Revisiting Sensitivity to Horizontal Grid Spacing in Convection-Allowing Models over the Central and Eastern United States

CRAIG S. SCHWARTZ AND RYAN A. SOBASH

National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 16 April 2019, in final form 30 July 2019)

ABSTRACT

Hourly accumulated precipitation forecasts from deterministic convection-allowing numerical weather prediction models with 3- and 1-km horizontal grid spacing were evaluated over 497 forecasts between 2010 and 2017 over the central and eastern conterminous United States (CONUS). While precipitation biases varied geographically and seasonally, 1-km model climatologies of precipitation generally aligned better with those observed than 3-km climatologies. Additionally, during the cool season and spring, when large-scale forcing was strong and precipitation entities were large, 1-km forecasts were more skillful than 3-km forecasts, particularly over southern portions of the CONUS where instability was greatest. Conversely, during summertime, when synoptic-scale forcing was weak and precipitation entities were small, 3- and 1-km forecasts had similar skill. These collective results differ substantially from previous work finding 4-km forecasts had comparable springtime precipitation forecast skill as 1- or 2-km forecasts over the central–eastern CONUS. Additional analyses and experiments suggest the greater benefits of 1-km forecasts documented here could be related to higher-quality initial conditions than in prior studies. However, further research is needed to confirm this hypothesis.

1. Introduction

Convection-allowing numerical weather prediction (NWP) models[1] with horizontal grid spacings (Δx) of approximately 4 km or less have dramatically improved precipitation and severe weather forecasting and are routinely used in research and operations (e.g., Clark et al. 2016). As substantial computational resources are needed to produce convection-allowing forecasts over large domains, an important consideration regards resolution—how fine should Δx be, especially for operational purposes?

Thus, many studies have examined sensitivity to Δx at convection-allowing scales for precipitation and severe weather applications (e.g., Tables 1 and 2), and some consensus appears to have emerged. For instance, decreasing Δx to ~1 km seems necessary to capture intense precipitation when topographic gradients are steep (e.g., Colle and Mass 2000; Colle et al. 2005; Garvert et al. 2005; Buzzi et al. 2014; Schwartz 2014; Bartsotas et al. 2017), and continually reducing Δx clearly yields more realistic structures. There also appears to be agreement that coarser-resolution (but still convection-allowing) ensemble forecasts produce better guidance than deterministic forecasts with even higher resolution (e.g., Hagelin et al. 2017; Loken et al. 2017; Mittermaier and Csima 2017; Schwartz et al. 2017).

However, there remains substantial disagreement about the necessity of decreasing Δx below 3–4 km when strong orographic forcing does not primarily drive precipitation. For example, studies over Europe and Japan found forecasts with $\Delta x = 1$ or 2 km were better than those with $\Delta x = 4$ or 5 km (e.g., Lean et al. 2008; Roberts and Lean 2008; Ito et al. 2017). Accordingly, operational European and Japanese convection-allowing models have Δx between 1.1 and 2.8 km (e.g., Baldauf et al. 2011; Hirahara et al. 2011; Seity et al. 2011; Tang et al. 2013; Kühnlein et al. 2014; Brousseau et al. 2016; Hagelin et al. 2017; Raynaud and Bouttier 2017; Klasa et al. 2018; MeteoSwiss 2019), and research in Europe and Japan has now largely turned to assessing whether reducing Δx to <2 km warrants the cost, which has collectively

[1] Deep cumulus parameterization schemes are typically removed with $\Delta x \leq$ ~4 km to allow explicit representation of deep convection, hence the nomenclature "convection-allowing."

Corresponding author: Craig Schwartz, schwartz@ucar.edu

DOI: 10.1175/MWR-D-19-0115.1

TABLE 1. Real-data studies that systematically examined sensitivity to Δx over relatively flat terrain. Studies are grouped chronologically by region. While some studies additionally examined sensitivity to Δx for nonprecipitation variables (Brousseau et al. 2016; Hagelin et al. 2017; Ito et al. 2017; Raynaud and Bouttier 2017; Stein et al. 2019), as our focus is on precipitation, findings regarding other variables are not included in the table.

