

Next-Day Prediction of Tornadoes Using Convection-Allowing Models with 1-km Horizontal Grid Spacing

RYAN A. SOBASH, CRAIG S. SCHWARTZ, GLEN S. ROMINE, AND MORRIS L. WEISMAN

National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 1 March 2019, in final form 5 June 2019)

ABSTRACT

Explicit attributes of convective storms within convection-allowing model (CAM) forecasts are routinely used as surrogates for convective weather hazards. The ability of 3- and 1-km horizontal grid spacing CAM forecasts to anticipate tornadoes using surrogates was examined for 497 severe weather events. Five diagnostics were used as tornado surrogates, including 0–1 km above ground level (AGL) updraft helicity (UH01), 2–5 km AGL UH (UH25), 0–3 km AGL UH (UH03), and 500 m and 1 km AGL relative vorticity. Next-day surrogate severe probability forecasts (SSPFs) for tornadoes were produced by thresholding the diagnostics and smoothing the resulting binary field. SSPFs were verified against SPC tornado reports and NWS tornado warnings. The 1-km SSPFs were more skillful than 3-km SSPFs across all diagnostics with statistically significant differences in skill that were largest on the mesoscale. UH01 outperformed the other four diagnostics, in part because UH01 best represented regional variations in observed tornado report totals. Filtering forecasts based on the significant tornado parameter benefited the 3-km SSPFs much more than the 1-km SSPFs, with filtered 3-km SSPFs having similar skill to the filtered 1-km SSPFs. SSPFs verified with a combination of tornado warnings and reports were more skillful than when verified against reports alone, indicating that CAMs can better predict intense low-level rotation events than tornadoes. When verifying all severe hazards, UH25 SSPFs were more skillful than UH01 SSPFs; UH01 and UH25 appear to be the most useful pair for anticipating tornadoes and the combined severe threat on a given forecast day.

1. Introduction

The suite of operational numerical weather prediction (NWP) guidance presently available to forecasters includes models configured with horizontal grid spacing, Δx , fine enough to at least partially resolve convective storms [e.g., the High-Resolution Rapid Refresh (HRRR) model (Benjamin et al. 2016) or the High-Resolution Ensemble Forecast (HREF) system (Roberts et al. 2019), both run operationally at the U.S. National Centers for Environmental Prediction (NCEP) with Δx of ~ 3 km]. These convection-allowing models (CAMs) yield improved predictions of precipitation extremes compared to convection-parameterizing NWP models (e.g., Mass et al. 2002; Kain et al. 2006; Clark et al. 2009; Schwartz et al. 2009) and provide explicit representations of severe convective modes, such as supercells and mesoscale convective systems (e.g., Done et al. 2004; Weisman et al. 2008).

While CAMs with ~ 3 -km Δx are capable of producing realistic convective structures (Weisman et al. 1997, 2008), severe weather hazards are not fully resolved and must be identified indirectly through surrogate diagnostics. Diagnostic fields related to the presence and intensity of mesocyclones have proved especially valuable [e.g., updraft helicity (UH), Kain et al. 2008, 2010; Sobash et al. 2011; Naylor et al. 2012], since supercells are prolific severe weather producers in the United States (Gallus et al. 2008; Duda and Gallus 2010, Smith et al. 2012). Extreme values of UH are used to identify supercells in CAMs and have been used to produce next-day (i.e., forecast lead times > 12 h) guidance (e.g., Sobash et al. 2011, 2016) for the combined threat from all severe hazards [i.e., hail ≥ 1 in., wind gusts ≥ 50 kt ($1 \text{ kt} \approx 0.51 \text{ m s}^{-1}$), and/or tornadoes]. Guidance for individual hazards has also been explored, such as using 10-m wind speed and vertically integrated graupel diagnostics to predict wind and hail (e.g., Clark et al. 2012; Hepper et al. 2016); UH with calibrated NCEP Short-Range Ensemble Forecast (SREF) environmental information to produce hail, wind, and tornado forecasts

Corresponding author: Dr. Ryan A. Sobash, sobash@ucar.edu

DOI: 10.1175/WAF-D-19-0044.1

© 2019 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy](#) (www.ametsoc.org/PUBSReuseLicenses).

(Jirak et al. 2014); machine-learning and hail models within CAMs to predict severe hail (e.g., Gagne et al. 2017; Adams-Selin et al. 2019); UH to identify tornadoes (e.g., Clark et al. 2013, Sobash et al. 2016); and combining UH with the significant tornado parameter (STP; Thompson et al. 2003) and climatological tornado frequencies, to produce tornado guidance (e.g., Gallo et al. 2016, 2018, 2019).

These studies have noted deficiencies of CAMs with 3–4-km Δx at predicting the physical processes related to convective hazards. Hepper et al. (2016) described two cases where convective winds were underforecast in their 4-km forecasts, suggesting the issue was due to the inability to resolve processes related to the production of near-surface convective gusts. Additionally, in their tornado forecasts, Gallo et al. (2016, 2018, 2019) used STP to reduce overforecasting when using UH alone, commenting that 3–4-km CAMs were unable to represent processes leading to intense low-level rotation. A similar overforecasting issue was present in the tornado forecasts of Jirak et al. (2014). Errant storm motions have also been noted in 3–4-km forecasts of supercells (VandenBerg et al. 2014) and MCSs (Schwartz et al. 2017). Similar deficiencies have also been identified in idealized studies using simplified cloud models, with mesocyclone cycling (Adlerman and Droegemeier 2002), low-level vorticity intensification (Potvin and Flora 2015), and updraft and downdraft strength (e.g., Bryan et al. 2003; Bryan and Morrison 2012) being poorly handled at 3–4-km Δx , but better represented at $\Delta x \leq 1$ km.

Even though moving toward finer Δx improves storm-scale processes in idealized simulations, studies of next-day CAM forecasts have documented minimal benefit to moving toward $\Delta x = 1$ km for forecasts of precipitation and severe weather (e.g., Kain et al. 2008; Schwartz et al. 2009; Clark et al. 2012; Johnson et al. 2013; Loken et al. 2017). Among the studies that focused on severe weather phenomena, Kain et al. (2008) determined that while 2-km forecasts possessed more detailed structures than 4-km forecasts, the forecasts were qualitatively similar in their placement of storms on a given day, and the climatology of UH objects were similar after adjusting for differences in the scaling of the UH diagnostics. Clark et al. (2012) arrived at similar conclusions, with participants in the 2010 NOAA Hazardous Weather Testbed (HWT) Spring Experiment subjectively rating 1-km forecasts similarly to 4-km forecasts for a majority of events. Loken et al. (2017) quantitatively evaluated ~60 next-day CAM forecasts of severe weather from the 2010 and 2011 HWT Spring Experiments and determined that, while 1-km forecasts were slightly superior to 4-km forecasts, the results were not statistically significant.

The aforementioned studies have largely focused on the placement of rotating storms in CAM output, which appears to be relatively insensitive to reductions in Δx below 4 km, at least within springtime convective environments. We hypothesize that the development of near-surface rotation (e.g., associated with low-level mesocyclones) will be better captured as Δx decreases, leading to diagnostics that can better discriminate between tornadic and nontornadic events. To investigate the impact of reducing Δx to 1 km on next-day tornado forecasts, we extend the work of Sobash et al. (2016) by verifying 1-km forecasts using diagnostics more closely tied to the development of low-level rotation within supercells. We also expand on Sobash et al. (2016) by 1) using 497 deterministic forecasts of specific events occurring between 2010 and 2017, rather than 91 ensemble forecasts of consecutive events from May to July 2015, to provide a statistically robust sample of diverse events across a range of seasons and environments, 2) combining the diagnostics with environmental parameters (i.e., STP), similar to the methods used by Gallo et al. (2016, 2018), and 3) supplementing tornado reports with tornado warnings as a novel verification dataset to capture low-level rotation events that did not produce tornadoes.

2. Methods

a. Case selection

A total of 497 events were simulated with both 3- and 1-km Δx on days with severe thunderstorms east of the Rocky Mountains. Cases were selected from the Storm Prediction Center's (SPC's) severe thunderstorm event archive, which includes events meeting certain criteria, including the total number of severe weather reports, monetary losses, and fatalities.¹ Forecasts were produced for all events in the SPC archive occurring between 15 March and 15 July (the "warm season") each year between 2011 and 2016 (inclusive). We also produced forecasts between 15 October and 14 March (the "cool season") for the 2010–11, 2011–12, . . . , 2015–16, and 2016–17 periods, when severe weather occurred less frequently.

The criteria for including cool-season events in the SPC's archive was relaxed, and some cool-season events had very few storm reports over localized areas that we did not simulate. We chose to neglect cool-season events with <20 storm reports, and included nearly all events

¹SPC event archive and selection criteria available at <http://www.spc.noaa.gov/exper/archive/events/>.

TABLE 1. Dates of 497 WRF forecasts and number of events per month (italics). Forecasts were initialized at 0000 UTC on the date shown. All 419 warm-season events (15 Mar–15 Jul) in the SPC storm event archive between 2010 and 2017 were simulated, in addition to 78 selected cool-season (15 Oct–14 Mar) events.

| | No. of events per month | 2010–11 | 2011–12 | 2012–13 | 2013–14 | 2014–15 | 2015–16 | 2016–17 |
|-----|-------------------------|-------------------------------------|--|-----------------------------------|---------------------------------|---------------------------------|--------------------------------------|----------------------------|
| Oct | <i>8</i> | 24–27 | | 17 | 31 | | 30, 31 | |
| Nov | <i>17</i> | 16, 22, 29, 30 | 7, 8, 14, 16 | | 17 | 16, 23 | 11, 16, 17 | 28–30 |
| Dec | <i>11</i> | 31 | 22 | 17, 19, 20, 25 | 21 | 23 | 23, 27 | 17 |
| Jan | <i>13</i> | 25 | 17, 22, 25 | 29, 30 | 11 | 3 | 21 | 2, 20–22 |
| Feb | <i>20</i> | 1, 24, 27, 28 | 18, 24, 28, 29 | 10, 18 | 20, 21 | | 15, 16, 23, 24 | 7, 19, 25, 28 |
| Mar | <i>54</i> | 5, 8, 9, 10, 22, 23, 26, 27, 29, 31 | 2, 15, 18–21, 23, 24, 28, 29, 31 | 18, 23, 24, 29–31 | 16, 26–28 | 24, 25, 31 | 8, 15, 17, 18, 23, 24, 27, 30, 31 | 1, 6, 9, 21, 23, 24, 26–30 |
| Apr | <i>82</i> | 3, 4, 8–10, 11, 14–16, 18–28, 30 | 1–3, 5, 7, 9, 13, 14, 15, 21, 26, 28, 30 | 1, 7, 9–11, 14, 17–19, 29 | 2–4, 6, 7, 12–14, 24, 25, 27–29 | 2, 3, 8, 9, 16–20, 22, 24–27 | 1, 6, 7, 10, 11, 15, 24, 26–29 | |
| May | <i>101</i> | 9–12, 21–27, 29, 30 | 1–6, 9, 10, 19, 25, 27–31 | 2, 8–10, 13, 15–23, 27–31 | 7–11, 13, 14, 20–22, 25–27, 30 | 5–11, 15, 16, 19, 23–28, 30, 31 | 1–5, 7–12, 16, 17, 21, 22, 24–29, 31 | |
| Jun | <i>123</i> | 1, 4–26 | 1, 4, 6, 7, 11, 13, 14, 22–24, 29, 30 | 2, 5, 6, 10, 12, 13, 15–22, 24–29 | 1, 3–11, 14, 16–22, 24–30 | 1, 4, 5, 7–10, 12, 14, 17–30 | 2, 5, 8, 10, 13–17, 19–26, 28, 29 | |
| Jul | <i>68</i> | 1–6, 10–13 | 1–8, 10, 12, 15 | 1–4, 7–12, 15 | 2, 3, 5–10, 13–15 | 1–3, 5–10, 12–15 | 4–15 | |

that produced >100 storm reports (45 events; 9 March 2013 was the sole exception). The remaining cool-season events generated between 20 and 100 storm reports and were selected to capture events across a diversity of months and regions. These criteria resulted in forecasts of 419 warm-season and 78 cool-season events (Table 1), with forecast initializations ranging from 0000 UTC 24 October 2010 through 0000 UTC 30 March 2017. The case selection strategy neglected events between 16 July and 14 October. These events were not emphasized due to many being weakly forced and occurring when CAM severe weather predictability is reduced (Sobash and Kain 2017). Avoiding these events also mitigated issues associated with landfalling tropical cyclones. The case selection strategy also neglected the inclusion of “false alarm” events, that is, forecast severe weather events that were not observed, although, an assessment of the ability of 3- and 1-km forecasts to discriminate between tornadic and nontornadic episodes should be maintained, due to the existence of many null tornado events.

