARPpy: An Open-Source sounding analysis toolkit for the atmospheric sciences. *Bull. Amer. Meteor. Soc.*, **98**, 1625–1636, https://doi.org/10.1175/BAMS-D-15-00309.1.

Brooks, H. E., J. W. Lee, and J. P. Craven, 2003: The spatial distribution of severe thunderstorm and tornado environments from global reanalysis data. *Atmos. Res.*, **67–68**, 73–94, https://doi.org/10.1016/S0169-8095(03)00045-0.

Brotzge, J., and S. Erickson, 2010: Tornadoes without NWS warning. *Wea. Forecasting*, **25**, 159–172, https://doi.org/10.1175/2009WAF2222270.1.

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

Coffer, B. E., and M. D. Parker, 2018: Is there a ''tipping point'' between simulated nontornadic and tornadic supercells in VORTEX2 environments? *Mon. Wea. Rev.*, **146**, 2667–2693, https://doi.org/10.1175/MWR-D-18-0050.1.

Cover, T. M., and J. A. Thomas, 2012: *Elements of Information Theory*. John Wiley & Sons, 792 pp.

Davis, C., B. Brown, and R. Bullock, 2006: 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.

Ding, R., B. Liu, B. Gu, J. Li, and X. Li, 2019: Predictability of ensemble forecasting estimated using the Kullback–Leibler divergence in the Lorenz model. *Adv. Atmos. Sci.*, **36**, 837–846, https://doi.org/10.1007/s00376-019-9034-9.

Dixon, P. G., A. E. Mercer, J. Choi, and J. S. Allen, 2011: Tornado risk analysis: Is Dixie Alley an extension of Tornado Alley? *Bull. Amer. Meteor. Soc.*, **92**, 433–441, https://doi.org/10.1175/2010BAMS3102.1.

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.

Duc, L., K. Saito, and H. Seko, 2013: Spatial-temporal fractions verification for high-resolution ensemble forecasts. *Tellus*, **65A**, 18171, https://doi.org/10.3402/tellusa.v65i0.18171.

Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteor. Apps*, **15**, 51–64, https://doi.org/10.1002/met.25.

Faggian, N., B. Roux, P. Steinle, and B. Ebert, 2015: Fast calculation of the fractions skill score. *Mausam*, **66**, 457–466.

Flora, M. L., C. K. Potvin, and L. J. Wicker, 2018: Practical predictability of supercells: Exploring ensemble forecast sensitivity to initial condition spread. *Mon. Wea. Rev.*, **146**, 2361–2379, https://doi.org/10.1175/MWR-D-17-0374.1.

——, 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.

Geerts, B., T. Andretta, S. Luberda, J. Vogt, Y. Wang, L. Oolman, J. Finch, and D. Bikos, 2009: A case study of a long-lived tornadic mesocyclone in a low-CAPE complex-terrain environment. *Electron. J. Severe Storms Meteor.*, **4** (3), https://ejssm.org/ojs/index.php/ejssm/article/viewArticle/59/80.

Gilleland, E., D. A. Ahijevych, B. G. Brown, and E. E. Ebert, 2010: Verifying forecasts spatially. *Bull. Amer. Meteor. Soc.*, **91**, 1365–1376, https://doi.org/10.1175/2010BAMS2819.1.

Gleick, J., 2012: *The Information: A History, a Theory, a Flood*. Vintage Books, 544 pp.

Good, I. J., 1952: Rational decisions. *J. Roy. Stat. Soc.*, **14B**, 107–114, https://www.jstor.org/stable/2984087.

Green, D. M., and J. A. Swets, 1966: *Signal Detection Theory and Psychophysics*. Vol. 1, Wiley, 455 pp.

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

James, R. P., and P. M. Markowski, 2010: A numerical investigation of the effects of dry air aloft on deep convection. *Mon. Wea. Rev.*, **138**, 140–161, https://doi.org/10.1175/2009MWR3018.1.

Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley & Sons, 254 pp.

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. Skinner, K. Knopfmeier, E. Mansell, P. Minnis, R. Palikonda, and W. Smith, 2018: Comparison of cloud microphysics schemes in a Warn-on-Forecast system using synthetic satellite objects. *Wea. Forecasting*, **33**, 1681–1708, https://doi.org/10.1175/WAF-D-18-0112.1.

Kain, J. S., and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 09/02/21 02:14 PM UTC

operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931–952, https://doi.org/10.1175/WAF2007106.1.

Kerr, B. W., and G. L. Darkow, 1996: Storm-relative winds and helicity in the tornadic thunderstorm environment. *Wea. Forecasting*, **11**, 489–505, https://doi.org/10.1175/1520-0434(1996)011<0489:SRWAHI>2.0.CO;2.