Study	Sample size	Δx (km)	Region	Precipitation-related variables	Forecast range (h)	Relevant conclusions
Roberts and Lean (2008)	40[a]	4, 1	Southern England	Precipitation	1–7 h	1 km better than 4 km
Lean et al. (2008)	63[a]	4, 1	Southern England	Precipitation	1–7 h	1 km better than 4 km
Hagelin et al. (2017)	228[b]	2.2, 1.5	United Kingdom	Precipitation	1–36 h	2.2- and 1.5-km probabilistic precipitation forecasts similar
Brousseau et al. (2016)	48	2.5, 1.3	France	Precipitation, reflectivity	1–30 h	1.3 km better than 2.5 km
Raynaud and Bouttier (2017)	91	2.5, 1.3	France	Precipitation	1–45 h	2.5- and 1.3-km probabilistic precipitation forecasts similar
Kain et al. (2008)	35	4, 2	Central–eastern CONUS	Precipitation, reflectivity, updraft helicity	Primarily 12–30 h	2- and 4-km forecasts similar and provided comparable value
Schwartz et al. (2009)	35	4, 2	Central–eastern CONUS	Precipitation, reflectivity	Primarily 21–33 h	2- and 4-km forecasts similar and provided comparable value
Clark et al. (2012)	24	4, 1	Central–eastern CONUS	Reflectivity	20–28 h	1- and 4-km forecasts subjectively rated similarly
Johnson et al. (2013)	91[c]	4, 1	Central–eastern CONUS	Precipitation	1–30 h	4- and 1-km forecasts similar, except on scales unresolvable by 4 km, where 1 km was better
VandenBerg et al. (2014)	22	4, 1	Central CONUS	Storm motion, where storms were identified with reflectivity	6–30 h	1 km better than 4 km
Schwartz et al. (2017)	32	3, 1	Central–eastern CONUS	Precipitation	Primarily 1–12 h, 18–36 h	1 km generally better than 3 km, especially for heavier rainfall rates
Loken et al. (2017)	63	4, 1	Central–eastern CONUS	Updraft helicity	12–36 h	4 and 1 km similar
Ito et al. (2017)	100	5, 2, 1, 0.5	Japan	Precipitation	3–16 h	2 km better than 5 km with little improvement going below 2 km
Stein et al. (2019)	30	4.4, 1.5	South Africa	Precipitation	12–36 h	4.4 and 1.5 km similar

[a] Roberts and Lean (2008) and Lean et al. (2008) selected 10 and 16 cases to study, respectively, and for each case, four forecasts were initialized 3 h apart [only three forecasts were run for one case in Lean et al. (2008)]. Thus, their effective sample sizes may have been smaller due to possible meaningful correlations between rapidly updated forecasts.

[b] Hagelin et al. (2017) produced forecasts 4 times per day, initialized 6 h apart, for 57 days. Thus, their effective sample size may have been smaller due to possible meaningful correlations between adjacent forecasts.

[c] Johnson et al. (2013, hereafter J13) evaluated forecasts over 22, 36, and 33 days in 2009, 2010, and 2011, respectively, and model configurations changed each year.

TABLE 2. Idealized studies performed within Great Plains–like environments that examined sensitivity to Δx between 4 and 1 km. All simulations were in deterministic frameworks. See S17 for further discussion of these studies.