b. Model configurations and diagnostics

Independent forecasts with 3- and 1-km Δx were produced with version 3.6.1 of the Advanced Research version of the Weather Research and Forecasting

(WRF) Model (WRF-ARW; Skamarock et al. 2008; Powers et al. 2017). Both sets of forecasts were initialized by interpolating 0000 UTC 0.5° Global Forecast System (GFS) analyses onto the 3- and 1-km domains and used 3-hourly GFS forecasts as lateral boundary conditions. Although 0.25° GFS grids became available in 2015, 0.5° GFS fields were used for all 497 events. All forecasts used a computational domain spanning the entire continental United States (CONUS) (Fig. 1), used identical physical parameterizations (Table 2), and had 40 vertical levels and a 50-hPa model top. The time step was set to $4\Delta x$ for both sets of forecasts (i.e., 4 s for the 1-km forecasts and 12 s for the 3-km forecasts). The 3-km forecasts using the same time step (i.e., 4 s) were less skillful than the 3-km, 12-s time step forecasts, due to increased WRF sixth-order diffusion (Knievel et al. 2007). The 3-km, 4-s time step forecasts with reduced diffusion produced results nearly identical results to the 3-km, 12-s time step forecasts, thus the skill differences between the 1-km, 4-s time step forecasts and 3-km, 12-s time step forecasts were attributed solely to Δx , and not time step differences.

Five diagnostics related to rotation were computed: 2–5 km AGL updraft helicity (UH25), 0–3 km AGL updraft helicity (UH03), 0–1 km AGL updraft helicity

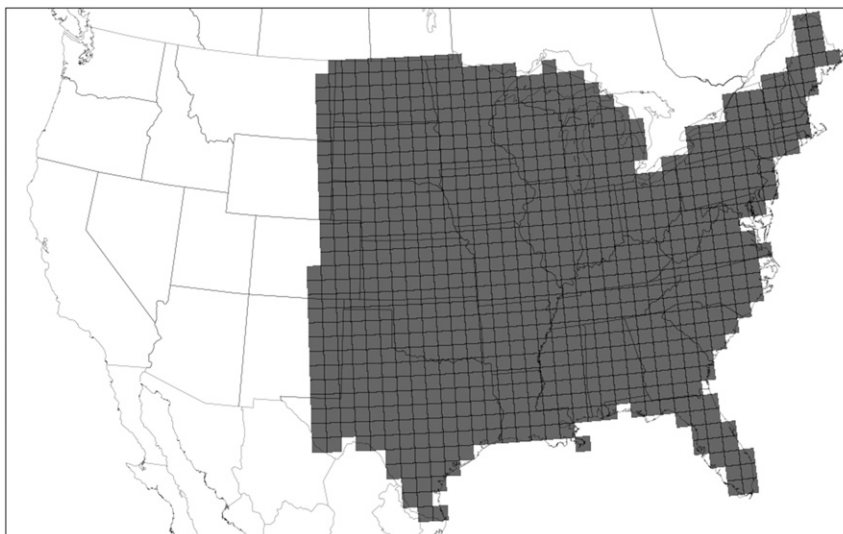


FIG. 1. The WRF domain for 1- and 3-km forecasts. The 80-km grid boxes used for verification are shaded. The 1-km domain consisted of exactly 9 times the number of grid points (4743×2958) as the 3-km domain (1581×986).

(UH01), 1 km AGL vertical vorticity (RVORT1), and 500 m AGL vertical vorticity (RVORT0.5) (Table 3). While UH25, UH03, and RVORT1 were used in Sobash et al. (2016), UH01 and RVORT0.5 diagnostics have not been used in previous CAM verification studies. Generally, UH is defined as the integral of updraft speed multiplied by vertical vorticity between two levels (Kain et al. 2008), which is approximated in WRF by numerical integration, using updraft speed and vertical vorticity computed on model levels between 2 and 5 km (UH25), between 0 and 3 km (UH03), and between 0 and 1 km (UH01) AGL. RVORT1 and RVORT0.5 were computed by using the vertical vorticity within the WRF UH computation, interpolated to 1 km and 500 m AGL. A nine-point smoother was applied after each time step to the UH25, UH03, and UH01 fields, but not to the RVORT1 and RVORT0.5 fields, since these fields used the vorticity fields computed as a part of the UH computation, prior to smoothing. Diagnostics were computed each time step during WRF integration and stored as hourly maximum values as in Kain et al. (2010). Since the maximum value of each diagnostic was stored each hour, and each field was initialized with zeros, only positive magnitudes of each diagnostic, representative of cyclonic rotation, were preserved.

c. Producing tornado SSPFs

The 3- and 1-km forecasts were evaluated by producing surrogate severe weather reports (SSRs). SSRs are designed to be the model counterpart to observed severe reports (OSRs), representing the anticipated

locations of OSRs based on extremes in CAM diagnostics such as UH25 (e.g., Sobash et al. 2011, 2016; Gallo et al. 2016; Loken et al. 2017). To produce SSRs for each forecast for a particular diagnostic, the two-dimensional grid of maximum gridpoint diagnostic values was computed using model output from forecast hours 12–36 (1200 UTC–1200 UTC). A threshold was applied to convert the 24-h maximum field into a binary grid of ones and zeroes; these are the locations of the native model grid SSRs. The native grid SSRs were upscaled to an 80-km grid² of SSRs, by flagging an 80-km grid box if at least one 3- or 1-km grid SSR occurred within the 80-km grid box. For the rest of this work, each forecast's 80-km grid SSRs are used. The usage of the maximum diagnostic value over the 24-h forecast period eliminates most timing errors, focusing only on where 3- and 1-km forecasts predict severe weather on a given day, determined by the SSRs.

The diagnostics used here, which all incorporate vertical vorticity, exhibit large sensitivities to Δx (Adlerman and Droegemeier 2002), preventing the usage of a fixed threshold between the 3- and 1-km forecasts. To equitably compare the 3- and 1-km forecasts, thresholds were chosen so the total number of SSRs, when summed

² An 80-km grid was chosen to be consistent with SPC's probabilistic outlooks, which forecast event occurrence within 25 mi of a point, as well as reducing the impact of under- or over-reporting biases of severe storm reports (e.g., Weiss and Vescio 1998). The specific grid used is the NCEP 211 grid, which has a grid-spacing of 80 km at 35°N latitude.

TABLE 2. WRF parameterization schemes used in both 3- and 1-km forecasts.

| Parameterization | WRF option | Reference |
|----------------------------------|---|---|
| Cumulus | None | |
| Microphysics | Thompson | Thompson et al. (2008) |
| Planetary boundary layer | Mellor–Yemada–Janjić (MYJ) | Mellor and Yamada (1982); Janjić (1994, 2002) |
| Longwave and shortwave radiation | Rapid Radiative Transfer Model for Global Climate Models (RRTMG) with ozone and aerosol climatologies | Mlawer et al. (1997); Iacono et al. (2008); Tegen et al. (1997) |
| Land surface model | Noah | Chen and Dudhia (2001) |

across all 497 events, was equal to the total number of OSRs (i.e., 3096 OSRs and SSRs, an SSR bias of 1; Table 3). Additional thresholds were also chosen at SSR biases between 2.0 and 0.25, in increments of 0.25, producing total SSR counts between 6192 and 774, in increments of ~800 (Table 3). Each 24-h SSR field was smoothed using an isotropic Gaussian filter (Sobash et al. 2011; Hitchens et al. 2013) with standard deviation σ between 20 and 300 km, in 20-km increments. After smoothing, the binary SSR field values fall between 0 and 1 and can be interpreted as a probability (Theis et al. 2005); this product is henceforth referred to as a surrogate severe probability forecast (SSPF) for tornadoes.

d. Verification

SSPFs were verified using National Centers for Environmental Information Storm Data tornado reports, contained within the SPC storm report database. Additionally, NWS tornado warnings were used for verification since the rotation diagnostics should be more skillful as surrogates for intense low-level rotation than for tornadoes. The start location of each tornado report was mapped to an 80-km grid and aggregated over each 24-h forecast period, to produce an OSR field analogous to the SSR field. Tornado warnings

were retrieved from the Iowa Environmental Mesonet website (<https://mesonet.agron.iastate.edu>) and, like the reports, were mapped to an 80-km grid using the centroid of each warning rather than the report start location. If the tornado warning was valid for any portion of the SSPF period, it was included in the verification data. Tornado warnings and reports were then merged to produce a gridded 80-km dataset of combined tornado report and tornado warning locations.

SSRs were verified with OSRs using contingency table metrics such as the probability of detection (POD) and false alarm ratio (FAR). POD and the success ratio (SR; $1 - FAR$) for each SSR threshold were plotted on a performance diagram (Roebber 2009) to visually assess forecast quality. SSPFs were verified with the fractions skill score (FSS; Roberts and Lean 2008), area under the relative operating characteristic (ROC) curve (AUC; Mason 1982; Marzban 2004), and attributes diagrams, which assess forecast reliability (Wilks 2011). For each FSS curve, the scale where the FSS exceeded $0.5 + f_o/2$, where f_o is the sample climatology (0.008), was computed via interpolation between length scales. This scale represents the minimum useful scale of the forecast (l_{min} ; Roberts and Lean 2008) and is defined as the FSS of a uniform forecast of the climatological event

TABLE 3. Thresholds for each diagnostic that produce SSR biases (number of SSRs/number of OSRs) between 2.0 and 0.25. Percentiles listed are for distributions that included all 80-km model grid points within the verification region, including zeros, for forecast hours between 13 and 36. Bold numbers indicate the thresholds that produce an SSR bias of 1.