Kirkpatrick, C., E. W. McCaul Jr., and C. Cohen, 2011: Sensitivities of simulated convective storms to environmental CAPE. *Mon. Wea. Rev.*, **139**, 3514–3532, https://doi.org/10.1175/2011MWR3631.1.

Kis, A. K., and J. M. Straka, 2010: Nocturnal tornado climatology. *Wea. Forecasting*, **25**, 545–561, https://doi.org/10.1175/2009WAF2222294.1.

Lawson, J. R., 2019: Predictability of idealized thunderstorms in buoyancy–shear space. *J. Atmos. Sci.*, **76**, 2653–2672, https://doi.org/10.1175/JAS-D-18-0218.1.

——, J. S. Kain, N. Yussouf, D. C. Dowell, D. M. Wheatley, K. H. Knopfmeier, and T. A. Jones, 2018: Advancing from convection-allowing NWP to Warn-on-Forecast: Evidence of progress. *Wea. Forecasting*, **33**, 599–607, https://doi.org/10.1175/WAF-D-17-0145.1.

Loken, E. D., A. J. Clark, M. Xue, and F. Kong, 2017: Comparison of next-day probabilistic severe weather forecasts from coarse- and fine-resolution CAMs and a convection-allowing ensemble. *Wea. Forecasting*, **32**, 1403–1421, https://doi.org/10.1175/WAF-D-16-0200.1.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

Mahalik, M. C., B. R. Smith, K. L. Elmore, D. M. Kingfield, K. L. Ortega, and T. M. Smith, 2019: Estimates of gradients in radar moments using a linear least squares derivative technique. *Wea. Forecasting*, **34**, 415–434, https://doi.org/10.1175/WAF-D-18-0095.1.

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. M., and J. M. Straka, 2000: Some observations of rotating updrafts in a low-buoyancy, highly sheared environment. *Mon. Wea. Rev.*, **128**, 449–461, https://doi.org/10.1175/1520-0493(2000)128<0449:SOORUI>2.0.CO;2.

——, and Y. Richardson, 2010: *Mesoscale Meteorology in Midlatitudes*. Wiley-Blackwell, 407 pp.

——, J. M. Straka, E. N. Rasmussen, and D. O. Blanchard, 1998: Variability of storm-relative helicity during VORTEX. *Mon. Wea. Rev.*, **126**, 2959–2971, https://doi.org/10.1175/1520-0493(1998)126<2959:VOSRHD>2.0.CO;2.

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.

Peirolo, R., 2011: Information gain as a score for probabilistic forecasts. *Meteor. Appl.*, **18**, 9–17, https://doi.org/10.1002/met.188.

Pierce, J. R., 1980: *An Introduction to Information Theory: Symbols, Signals and Noise*. Dover Publications, 336 pp.

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.

Powers, J. G., and Coauthors, 2017: The Weather Research and Forecasting model: Overview, system efforts, and future directions. *Bull. Amer. Meteor. Soc.*, **98**, 1717–1737, https://doi.org/10.1175/BAMS-D-15-00308.1.

Radanovics, S., J.-P. Vidal, and E. Sauquet, 2018: Spatial verification of ensemble precipitation: An ensemble version of SAL. *Wea. Forecasting*, **33**, 1001–1020, https://doi.org/10.1175/WAF-D-17-0162.1.

Rasmussen, E. N., 2003: Refined supercell and tornado forecast parameters. *Wea. Forecasting*, **18**, 530–535, https://doi.org/10.1175/1520-0434(2003)18<530:RSATFP>2.0.CO;2.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, https://doi.org/10.1175/2007MWR2123.1.

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

Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.*, **130**, 1653–1660, https://doi.org/10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2.

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.

——, and ——, 2019: Revisiting sensitivity to horizontal grid spacing in convection-allowing models over the central-eastern United States. *Mon. Wea. Rev.*, **147**, 4411–4435, https://doi.org/10.1175/MWR-D-19-0115.1.

——, and Coauthors, 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263–280, https://doi.org/10.1175/2009WAF2222267.1.

——, G. S. Romine, K. R. Fossell, R. A. Sobash, and M. L. Weisman, 2017: Toward 1-km ensemble forecasts over large domains. *Mon. Wea. Rev.*, **145**, 2943–2969, https://doi.org/10.1175/MWR-D-16-0410.1.

——, ——, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2019: NCAR's real-time convection-allowing ensemble project. *Bull. Amer. Meteor. Soc.*, **100**, 321–343, https://doi.org/10.1175/BAMS-D-17-0297.1.

Shannon, C. E., 1948: A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423, https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Sherburn, K. D., and M. D. Parker, 2014: Climatology and ingredients of significant severe convection in high-shear, low-CAPE environments. *Wea. Forecasting*, **29**, 854–877, https://doi.org/10.1175/WAF-D-13-00041.1.