Study	Δx (km)	Phenomena	Relevant conclusions
Weisman et al. (1997)	12, 8, 4, 2, 1	Squall line	1 km best, but 4 km good enough to capture most of the structure and evolution
Bryan and Morrison (2012)	4, 1, 0.25	Squall line	4 and 1 km broadly similar in terms of rainfall, but 0.25 km best
Potvin and Flora (2015)	4, 3, 2, 1	Supercell	4 km clearly worst, 1 km clearly best, but 3 km good enough for operations
Verrelle et al. (2015)	4, 2, 1, 0.5	Supercell–multicell	1 km better than 2 or 4 km, with little benefit of going from 1 km to 500 m; recommended 1 km for operations

yielded mixed conclusions (e.g., Brousseau et al. 2016; Barthlott et al. 2017; Hagelin et al. 2017; Ito et al. 2017; Raynaud and Bouttier 2017).

Conversely, operational convection-allowing NWP models over the conterminous United States (CONUS) currently have Δx around 3 km, in part because of conflicting conclusions regarding the necessity to further increase resolution over the relatively flat terrain of the central and eastern CONUS. For instance, while several studies found 1- or 2-km forecasts provided comparable next-day precipitation and severe weather guidance as 4-km forecasts east of the Rocky Mountains (e.g., Kain et al. 2008; Schwartz et al. 2009; Clark et al. 2012; Johnson et al. 2013; Loken et al. 2017), VandenBerg et al. (2014) suggested convective storm motion could be improved by decreasing Δx from 4 to 1 km. Moreover, within an ensemble context, Schwartz et al. (2017, hereafter S17) found 1-km 18–36-h forecasts of 1-h accumulated precipitation $\geq$5.0 mm were statistically significantly better than 3-km forecasts, which clashes with Kain et al. (2008, hereafter K08) and Schwartz et al. (2009, hereafter S09), who found deterministic 2- and 4-km precipitation forecasts were remarkably similar at those precipitation rates and forecast ranges.

These disagreements across systematic studies (e.g., Table 1) have also been manifested by real-data case studies over the CONUS. For example, Xue et al. (2013) described a case where 4-km forecasts failed to initiate a supercell but a 1-km forecast did, while Schumacher (2015) detailed the opposite: a supercell and subsequent mesoscale convective system (MCS) were accurately reproduced with $\Delta x = 4$ km but simulations with finer Δx failed to initiate the supercell. Moreover, Clark et al. (2013) noted a case where $\Delta x = 1$ km was needed to capture strong supercellular updraft rotation when $\Delta x = 4$ km was insufficient, even though strong rotating updrafts were produced with $\Delta x = 4$ km for other cases. Likewise, idealized studies in Great Plains–like environments have yielded varied conclusions and recommendations regarding Δx for operational applications (Table 2).

Thus, approximately ten years after the first systematic evaluations of sensitivity to Δx at convection-allowing scales, questions still remain about resolution requirements, particularly over relatively flat terrain. While the differing conclusions regarding necessary Δx between the European and American studies could be due to disparities related to geography, climatology, and NWP models, differences amongst the CONUS studies (e.g., Table 1) are more difficult to reconcile, as they all used a common NWP model dynamic core and examined forecasts over similar regions.

Might differences across the various CONUS studies be due to small sample sizes? Because high-resolution simulations can be computationally expensive, previous work systematically assessing sensitivity to Δx often employed modest sample sizes; for CONUS-based investigations, at most 91 cases were examined. Although many studies listed in Table 1 found statistically significant results, larger sample sizes are arguably needed to attempt to reconcile different conclusions regarding whether decreasing Δx below 3–4 km over the central and eastern CONUS yields forecast improvements.

Therefore, this paper examines sensitivity to Δx over the CONUS for next-day precipitation forecasts with an unprecedented sample size. Specifically, 497 corresponding 3- and 1-km forecasts were evaluated with hopes of more definitively determining whether Δx should be decreased toward 1 km in future operational systems covering the CONUS. While Sobash et al. (2019) used an identical dataset and found 1-km forecasts provided better tornado guidance than 3-km forecasts, this paper focuses on precipitation and goes further by assessing seasonal and geographic variations of sensitivity to Δx and relating differences between 3- and 1-km forecast skill to broader environmental characteristics.