| Bias | 2.0 | 1.75 | 1.5 | 1.25 | 1.0 | 0.75 | 0.5 | 0.25 |
|---|--------|--------|--------|--------|---------------|--------|--------|--------|
| No. of SSRs | 6192 | 5418 | 4644 | 3870 | 3096 | 2322 | 1548 | 774 |
| Percentile | 99.794 | 99.820 | 99.845 | 99.871 | 99.897 | 99.922 | 99.948 | 99.967 |
| 3-km UH25 ($m^2 s^{-2}$) | 127.3 | 135.55 | 145.25 | 156.65 | 170.95 | 191.8 | 222.3 | 275.45 |
| 3-km UH03 ($m^2 s^{-2}$) | 78.6 | 83.25 | 89.0 | 96.15 | 105.0 | 116.8 | 135.75 | 169.85 |
| 3-km UH01 ($m^2 s^{-2}$) | 10.937 | 11.55 | 12.28 | 13.197 | 14.362 | 16.002 | 18.299 | 22.71 |
| 3-km RVORT1 ($\times 10^{-3} s^{-1}$) | 9.176 | 9.443 | 9.745 | 10.107 | 10.555 | 11.11 | 11.943 | 13.362 |
| 3-km RVORT0.5 ($\times 10^{-3} s^{-1}$) | 8.193 | 8.451 | 8.734 | 9.06 | 9.473 | 9.999 | 10.765 | 12.332 |
| 1-km UH25 ($m^2 s^{-2}$) | 670.1 | 705.9 | 746.9 | 794.4 | 861.0 | 948.1 | 1079.0 | 1307.1 |
| 1-km UH03 ($m^2 s^{-2}$) | 423.0 | 449.0 | 480.2 | 516.1 | 566.2 | 630.6 | 722.4 | 917.5 |
| 1-km UH01 ($m^2 s^{-2}$) | 76.02 | 80.36 | 85.37 | 91.88 | 100.03 | 111.76 | 128.73 | 162.6 |
| 1-km RVORT1 ($\times 10^{-3} s^{-1}$) | 25.4 | 26.14 | 26.975 | 28.095 | 29.435 | 30.99 | 33.27 | 37.58 |
| 1-km RVORT0.5 ($\times 10^{-3} s^{-1}$) | 25.425 | 26.19 | 27.035 | 28.095 | 29.5 | 31.295 | 33.765 | 38.77 |

frequency f_o . ROC curves were constructed by computing POD and probability of false detection (POFD) for probability thresholds between 0% and 100% in 5% increments, as well as at 1% and 2%. The ROC AUC was then computed with the trapezoidal approximation. All metrics were computed using CONUS land points east of 105°W (i.e., east of the Rocky Mountains). SSRs and OSRs outside of the verification domain were removed prior to producing the SSPFs, to avoid the issue of SSRs outside the verification domain impacting interior verification points after smoothing.

Statistical significance of the FSS and ROC AUCs were determined by a bootstrap technique applied to differences between pairs of experiments (e.g., Wolff et al. 2014). Specifically, for a particular forecast pair (e.g., 1-km UH01 SSPFs and 3-km UH01 SSPFs) distributions of FSS or ROC AUC differences were constructed by randomly sampling the 497 cases, with replacement, and computing the FSS or ROC AUC differences using those dates for each set of forecasts. This procedure was repeated 10 000 times to estimate bounds of 99% confidence intervals (CIs). Using a one-tailed hypothesis test, if zero was not included within bounds of the 99% CIs, then differences were statistically significant at the 99% level or higher. The permutation test outlined in Hamill (1999) was also used, producing similar results to the paired difference test.

3. Verification of tornado SSPFs at 3- and 1-km Δx

Verification of the 3- and 1-km binary SSRs and probabilistic SSPFs is presented on the 80-km grid scale for all SSR threshold values for each diagnostic to examine the impact of SSR bias on forecast verification using contingency table based metrics. SSPF evaluations will be restricted to SSPFs produced with an SSR bias of 1 (Table 3) using probabilistic metrics (e.g., FSS, ROC AUC, attributes diagram). The FSS evaluates the scale dependence of skill up to smoothing length scales of ~ 300 km,³ covering most of the mesoscale.

a. Verification of SSRs and SSPFs with tornado reports

Evidence of differences between the 3- and 1-km forecasts existed on the 80-km grid scale, when verifying SSRs against tornado OSRs. For a given threshold,

³ This length scale is the standard deviation of the Gaussian σ used to smooth the forecasts, differing from the traditional “box width” neighborhood presented in Roberts and Lean (2008). In general, the corresponding box width neighborhood will be larger than the Gaussian standard deviation.

the accuracy of the SSRs increased as the grid spacing was reduced and the depth of UH integration was positioned closer to the surface, in both the 1- and 3-km forecasts, with the decreased grid spacing having a larger impact on forecast skill than the change in UH integration depth (Fig. 2). Overall, the 1-km UH01 and UH03 SSRs produced the largest PODs and smallest FARs among the forecasts, resulting in the largest CSIs (Fig. 2). The 1-km UH03 SSRs were more accurate than the 1-km UH01 SSRs at large UH thresholds, although the number of SSRs is quite small (~ 500 – 1000). Among the 3-km forecasts, the UH01 SSRs were the most accurate, producing the largest CSIs, while the 3-km UH25 SSRs had the lowest CSIs among all SSPFs. Of the 8 thresholds used to produce SSRs in each of the forecasts, the CSI increased as the UH threshold was reduced, with the maximum CSI produced for the smallest UH threshold (i.e., an SSR bias of 2.0), although CSI differences were generally small for SSR biases between 1.0 and 2.0. CSI values increased slightly for biases greater than 2.0, then decreased (not shown). The improvement of CSI for biases greater than one is a result of the sensitivity of CSI to bias, with overforecasting, or “hedging” being rewarded, with this effect being particularly acute in the verification of rare events, such as tornadoes (Baldwin and Kain 2006).

Even though thresholds were selected so that UH25 and UH01 produced the same number of SSRs (Table 3), differences in regional coverage of SSRs contributed to differences in forecast accuracy. UH25 SSRs were overpredicted across the central United States and underpredicted across the eastern United States, when compared to tornado OSRs (Fig. 3a). In both the 3- and 1-km forecasts, UH01 better identified the locations of tornado OSRs regionally, with a reduction in SSRs over the central United States and an increase within the east and southeast United States, in better correspondence with tornado OSRs (Fig. 3b). Tornadoic environments in the east and southeast United States often differ from those in the central United States, with eastern U.S. tornadoes more likely to occur within high-shear, low-instability regimes (Guyer and Dean 2010; Dean and Schneider 2008). Supercells in these environments are often spatially compact (Davis and Parker 2014), with reduced UH25 magnitudes (Guyer and Jirak 2014). While regional or environmental varying thresholds could improve the performance of UH25 tornado SSPFs in the southeastern United States (Sobash and Kain 2017), it appears that UH01 can better discriminate between tornadoes in these two regimes.

Among the three diagnostics that were used to identify rotation < 1 km AGL (i.e., RVORT1, RVORT0.5, and UH01), UH01 SSRs were more closely associated

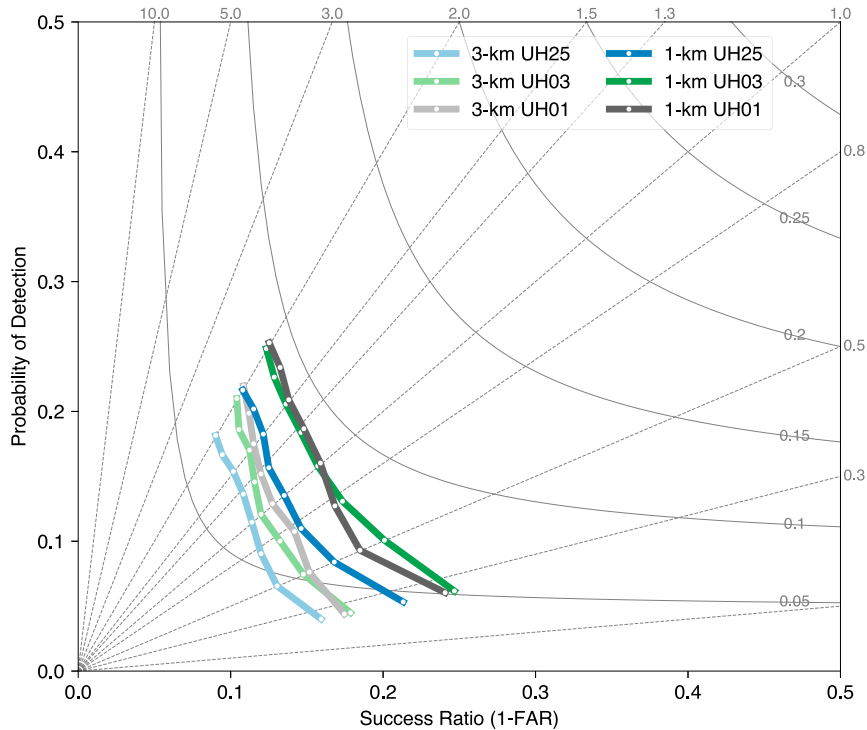


FIG. 2. Performance diagram for the 3-km UH25, 3-km UH03, 3-km UH01, 1-km UH25, 1-km UH03, and 1-km UH01 SSRs. The POD and SR for the eight bias-determined UH thresholds (Table 4) are marked by a white dot within each line, with the lowest UH threshold (bias of 2.0) producing the largest POD. Solid gray lines are lines of constant CSI, while dashed gray lines are lines of constant bias.

with tornado OSRs than RVORT1 or RVORT0.5 and had larger CSI values at both 1- and 3-km Δx across all thresholds (Fig. 4). While RVORT1 and RVORT0.5 are components of the UH01 computation, and due to this are strongly correlated with UH01, UH01 appears to have an advantage in that it identifies model grid points where intense vertical vorticity is collocated with an updraft, such as within a mesocyclone, rather than only identifying the presence of intense vertical vorticity. The usage of a nine-point smoother when computing UH01, and not when computing RVORT1 or RVORT0.5, may also contribute to some skill differences. Between RVORT1 and RVORT0.5, RVORT0.5 was less skillful than RVORT1 for all thresholds. The reduction in skill as the vertical vorticity level decreased from 1 km to 500 m AGL is potentially due to RVORT1 being more reflective of the presence of a low-level mesocyclone, while large RVORT0.5 magnitudes may arise due to horizontal shear along surface boundaries, unrelated to the presence of rotation aloft.

Differences between the 1- and 3-km SSRs were also present on larger spatial scales, when smoothed into SSPFs and compared to the smoothed OSR field with the FSS. Overall, FSS differences between the SSPFs

increased with spatial scale, with differences due to grid spacing and diagnostic choice being maximized on the mesoscale (Fig. 5). The FSS magnitudes for the 80-km gridscale SSPFs (i.e., those produced with minimal smoothing) were not useful (FSSs < 0.5), with small differences among the SSPFs, corroborating the gridscale SSR evaluation. The spatial scale that produced skillful forecasts (l_{\min}) was smallest for the 1-km SSPFs and largest for the 3-km SSPFs (Table 4). For example, l_{\min} for the 1-km UH01 SSPFs was 100 km, whereas the 3-km UH25, UH03, and UH01 SSPFs had l_{\min} values of 365, 195, and 140 km, respectively (Table 4). The 1-km UH01 SSPFs were most skillful at all scales, with the combined effects of the low-level mesocyclone diagnostic and finer resolution combining to produce the largest increases in forecast skill over the 3-km UH25 SSPFs (Fig. 5). FSS differences between forecasts using the same grid spacing (e.g., 1-km UH01 and 1-km UH25) or the same diagnostic (e.g., 1-km UH01 and 3-km UH01), were all statistically significant at the 99% level at all length scales. As in the gridscale results, the RVORT0.5 and RVORT1 SSPFs had inferior FSS to the UH01 SSPFs at any scale, with the disparity in skill increasing with scale (not shown). The RVORT1 and