Skamarock, W. C., 2004: Evaluating mesoscale NWP models using kinetic energy spectra. *Mon. Wea. Rev.*, **132**, 3019–3032, https://doi.org/10.1175/MWR2830.1.

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

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 09/02/21 02:14 PM UTC

Smart, D. J., and K. A. Browning, 2009: Morphology and evolution of cold-frontal misocyclones. *Quart. J. Roy. Meteor. Soc.*, **135**, 381–393, https://doi.org/10.1002/qj.399.

Smith, B. T., R. L. Thompson, J. S. Grams, C. Broyles, and H. E. Brooks, 2012: Convective modes for significant severe thunderstorms in the contiguous United States. Part I: Storm classification and climatology. *Wea. Forecasting*, **27**, 1114–1135, https://doi.org/10.1175/WAF-D-11-00115.1.

Smith, T. M., and K. L. Elmore, 2004: The use of radial velocity derivatives to diagnose rotation and divergence. *11th Conf. on Aviation, Range, and Aerospace*, Hyannis, MA, Amer. Meteor. Soc., P5.6, https://ams.confex.com/ams/11aram22sls/techprogram/paper_81827.htm.

——, 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.

Snyder, C., and F. Zhang, 2003: Assimilation of simulated Doppler radar observations with an ensemble Kalman filter. *Mon. Wea. Rev.*, **131**, 1663–1677, https://doi.org/10.1175//2555.1.

Sobash, R. A., C. S. Schwartz, G. S. Romine, and M. L. Weisman, 2019: Next-day prediction of tornadoes using convection-allowing models with 1-km horizontal grid spacing. *Wea. Forecasting*, **34**, 1117–1135, https://doi.org/10.1175/WAF-D-19-0044.1.

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.

Sterk, A. E., D. B. Stephenson, M. P. Holland, and K. R. Mylne, 2016: On the predictability of extremes: Does the butterfly effect ever decrease? *Quart. J. Roy. Meteor. Soc.*, **142**, 58–64, https://doi.org/10.1002/qj.2627.

Trapp, R. J., and M. L. Weisman, 2003: Low-level mesovortices within squall lines and bow echoes. Part II: Their genesis and implications. *Mon. Wea. Rev.*, **131**, 2804–2823, https://doi.org/10.1175/1520-0493(2003)131<2804:LMWSLA>2.0.CO;2.

——, 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.

——, G. R. Marion, and S. W. Nesbitt, 2017: The regulation of tornado intensity by updraft width. *J. Atmos. Sci.*, **74**, 4199–4211, https://doi.org/10.1175/JAS-D-16-0331.1.

van der Walt, S., J. L. Schönberger, 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.

Wang, Y., J. Gao, P. S. Skinner, K. Knopfmeier, T. Jones, G. Creager, P. L. Heiselman, and L. J. Wicker, 2019: Test of a weather-adaptive dual-resolution hybrid Warn-on-Forecast analysis and forecast system for several severe weather events. *Wea. Forecasting*, **34**, 1807–1827, https://doi.org/10.1175/WAF-D-19-0071.1.

Weijs, S. V., R. van Nooijen, and N. van de Giesen, 2010: Kullback–Leibler divergence as a forecast skill score with classic reliability–resolution–uncertainty decomposition. *Mon. Wea. Rev.*, **138**, 3387–3399, https://doi.org/10.1175/2010MWR3229.1.

Weisman, M. L., and R. J. Trapp, 2003: Low-level mesovortices within squall lines and bow echoes. Part I: Overview and dependence on environmental shear. *Mon. Wea. Rev.*, **131**, 2779–2803, https://doi.org/10.1175/1520-0493(2003)131<2779:LMWSLA>2.0.CO;2.

Wernli, H., M. Paulat, M. Hagen, and C. Frei, 2008: SAL—A novel quality measure for the verification of quantitative precipitation forecasts. *Mon. Wea. Rev.*, **136**, 4470–4487, https://doi.org/10.1175/2008MWR2415.1.

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.

Williams, G. P., 1997: *Chaos Theory Tamed*. Joseph Henry Press, 520 pp.

Wilson, K. A., P. L. Heinselman, P. S. Skinner, J. J. Choate, and K. E. Klockow-McClain, 2019: Meteorologists' interpretations of storm-scale ensemble-based forecast guidance. *Wea. Climate Soc.*, **11**, 337–354, https://doi.org/10.1175/WCAS-D-18-0084.1.

Zeeman, E. C., 1979: Catastrophe theory. *Structural Stability in Physics*, W. Güttinger and H. Eikemeier, Eds., Springer, 12–22.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 09/02/21 02:14 PM UTC