2. Model configurations and case selection

Sobash et al. (2019) described case selection and model configurations. For completeness, descriptions are briefly

TABLE 3. Cool season severe weather events for which 3- and 1-km forecasts were produced. Forecasts were initialized at 0000 UTC each day. The cool season (defined as 15 Oct–14 Mar) spans months in two years, with Oct–Dec in the earlier year and Jan–Mar in the later year (e.g., the 2010/11 cool season spanned 15 Oct 2010–14 Mar 2011).

		Month					
		Oct	Nov	Dec	Jan	Feb	Mar
Cool season	2010/11	24, 25, 26, 27	16, 22, 29, 30	31	25	1, 24, 27, 28	5, 8, 9, 10
	2011/12		7, 8, 14, 16	22	17, 22, 25	18, 24, 28, 29	2
	2012/13	17		17, 19, 20, 25	29, 30	10, 18	
	2013/14	31	17	21	11	20, 21	
	2014/15		16, 23	23	3		
	2015/16	30, 31	11, 16, 17	23, 27	21	15, 16, 23, 24	8
	2016/17		28, 29, 30	17	2, 20, 21, 22	7, 19, 25, 28	1, 6, 9

repeated here, and at times the text parallels that of Sobash et al. (2019).

a. Case selection

A total of 497 36-h forecasts were produced with both 3- and 1-km Δx on days with severe thunderstorms east of the Rocky Mountains. Cases were determined by consulting the Storm Prediction Center's (SPC) archive of severe thunderstorm events; events were defined based on several criteria including number of severe weather reports within a particular area, monetary damage, and fatalities (http://www.spc.noaa.gov/exper/archive/events/introduction.html). This case selection strategy meant forecasts were not produced for continuous time periods, differing from many previous studies of sensitivity to Δx over the CONUS (Table 1). Thus, while our results are somewhat conditioned upon observations of extreme weather, a large evaluation domain meant there were many null events.

Forecasts were produced for all events in the SPC archive between 15 March and 15 July each year between 2011 and 2016 (inclusive), corresponding to peak severe weather season over the CONUS east of the Rockies. We classified forecasts between 15 March and 14 June as "spring" and between 15 June to 15 July as "summer." Forecasts were also produced for the 2010/11, 2011/12, . . . , 2015/16, and 2016/17 "cool seasons," defined as 15 October–14 March. During the cool season, severe weather occurs less frequently and the SPC's criteria for including events in its archive are relaxed; some cataloged events had very few (i.e., <10) storm reports over the CONUS that we did not simulate. Instead, we only produced cool season forecasts for selected events in the SPC archive, focusing on those with relatively large numbers of storm reports (Table 3). In total, there were 279, 140, and 78 springtime, summertime, and cool season forecasts, respectively.

b. Model configurations and initialization

Independent forecasts with 3- and 1-km Δx were produced by version 3.6.1 of the Advanced Research Weather Research and Forecasting (WRF) Model (Skamarock et al. 2008; Powers et al. 2017). All forecasts ran over an identical computational domain (Fig. 1) spanning the entire CONUS; 1-km forecasts had exactly 9 times the number of grid points as 3-km forecasts. All forecasts used common physical parameterizations (Table 4), employed positive-definite moisture and scalar advection (Skamarock and Weisman 2009), and had 40 vertical levels with a 50-hPa top. Identical vertical levels were used in the 3- and 1-km forecasts to fully isolate sensitivity to Δx. We compared 3- and 1-km forecasts rather than say, 2- and 4-km forecasts, because most operational convection-allowing models over the CONUS currently have $\Delta x = 3$ km (e.g., Benjamin et al. 2016; Rogers et al. 2017).