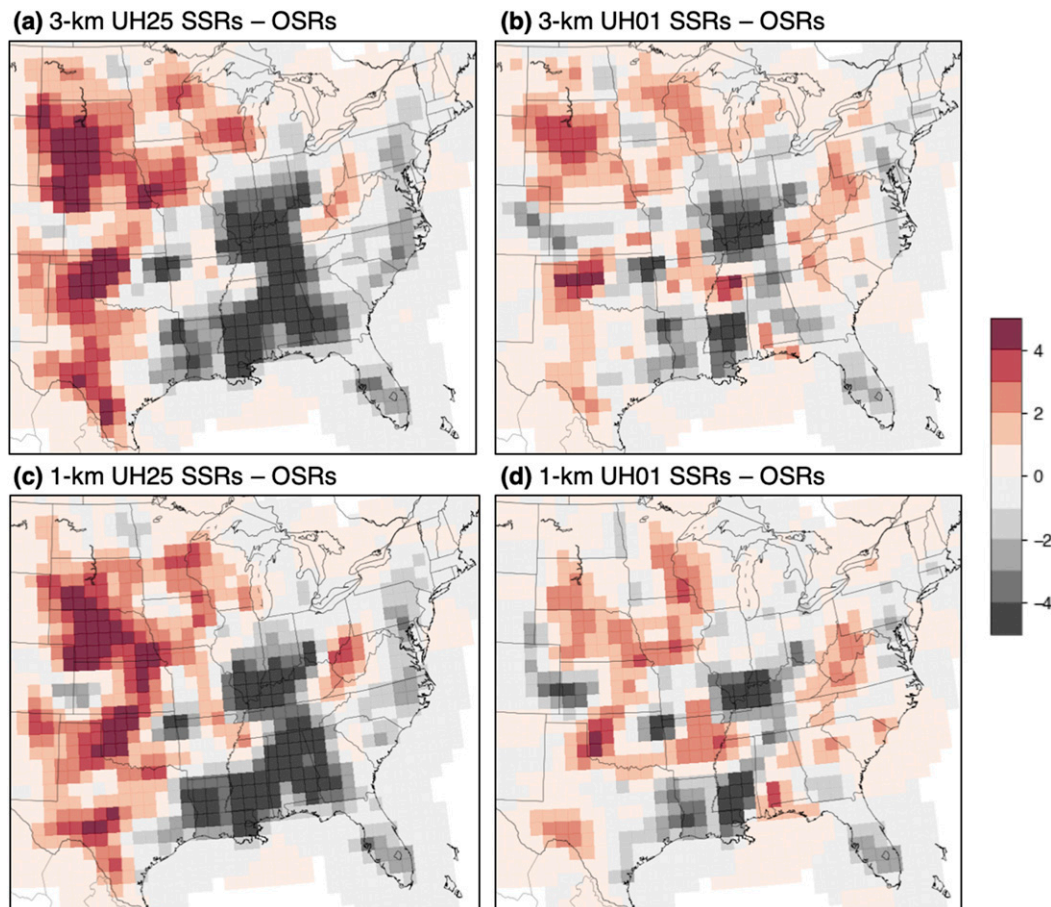


FIG. 3. Difference between total number of SSRs and tornado OSRs at each 80-km grid box accumulated over all 497 forecasts for the (a) 3-km UH25, (b) 3-km UH01, (c) 1-km UH25, and (d) 1-km UH01 diagnostics using the threshold for each field that produced a bias of 1 (Table 4).

RVORT0.5 SSPFs will not be evaluated further, since these diagnostics do not appear to provide any added value over UH01.

Two independent aspects of probabilistic forecast skill, reliability, and resolution, were evaluated for the UH01 and UH25 SSPFs (the UH03 SSPFs generally produced results falling between the UH25 and UH01 SSPFs; Fig. 6). All SSPFs produced overconfident probability values at small length scales (Fig. 6a). Providing smoothed SSPF guidance was necessary to produce reliable probabilities for all forecast sets, although the 3-km UH25 SSPFs were not reliable even with fairly aggressive smoothing (Fig. 6d). For SSPF probabilities $< 40\%$ the 1-km UH01 SSPFs tended to be most reliable and required less smoothing to achieve a given level of reliability, retaining the most sharpness in the forecasts. For SSPF probabilities $> 40\%$, the observed frequencies were often sensitive to the small sample sizes, especially when SSPFs were smoothed using $\sigma > 160$ km, although the 1-km UH01 SSPFs were still the most reliable.

As for forecast resolution, the ROC AUC was largest for the 1-km UH01 SSPFs and smallest for the 3-km UH25 SSPFs, with differences ≥ 0.1 on scales > 80 km (Fig. 7). Similar to the gridscale and FSS results, ROC AUCs increased as grid spacing was decreased and as the UH diagnostic incorporated model levels closer to the surface. The 1-km UH01 SSPFs had ROC AUCs ≥ 0.8 for length scales > 140 km, indicating good ability to discriminate between events and nonevents, while the 3-km UH25 SSPFs had ROC AUCs ~ 0.7 for the same length scales (Fig. 7). ROC AUCs were reduced to near 0.5 for all forecasts at the 80-km grid scale, since the SSPFs were minimally smoothed and only consisted of 0% or 100% probability values.

The usage of low-level UH diagnostics (e.g., UH01) and the finer grid spacing within the 1-km forecasts both resulted in improvements to forecast reliability and resolution. Improvements in forecast resolution are often more difficult to achieve, since they are fundamentally related to the underlying value of the forecast system

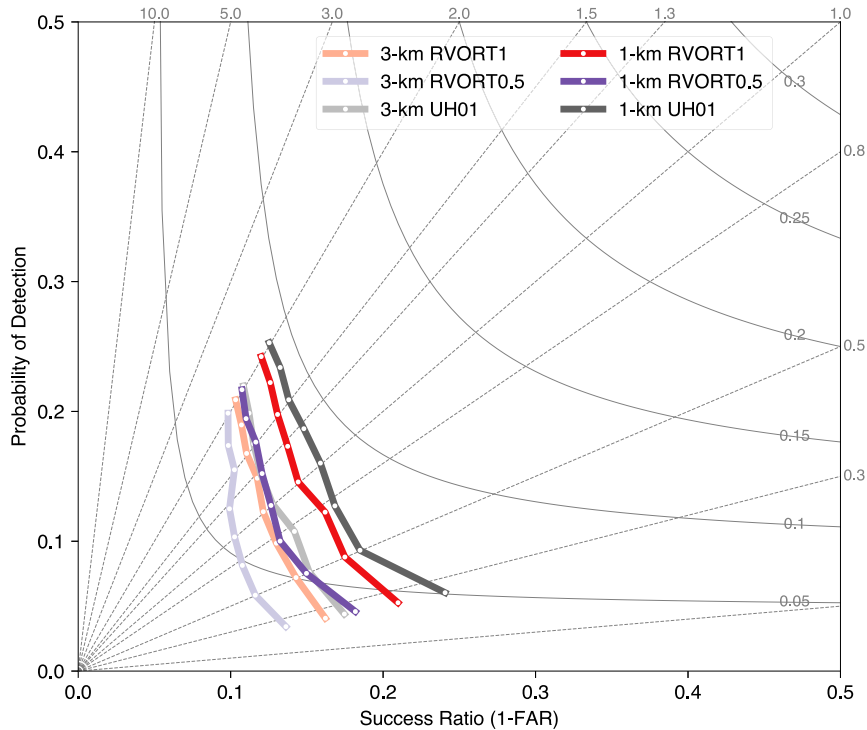


FIG. 4. As in Fig. 2, but for 3-km UH01, 3-km RVORT1, 3-km RVORT0.5, 1-km UH01, 1-km RVORT1, and 1-km RVORT0.5 SSRs.

(rather than reliability, which can be calibrated without changing the underlying components or configuration of the system). Here, the choice of diagnostic and grid spacing appear to both be factors in increasing forecast resolution, improving the underlying ability of CAMs to discriminate between tornadic and non-tornadic events.

b. Verification of SSRs and SSPFs with tornado warnings

In this section, we use the most skillful tornado forecasts from the previous section as a baseline (i.e., 1-km UH01 SSPFs) and evaluate differences in skill when verifying against tornado reports and warnings, rather than tornado reports alone. The inclusion of tornado warnings to the tornado OSRs approximately doubled the total number of “hit” 80-km grid boxes (i.e., 6420 versus 3096 OSRs), indicating a significant number of false alarm tornado warnings. This is not unexpected, considering NWS tornado warnings typically have FARs of ~ 0.7 (Brooks and Correia 2018). Even though many tornado warnings did not verify, they still provide valuable information about the existence of radar-indicated rotation of sufficient strength to issue a tornado warning, assuming sufficient radar coverage.

The larger number of OSRs when verifying with tornado reports and warnings required that the UH01 threshold be decreased to $75 \text{ m}^2 \text{ s}^{-2}$, producing 6420 SSRs, a bias of 1. The 1-km UH01 SSPFs corresponded better with the locations of tornado reports and warnings combined, compared to reports alone, reflected by larger FSSs across all scales (Fig. 8a). In fact, the 1-km UH01 SSPFs verified with both reports and warnings had the largest FSSs among all SSPFs examined in this work, with FSSs > 0.7 for length scales $> 200 \text{ km}$ and l_{\min} of $\sim 70 \text{ km}$. The improvement in FSS when combining tornado reports with warnings with the 1-km UH01 SSPFs was similar for the 3-km UH01, 3-km UH25, and 1-km UH25 SSPFs; verifying with reports and warnings for all forecasts increased FSS by ~ 0.1 across most scales (not shown). The larger FSS values were also associated with larger AUCs, indicating that the differences in skill were in part a result of better discrimination between events and nonevents and not just better calibration (Fig. 8b). The FSS and ROC AUC differences between forecasts verified with reports, or reports and warnings, were statistically significant at the 99% confidence level. Given the higher FSSs and larger AUCs for forecasts verified with tornado warnings, it is safe to say that UH01 is a more skillful proxy for the

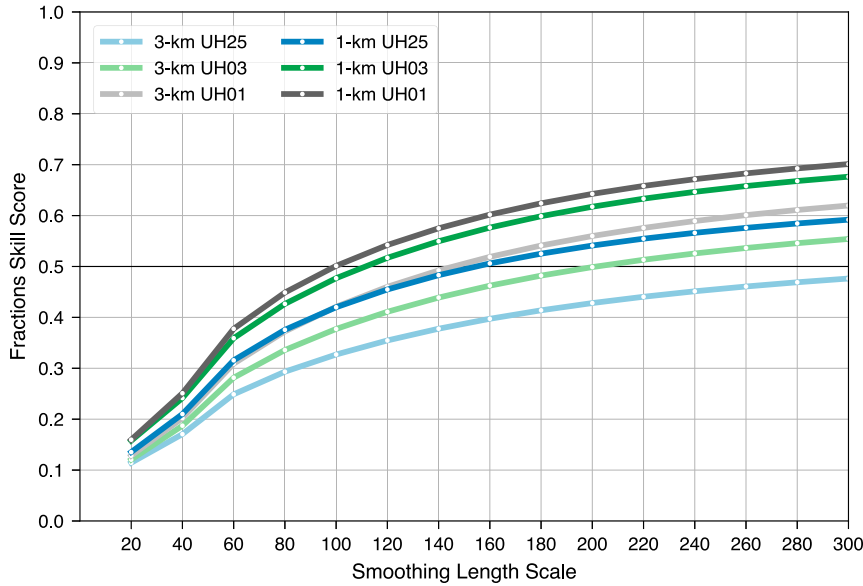


FIG. 5. Fractions skill score as a function of smoothing length scale σ (km) for the 3-km UH25, 3-km UH03, 3-km UH01, 1-km UH25, 1-km UH03, and 1-km UH01 tornado SSPFs. The black horizontal line at $FSS = 0.5 + f_o/2$, where f_o is the sample climatology (0.008), indicates the minimum useful scale.

locations where intense low-level rotation will occur, and tornado warnings will be issued on a given forecast day, rather than the locations where tornado reports will occur.

c. Verification of filtered SSRs and SSPFs

Previous work has applied the STP as a filter that improved the skill of CAM-based tornado forecasts by removing convection in environments not conducive for tornadoes (i.e., those with high LCLs or weak low-level shear; Thompson et al. 2012; Gallo et al. 2016, 2018). To test if removing SSRs in environments of low STP improved forecast skill, SSPFs were generated from a subset of the SSRs that occurred in environments where the STP was larger than 1.0. Specifically, at each hour, the representative STP within each 80-km grid box was computed by averaging the 3- or 1-km gridpoint STP values composing the grid box. Each hour’s binary SSRs were removed where the average STP within the 80-km grid box was <1.0 in the previous hour. STP thresholds of 1–4 were examined, and FSS values were maximized using an STP threshold of 1 in both the 1- and 3-km SSPFs. The filtered SSRs were smoothed using the same procedure to create the original SSPFs. As in the section 3b, we examine only the 3- and 1-km UH25 and UH01 SSPFs.

Using the UH thresholds in Table 3 that produced an SSR bias of 1 underforecast the number of OSRs after

removing SSRs where $STP < 1.0$. To compensate, the UH threshold was decreased to produce an SSR bias of 1 among the four sets of SSPFs that were filtered, that is, the UH thresholds used to produce the filtered 3-km UH25, 1-km UH25, 3-km UH01, and 1-km UH01 SSRs were 110.6, 631.3, 9.87, and 69.03 $m^2 s^{-2}$, respectively. Filtering resulted in statistically significant improvements in FSS and AUC for the 3-km UH25, 3-km UH01, and 1-km UH25 SSPFs (Fig. 9). FSS and AUC differences for

TABLE 4. Minimum useful scale l_{min} (km) among the various forecasts and verification datasets used in this work (ALL represents verification against hail, wind, and tornado reports, while TOR represents verification with tornado reports only).

| | Verification dataset | Minimum useful scale l_{min} (km) |
|---------------|----------------------|-------------------------------------|
| 1-km UH25 | ALL | 40 |
| 1-km UH01 | ALL | 45 |
| 3-km UH25 | ALL | 45 |
| 3-km UH01 | ALL | 50 |
| 1-km UH01 | TOR | 100 |
| 1-km UH03 | TOR | 110 |
| 1-km RVORT1 | TOR | 120 |
| 1-km RVORT0.5 | TOR | 135 |
| 3-km UH01 | TOR | 140 |
| 1-km UH25 | TOR | 150 |
| 3-km RVORT1 | TOR | 155 |
| 3-km RVORT0.5 | TOR | 170 |
| 3-km UH03 | TOR | 195 |
| 3-km UH25 | TOR | 365 |

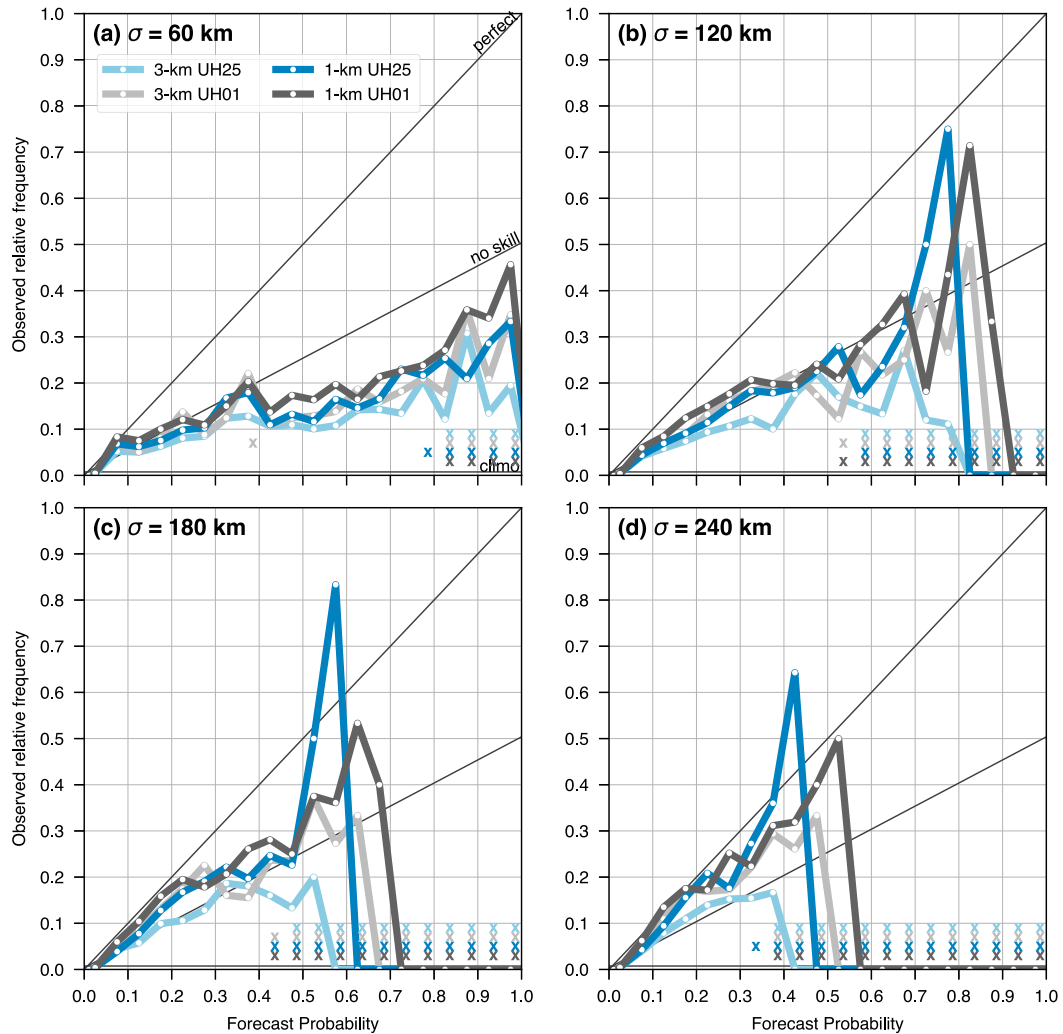


FIG. 6. Attributes diagrams for 3-km UH25, 3-km UH01, 1-km UH25, and 1-km UH01 tornado SSPFs with σ of (a) 60, (b) 120, (c) 180, and (d) 240 km using the UH threshold that produces an SSR bias of 1. The “x” markers denote where the number of SSPF grid boxes within the forecast probability bin is less than 100.

the 1-km UH01 SSPFs were small and not statistically significant (Fig. 9c,d). The 3-km UH25 SSPFs benefited the most from filtering, especially at large scales, since both the 3-km (compared to the 1-km) and UH25 (compared to UH01) forecasts preferentially benefited (Figs. 9c,d).

While the 1-km and UH01 SSPFs exhibited less need for supplementary environmental information, the inclusion of this environmental information in the 3-km SSPFs closed the skill gap between the 1-km and 3-km UH01 SSPFs (cf. Figs. 5 and 9). Specifically, STP-filtered 3-km UH01 SSPFs possessed more similar FSSs (Fig. 9a) and AUCs (Fig. 9b) to 1-km UH01 SSPFs compared to unfiltered SSPFs. In fact, the filtered 1-km UH01 SSPFs possessed statistically significant FSS differences to the filtered 3-km UH01 SSPFs only at scales ≤ 140 km; FSS

differences between these two filtered sets of SSPFs were *not statistically significant at scales > 140 km* (FSS differences between filtered 3-km UH01 SSPFs and unfiltered 1-km UH01 SSPFs were not statistically significant at *any* length scale). This was not true for the UH25 SSPFs; that is, the filtered 1-km UH25 SSPFs produced statistically significant improvements in FSS over the filtered 3-km UH25 SSPFs across all length scales (cf. Figs. 5, 9a), although ROC AUCs did not possess statistically significant differences (cf. Figs. 7, 9b).

Together, these results highlight the benefits of filtering UH25 and UH01 SSPFs from 3-km CAMs, as was done in Gallo et al. (2018; 2019), to improve the skill of next-day tornado forecasts. Assuming the availability of both UH25 and UH01 diagnostics in contemporary

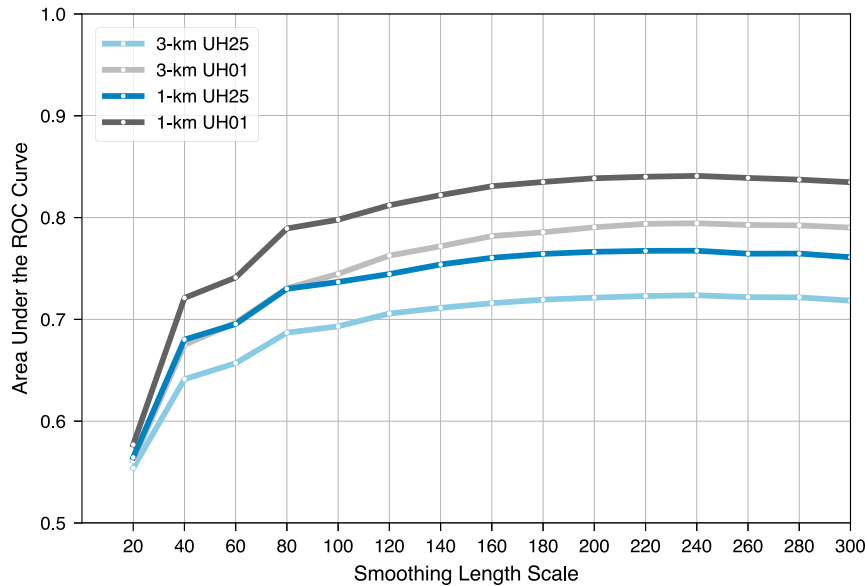


FIG. 7. Area under the relative operating characteristic curve as a function of σ for 3-km UH25, 3-km UH01, 1-km UH25, and 1-km UH01 tornado SSPFs for the UH threshold that produced an SSR bias of 1 (Table 4).

3-km CAMs, tornado SSPFs should incorporate UH01 with STP information to produce tornado SSPFs that are potentially competitive with tornado SSPFs derived from 1-km CAMs.

d. Daily variations in UH01 SSPF skill

When aggregated across all 497 cases, the 1-km UH01 SSPFs outperformed the 3-km UH01 SSPFs, especially at larger scales (e.g., Fig. 5). To examine if these aggregate statistics produced noticeable improvements in daily UH01 SSPF skill, the FSS was computed individually for the 497 SSPFs for the 120-km length scale. For this analysis, we removed days where no tornado OSRs occurred and no SSRs were present in either of the 1- and 3-km forecasts (the FSS would be undefined in this case); this occurred in 34 SSPFs. The remaining 463 forecast pairs were used to compute daily FSS differences between the 1-km UH01 and 3-km UH01 SSPFs.

The 202 forecast days (41% of forecasts) had UH01 SSPF FSS differences $< \pm 0.05$, or essentially no difference in skill between the 1- and 3-km UH01 SSPFs (Fig. 10). SSPFs within these bins either produced similar SSPFs or were events where no tornado OSRs occurred (FSSs will be zero for both sets of forecasts if no OSRs occurred). Of the 261 SSPFs that possessed FSS differences $> \pm 0.05$, the 1-km UH01 SSPFs were more skillful 63% of the time (163 forecasts). The largest FSS differences occurred preferentially in favor of the 1-km UH01 SSPFs. For example, $\sim 70\%$ of UH01 SSPFs with FSS differences $> \pm 0.3$ were associated with increased

skill of the 1-km UH01 SSPFs (Fig. 10). In other words, when the difference in skill between the 3- and 1-km SSPFs was large, the 1-km SSPFs were more likely to be superior, with the 3-km SSPFs only producing more skillful forecasts in a small fraction of cases. At scales larger than 120 km, where the aggregate FSS differences were largest, the distributions of daily FSS differences were similar to those at $\sigma = 120$ km. The daily FSS differences suggest that not all forecasts benefit equally from 1-km Δx , and in some cases, forecast skill was reduced in the 1-km forecasts compared to the 3-km forecasts. Work is ongoing to understand these variations in forecast skill as a function of Δx , as the benefits of reducing Δx from 3 to 1 km may occur in specific environmental situations. While 1-km UH01 SSPFs are more likely to produce larger daily FSS values than 3-km UH01 SSPFs, future work should evaluate if these objective differences translate into subjective differences that would be apparent to forecasters.

4. Verification of all-hazard SSPFs at 3- and 1-km Δx

The 1-km SSPFs evaluated in section 3 provided statistically significant improvements in forecast skill compared to the 3-km SSPFs when predicting the locations of tornadoes (and tornado warnings). Yet, previous studies have noted that 1-km Δx did not improve next-day severe weather forecasts of all severe phenomena (i.e., tornadoes, large hail, or strong wind gusts),

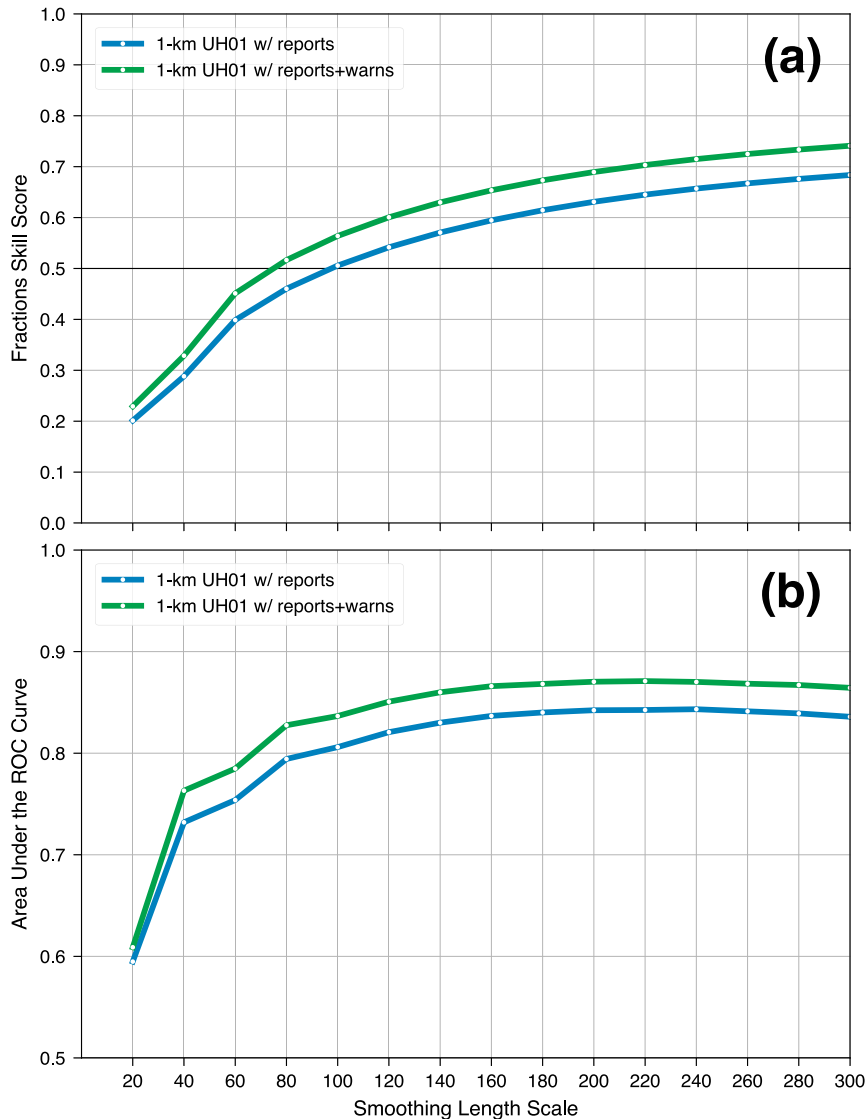


FIG. 8. (a) As in Fig. 5 and (b) as in Fig. 8, but for 1-km UH01 verified with only tornado reports (blue) and the combination of tornado reports and warnings (green).

both in subjective evaluations (e.g., Kain et al. 2008; Clark et al. 2012) and when verifying UH25 SSPFs (e.g., Loken et al. 2017). To assess if the present 1-km SSPFs produce more skillful all-severe hazard guidance than the 3-km SSPFs, we produced 3- and 1-km UH25 and UH01 SSPFs with adjusted thresholds based on the number of all-severe OSRs (27 445). The 3-km UH25, 3-km UH01, 1-km UH25, and 1-km UH01 thresholds that produced an SSR bias of 1 were 53.57, 4.715, 322.25, and $34.63 \text{ m}^2 \text{ s}^{-2}$, respectively. These SSPFs were verified similarly to tornado reports, but included reports of hail ≥ 1 in. in diameter and wind gusts ≥ 50 kt. The storm reports were mapped to the 80-km grid and aggregated over the 24-h forecast period.

The 1-km UH25 SSPF FSSs and ROC AUCs were larger at all scales than the 3-km UH25 SSPFs, with both SSPFs having l_{\min} values between 40 and 60 km (Fig. 11; Table 4). While FSS differences were small, they were statistically significant at all scales except 300 km, while ROC AUC differences were statistically significant at all smoothing length scales. While the ROC AUC magnitudes and differences between the 1-km UH25 SSPFs and the 3-km UH25 SSPFs were similar to those reported in Loken et al. (2017), they concluded based on similar resampling tests to those used here that the differences were not statistically significant at any of their UH25 thresholds. To test the impact of sample size, the FSS and ROC AUC

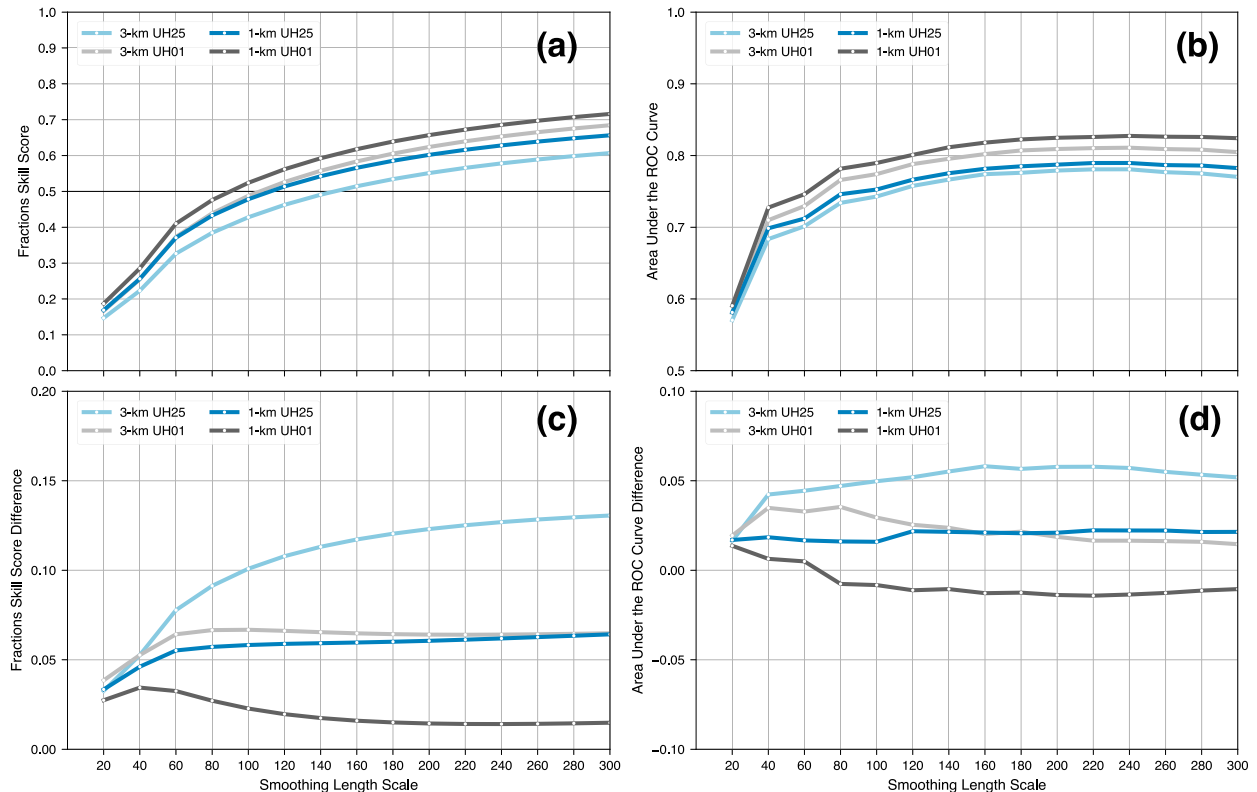


FIG. 9. (a) FSS and (b) ROC AUC for 3-km UH25, 3-km UH01, 1-km UH25 and 1-km UH01 STP-filtered SSPFs, along with (c) FSS and (d) ROC AUC differences between STP-filtered and unfiltered SSPFs (FSSs and ROC AUCs for unfiltered FSSs are found in Figs. 5 and 8, respectively). The 3- and 1-km UH thresholds were reduced as described in section 3c to produce STP-filtered SSR biases of 1.

statistics for the 1-km UH25 and 3-km UH25 SSPFs were computed for a subset of 58 forecasts between 1 April 2011 and 30 June 2011, corresponding to a portion of the 63 forecasts included in Loken et al. (2017). When evaluating this subset, the 1-km UH25 SSPFs possessed larger FSSs and ROC AUCs, with the ROC AUCs deemed statistically significant for all scales (not shown), while FSSs were not statistically significant except for the 300-km scale. Given the small differences in FSS and ROC AUC between the 1- and 3-km UH25 SSPFs, it may be that these differences in skill are not *practically* significant, even though statistical significance may exist.

Finally, while the UH01 SSPFs were more skillful than the UH25 SSPFs at both 1- and 3-km Δx for tornadoes, they were *less skillful* than UH25 SSPFs at both Δx when used as a surrogate for all severe hazards (Fig. 11). The reduction in FSS when using UH01 compared to UH25 was larger than the reduction in FSS when going from 1- to 3-km Δx , with skill differences being statistically significant at all scales for 3-km Δx and for scales ≤ 200 km for 1-km Δx . In other words, UH25 was most appropriate to construct all-severe

hazard SSPFs, whereas UH01 was best suited for tornado SSPFs. The degraded skill when using UH01 to anticipate all-severe hazards is likely due to a weaker correspondence between the presence of a low-level mesocyclone, which UH01 is designed to detect, and hazards such as large hail and intense wind gusts,

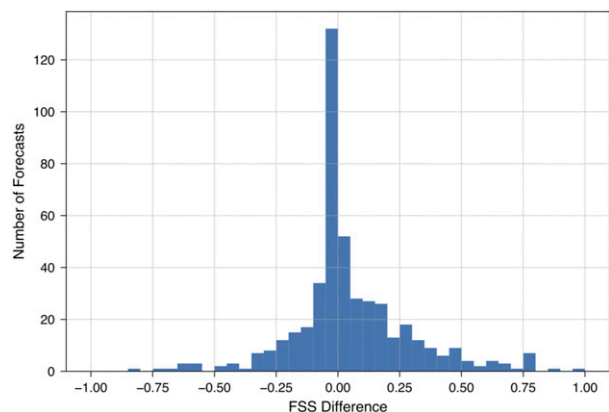


FIG. 10. Histogram of daily 1-km UH01 minus 3-km UH01 SSPF FSS differences for $\sigma = 120$ km.

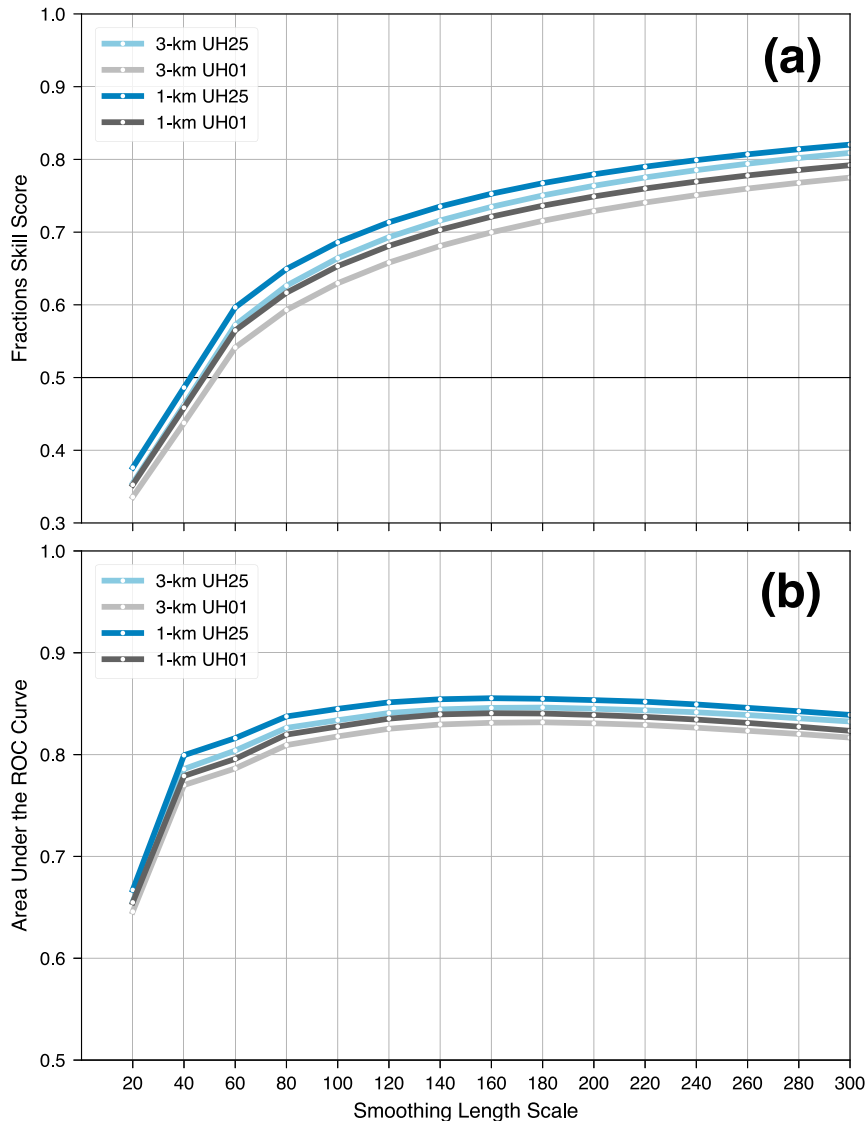


FIG. 11. As in Fig. 9, but for 3-km UH25, 3-km UH01, 1-km UH25, and 1-km UH01 SSPFs verified with all-severe OSRs. The threshold for each diagnostic, provided in the text, was chosen to produce the same number of SSRs as the total number of all severe OSRs.

compared to the more robust relationship between these hazards and the presence of a midlevel mesocyclone (i.e., UH25).

5. Summary

To assess the benefits of finer Δx on next-day forecasts of severe weather hazards such as tornadoes, WRF-ARW forecasts were produced for 497 high-impact severe weather events over the CONUS using 3- and 1-km Δx . For each forecast hour from 13 to 36 (i.e., 1200–1200 UTC), diagnostics related to low-level (UH01, UH03, RVORT1, and RVORT0.5) and midlevel

rotation (UH25) were used as surrogates for the occurrence of a tornado in the 3- and 1-km output. Thresholds were applied to make decisions on where tornadoes would occur on a given forecast day, with thresholds chosen based on the total number of tornado OSRs that occurred during the 497 forecasts. The locations where the surrogate diagnostics exceeded the threshold were upscaled to an 80-km grid, aggregated over a 24-h forecast period, and smoothed to produce SSPFs for tornadoes. SSPFs were verified against both observed tornado reports and NWS tornado warnings using gridscale and scale-dependent metrics. Additionally, SSPFs of all-severe hazards were created and

verified against all observed severe weather reports. The primary conclusions are summarized below:

- The 1-km tornado SSPFs were more skillful than the 3-km tornado SSPFs for the five diagnostics examined in this work, with statistically significant differences in skill occurring at all scales. Among the five diagnostics, UH01 produced the most skillful next-day tornado SSPFs at both 3- and 1-km Δx , possessing larger FSSs, better reliability, and larger ROC AUCs than 3-km tornado SSPFs. The superiority of UH01 among the five diagnostics was partly due to its ability to detect tornadic convection in presumably different environmental regimes. While the 1-km UH01 tornado SSPFs were the most skillful, their minimum useful scale was ~ 110 km, necessitating smoothing to produce useful forecast guidance.
- The 3-km UH01 tornado SSPFs were improved by filtering the binary SSRs by removing SSRs where the STP was less than 1.0, while the 1-km UH01 tornado SSPFs did not benefit as greatly. The filtered 3-km UH01 SSPFs were of similar quality to the unfiltered 1-km UH01 SSPFs, with the differences in the filtered 3-km UH01 SSPFs and the unfiltered 1-km UH01 SSPFs not statistically significant (the filtered 1-km UH01 SSPFs were statistically better than the filtered 3-km UH01 SSPFs at small scales, but not at large scales).
- Using tornado warnings may be a viable supplement to tornado reports when verifying forecasts of low-level rotation in next-day CAM forecasts. SSPFs were more skillful when verified with a combined tornado report and warning verification dataset, confirming that 1- and 3-km CAMs can better capture the occurrence of intense low-level rotation events than the occurrence of a tornado.
- The 1-km UH25 SSPFs were more skillful than 3-km UH25 SSPFs at predicting all severe hazards, with statistically significant differences in FSS and ROC AUC. While the UH01 diagnostic was the best discriminator between tornadic and nontornadic events, UH01 had decreased skill relative to UH25 at both 3- and 1-km Δx when forecasting all severe hazards.

6. Discussion

While previous studies of convective storm forecasts (e.g., Johnson et al. 2013; Loken et al. 2017) revealed little sensitivity of forecast skill to Δx , the current work demonstrates that finer Δx may be beneficial for forecast applications such as identifying tornadic convection. We suspect that the present focus on phenomena such as

low-level mesocyclones, which have smaller spatial scales than midlevel mesocyclones, leads to a larger skill gap between the 3- and 1-km forecasts. This is partly supported by the large scaling factor for UH01 compared to UH25, associated with the smaller-scale of low-level mesocyclonic rotation (Table 3). Thus, the biggest benefit of 1-km CAMs may occur for applications where forecasts of small-scale convective processes are insufficiently resolved in 3–4-km CAMs. While this added discrimination is promising, simple postprocessing methods, such as combining STP with UH01, may reduce most of the skill gap between 3- and 1-km next-day forecasts of tornadoes. It remains to be seen whether CAM forecasts with $\Delta x < 1$ km further improve the ability to discriminate tornadic convection to the point where postprocessed 3-km CAM forecasts cannot compete.

The largest differences between the 3- and 1-km tornado SSPFs occurred on scales between 150 and 300 km. We hypothesize that the peak in 3- and 1-km skill differences on the mesoscale exists since the minimum scales at which CAMs can accurately predict tornadic environments also exists on the mesoscale. Given that storm-scale processes that lead to tornadogenesis are not predictable within the forecast lead time range considered here, producing accurate next-day tornado forecasts with present-day CAMs requires well-predicted larger-scale tornadic environments, in addition to accuracy in storm locations. Assuming 3- and 1-km forecasts predict the mesoscale environment with similar skill and have similar skill at predicting storm locations,⁴ the 1-km forecasts benefit by being better able to generate low-level mesocyclones and mesovortices in tornadic environments, with magnitudes distinct from those in nontornadic environments. In other words, while the generation of the low-level rotation in severe convection is not intrinsically predictable at next-day lead times, the prediction of intense low-level rotation benefits from the enhanced predictability of the larger-scale environment, as well as the capability to generate intense vertical vorticity on the 1-km grid. Evidence of this exists when filtering by STP, which improved the 3-km SSPFs more so than the 1-km SSPFs, revealing deficiencies in the ability of 3-km forecasts to discriminate between events and non-events that were partially remedied by supplementing forecasts with accurate mesoscale environmental information.

⁴ Subjective comparisons of many of the 3- and 1-km forecasts reveal similar forecasts of storm placement, with timing errors eliminated by producing 24-hour SSPFs.

Even though the 1-km tornado SSPFs were more skillful than 3-km tornado SSPFs, the tornado SSPFs examined here remain insufficiently skillful to provide detailed guidance on scales below the mesoscale (i.e., FSSs are <0.5 for scales $< \sim 100$ km). As discussed in Roberts and Lean (2008), smoothing finescale NWP output to improve forecast reliability may be detrimental to the value of the forecast products, reducing the sharpness of the probabilities generated and ultimately the practical benefit of added resolution. It is not obvious whether forecasters would find value in smoothed tornado SSPFs, although smoothed all-severe hazard guidance on length scales of 120–160 km has been tested and deemed useful in operational forecasting environments (e.g., Clark et al. 2012). Incorporating initial condition and model uncertainty as part of a CAM ensemble will improve the SSPFs, allowing for information to be presented on smaller scales (e.g., Sobash et al. 2016).

Finally, the most skillful diagnostics for extracting CAM-based severe hazard guidance appear to be UH25 and UH01, the former being used as a surrogate for all severe weather hazards, while the latter being used as a surrogate for tornadoes. Given the growing volume of diagnostics emanating from CAM ensembles, it is recommended that UH25 and UH01 be given priority over others (e.g., UH03, RVORT1, and RVORT0.5) when considering which diagnostics forecasters should interrogate or should be included in CAM output. While the UH25 and UH01 thresholds used here may be useful as a starting point when building CAM tornado guidance, the thresholds are sensitive to not only Δx , but also vertical grid spacing, physics parameterizations, and smoothing choices. Future work should consider these factors when optimizing CAM tornado guidance skill using surrogates, and should consider the usage of machine learning techniques to blend multiple surrogates and environmental information into a skillful, reliable, postprocessing system for environmental hazards such as tornadoes (e.g., Gagne et al. 2017).

Acknowledgments. This work was partially supported by NOAA OAR Grant NA17OAR4590182 and the NCAR Short-term Explicit Prediction (STEP) program. We thank the Iowa Environmental Mesonet NWS warning archive, maintained by Daryl Herzmann, for providing tornado warning shapefiles. We would also like to acknowledge high-performance computing support from Cheyenne (Computational and Information Systems Laboratory 2017) provided by NCAR's Computational and Information Systems Laboratory. The National Center for Atmospheric Research is sponsored by the National Science Foundation.

REFERENCES

- Adams-Selin, R. D., 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 dependence 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](https://doi.org/10.1175/1520-0493(2002)130<2671:TSONSC>2.0.CO;2).
- Baldwin, M. E., and J. S. Kain, 2006: Sensitivity of several performance measures to displacement error, bias, and event frequency. *Wea. Forecasting*, **21**, 636–648, <https://doi.org/10.1175/WAF933.1>.
- 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>.
- 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](https://doi.org/10.1175/1520-0493(2003)131<2394:RRFTSO>2.0.CO;2).
- Chen, F., and J. Dudhia, 2001: Coupling an advanced land surface-hydrology model with the Penn State–NCAR MM5 modeling system: Part I: Model implementation and sensitivity. *Mon. Wea. Rev.*, **129**, 569–585, [https://doi.org/10.1175/1520-0493\(2001\)129<0569:CAALSH>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2).
- Clark, A. J., W. A. Gallus Jr., M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Wea. Forecasting*, **24**, 1121–1140, <https://doi.org/10.1175/2009WAF2222222.1>.
- , 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>.
- , J. Gao, P. Marsh, T. Smith, J. Kain, J. Correia, M. Xue, and F. Kong, 2013: Tornado pathlength forecasts from 2010 to 2011 using ensemble updraft helicity. *Wea. Forecasting*, **28**, 387–407, <https://doi.org/10.1175/WAF-D-12-00038.1>.
- Computational and Information Systems Laboratory, 2017: Cheyenne: HPE/SGI ICE XA System (Climate Simulation Laboratory). National Center for Atmospheric Research, Boulder, CO, doi:10.5065/D6RX99HX.
- Davis, J. M., and M. D. Parker, 2014: Radar climatology of tornadic and nontornadic vortices in high-shear, low-CAPE environments in the mid-Atlantic and southeastern United States. *Wea. Forecasting*, **29**, 828–853, <https://doi.org/10.1175/WAF-D-13-00127.1>.
- Dean, A. R., and R. S. Schneider, 2008: Forecast challenges at the NWS Storm Prediction Center relating to the frequency of favorable severe storm environments. Preprints, *24th Conf. on Severe Local Storms*, Savannah, GA, Amer.

- Meteor. Soc., 9A.2, https://ams.confex.com/ams/24SLS/techprogram/paper_141743.htm.
- Done, J., C. A. Davis, and M. L. 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>.
- Duda, J. D., and W. A. Gallus Jr., 2010: Spring and summer midwestern severe weather reports in supercells compared to other morphologies. *Wea. Forecasting*, **25**, 190–206, <https://doi.org/10.1175/2009WAF2222338.1>.
- Gagne, D. J., II, A. McGovern, J. B. Basara, and R. A. Brown, 2017: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Wea. Forecasting*, **32**, 1819–1840, <https://doi.org/10.1175/WAF-D-17-0010.1>.
- Gallo, B. T., A. J. Clark, and S. R. Dembek, 2016: Forecasting tornadoes using convection-permitting ensembles. *Wea. Forecasting*, **31**, 273–295, <https://doi.org/10.1175/WAF-D-15-0134.1>.
- , —, B. T. Smith, R. L. Thompson, I. Jirak, and S. R. Dembek, 2018: Blended probabilistic tornado forecasts: Combining climatological frequencies with NSSL-WRF ensemble forecasts. *Wea. Forecasting*, **33**, 443–460, <https://doi.org/10.1175/WAF-D-17-0132.1>.
- , —, —, —, —, and —, 2019: Incorporating UH occurrence time to ensemble-derived tornado probabilities. *Wea. Forecasting*, **34**, 151–164, <https://doi.org/10.1175/WAF-D-18-0108.1>.
- Gallus, W. A., N. A. Snook, and E. V. Johnson, 2008: Spring and summer severe weather reports over the Midwest as a function of convective mode: A preliminary study. *Wea. Forecasting*, **23**, 101–113, <https://doi.org/10.1175/2007WAF2006120.1>.
- Guyer, J. L., and A. R. Dean, 2010: Tornadoes within weak CAPE environments across the continental United States. Preprints, *25th Conf. on Severe Local Storms*, Denver, CO, Amer. Meteor. Soc., 1.5, https://ams.confex.com/ams/25SLS/techprogram/paper_175725.htm.
- , and I. L. Jirak, 2014: The utility of convection-allowing ensemble forecasts of cool season severe weather events from the SPC perspective. Preprints, *27th Conf. on Severe Local Storms*, Madison, WI, Amer. Meteor. Soc., 37, <https://ams.confex.com/ams/27SLS/webprogram/Paper254640.html>.
- Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, [https://doi.org/10.1175/1520-0434\(1999\)014<0155:HTFENP>2.0.CO;2](https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2).
- Hepper, R. M., I. L. Jirak, and J. M. Milne, 2016: Assessing the skill of convection-allowing ensemble forecasts of severe MCS winds from the SSEO. Preprints, *28th Conf. on Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., 16B.2, <https://ams.confex.com/ams/28SLS/webprogram/Paper300134.html>.
- Hitchens, N. M., H. E. Brooks, and M. P. Kay, 2013: Objective limits on forecasting skill of rare events. *Wea. Forecasting*, **28**, 525–534, <https://doi.org/10.1175/WAF-D-12-00113.1>.
- Iacono, M. J., J. S. Delamere, E. J. Mlawer, M. W. Shephard, S. A. Clough, and W. D. Collins, 2008: Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *J. Geophys. Res.*, **113**, D13103, <https://doi.org/10.1029/2008JD009944>.
- Janjić, Z. I., 1994: The step-mountain eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev.*, **122**, 927–945, [https://doi.org/10.1175/1520-0493\(1994\)122<0927:TSMECM>2.0.CO;2](https://doi.org/10.1175/1520-0493(1994)122<0927:TSMECM>2.0.CO;2).
- , 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>.
- Jirak, I. L., C. J. Melick, and S. J. Weiss, 2014: Combining probabilistic ensemble information from the environment with simulated storm attributes to generate calibrated probabilities of severe weather hazards. Preprints, *27th Conf. on Severe Local Storms*, Madison, WI, Amer. Meteor. Soc., 2.5, <https://ams.confex.com/ams/27SLS/webprogram/Paper254649.html>.
- Johnson, A., X. Wang, F. Kong, and M. Xue, 2013: Object-based evaluation of the impact of horizontal grid spacing on convection-allowing forecasts. *Mon. Wea. Rev.*, **141**, 3413–3425, <https://doi.org/10.1175/MWR-D-13-00027.1>.
- Kain, J. S., S. J. Weiss, J. J. Levit, M. E. Baldwin, and D. R. Bright, 2006: Examination of convection-allowing configurations of the WRF model for the prediction of severe convective weather: The SPC/NSSL Spring Program 2004. *Wea. Forecasting*, **21**, 167–181, <https://doi.org/10.1175/WAF906.1>.
- , 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>.
- , S. R. Dembek, S. J. Weiss, J. L. Case, J. J. Levit, and R. A. Sobash, 2010: Extracting unique information from high-resolution forecast models: Monitoring selected fields and phenomena every time step. *Wea. Forecasting*, **25**, 1536–1542, <https://doi.org/10.1175/2010WAF2222430.1>.
- Knievel, J. C., G. H. Bryan, and J. P. Hacker, 2007: Explicit numerical diffusion in the WRF Model. *Mon. Wea. Rev.*, **135**, 3808–3824, <https://doi.org/10.1175/2007MWR2100.1>.
- 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>.
- Marzban, C., 2004: The ROC curve and the area under it as performance measures. *Wea. Forecasting*, **19**, 1106–1104, <https://doi.org/10.1175/825.1>.
- Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.
- Mass, C. F., D. Owens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83**, 407–430, [https://doi.org/10.1175/1520-0477\(2002\)083<0407:DIHRPM>2.3.CO;2](https://doi.org/10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2).
- 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>.
- Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res.*, **102**, 16 663–16 682, <https://doi.org/10.1029/97JD00237>.
- Naylor, J., M. S. Gilmore, R. L. Thompson, R. Edwards, and R. B. Wilhelmson, 2012: Comparison of objective supercell identification techniques using an idealized cloud model. *Mon. Wea. Rev.*, **140**, 2090–2102, <https://doi.org/10.1175/MWR-D-11-00209.1>.
- 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>.
- 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>.
- Roberts, B., I. L. Jirak, A. J. Clark, S. J. Weiss, and J. S. Kain, 2019: Postprocessing and visualization techniques for

- convection-allowing ensembles. *Bull. Amer. Meteor. Soc.*, **100**, 1245–1258, <https://doi.org/10.1175/BAMS-D-18-0041.1>.
- 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>.
- Schwartz, C. S., and Coauthors, 2009: Next-day convection-allowing WRF Model guidance: A second look at 2-km versus 4-km grid spacing. *Mon. Wea. Rev.*, **137**, 3351–3372, <https://doi.org/10.1175/2009MWR2924.1>.
- , 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>.
- 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>.
- 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>.
- Sobash, R. A., and J. S. Kain, 2017: Seasonal variations in severe weather forecast skill in an experimental convection-allowing model. *Wea. Forecasting*, **32**, 1885–1902, <https://doi.org/10.1175/WAF-D-17-0043.1>.
- , —, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011: Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme phenomena in convection-allowing model forecasts. *Wea. Forecasting*, **26**, 714–728, <https://doi.org/10.1175/WAF-D-10-05046.1>.
- , G. S. Romine, C. S. Schwartz, D. J. Gagne II, and M. L. Weisman, 2016: Explicit forecasts of low-level rotation from convection-allowing models for next-day tornado prediction. *Wea. Forecasting*, **31**, 1591–1614, <https://doi.org/10.1175/WAF-D-16-0073.1>.
- Tegen, I., P. Hollrig, M. Chin, I. Fung, D. Jacob, and J. Penner, 1997: Contribution of different aerosol species to the global aerosol extinction optical thickness: Estimates from model results. *J. Geophys. Res.*, **102**, 23 895–23 915, <https://doi.org/10.1029/97JD01864>.
- Theis, S. E., A. Hense, and U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: A pragmatic approach. *Meteor. Appl.*, **12**, 257–268, <https://doi.org/10.1017/S1350482705001763>.
- Thompson, G., P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, <https://doi.org/10.1175/2008MWR2387.1>.
- 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](https://doi.org/10.1175/1520-0434(2003)018<1243:CPSWSE>2.0.CO;2).
- , B. T. Smith, J. S. Grams, A. R. Dean, and C. Broyles, 2012: Convective modes for significant severe thunderstorms in the contiguous United States. Part II: Supercell and QLCS tornado environments. *Wea. Forecasting*, **27**, 1136–1154, <https://doi.org/10.1175/WAF-D-11-00116.1>.
- VandenBerg, M. A., M. C. Coniglio, and A. J. Clark, 2014: Comparison of next-day convection-allowing forecasts of storm motion on 1- and 4-km grids. *Wea. Forecasting*, **29**, 878–893, <https://doi.org/10.1175/WAF-D-14-00011.1>.
- 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](https://doi.org/10.1175/1520-0493(1997)125<0527:TRDOEM>2.0.CO;2).
- , C. Davis, W. Wang, K. W. Manning, and J. B. Klemp, 2008: Experiences with 0–36-h explicit convective forecasts with the WRF-ARW Model. *Wea. Forecasting*, **23**, 407–437, <https://doi.org/10.1175/2007WAF2007005.1>.
- Weiss, S. J., and M. D. Vescio, 1998: Severe local storm climatology 1955–1996: Analysis of reporting trends and implications for NWS operations. Preprints, *18th Conf. on Severe Local Storms*, Minneapolis, MN, Amer. Meteor. Soc., 536–539.
- Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Academic Press, 676 pp.
- Wolff, J. K., M. Harrold, T. Fowler, J. H. Gotway, L. Nance, and B. G. Brown, 2014: Beyond the basics: Evaluating model-based precipitation forecasts using traditional, spatial, and object-based methods. *Wea. Forecasting*, **29**, 1451–1472, <https://doi.org/10.1175/WAF-D-13-00135.1>.