The 1-km forecasts had a 4-s time step, and it was unclear what time step to assign the 3-km forecasts to permit the best comparison with the 1-km forecasts. Thus, as described by Sobash et al. (2019), two 3-km forecast sets were produced. One used a 12-s time step, 3 times larger than the 1-km time step to maintain the 3:1 ratio of Δx between the 3- and 1-km forecasts, while the second 3-km forecast set used a 4-s time step, like the 1-km forecasts. Altering the 3-km time step made little difference regarding precipitation forecast skill and climatologies, so the remainder of this paper solely examines the 12-s time step 3-km forecasts.

Initial conditions (ICs) for all 36-h forecasts were produced by interpolating 0000 UTC 0.5° Global Forecast System (GFS) analyses onto the 3- and 1-km grids (Fig. 1). GFS forecasts at 3-h intervals provided lateral boundary conditions (LBCs). Although 0.25° GFS output became available in 2015, 0.5° GFS fields were used across all 497 forecasts for consistency. Because WRF model preprocessing discards GFS hydrometeor analyses, all forecasts began with

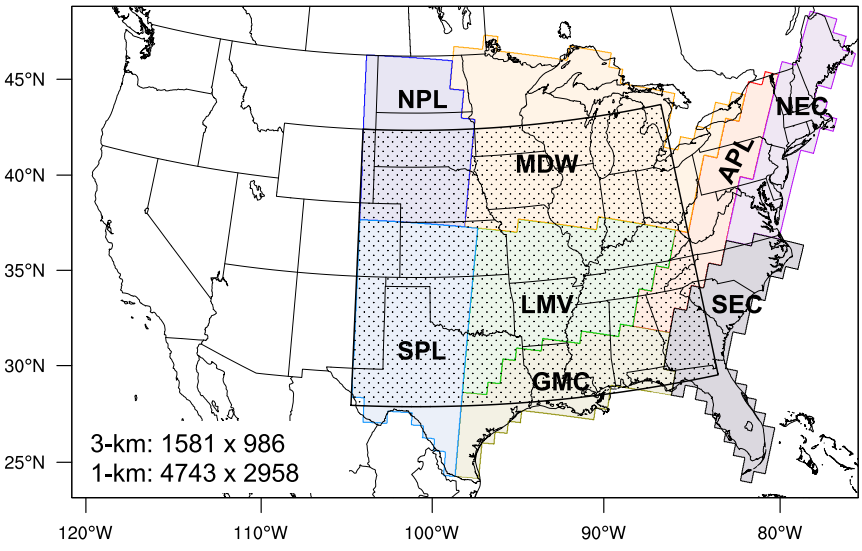

FIG. 1. Computational domain for the 3- and 1-km forecasts. Shadings and annotations denote eight regions used for model evaluation, while the stippled area represents a ninth region used to verify auxiliary forecasts revisiting previous studies (section 6). Forecasts were also verified over several metaregions composed of the union of various individual regions (Table 5). NPL: northern plains, SPL: southern plains, MDW: Midwest, LMV: lower Mississippi Valley, GMC: Gulf of Mexico coast, APL: Appalachians, NEC: Northeast coast, and SEC: Southeast coast.

no hydrometeors (i.e., zero values), meaning a substantial spinup period was expected.

3. Verification methods

Hourly accumulated precipitation from the 3- and 1-km forecasts was compared to Stage IV (ST4) observations produced at NCEP on a ~4.763-km grid (Lin and Mitchell 2005). To facilitate this comparison, 3- and 1-km precipitation fields were interpolated to the ST4 grid using a precipitation-conserving budget interpolation algorithm (e.g., Accadia et al. 2003), and these interpolated fields were used to produce all verification statistics. S17 and VandenBerg et al. (2014) both showed how interpolating 1-km forecasts to the ST4 grid does not meaningfully impact 1-km precipitation structures.

Objective verification statistics were produced separately for eight geographic regions east of 105°W (Fig. 1), where ST4 data were most robust (e.g., Nelson et al. 2016). The eight regions were nearly identical to NCEP's Weather Prediction Center's verification regions (e.g., Blake et al. 2018) that have "approximate uniformity of climatology and terrain" (http://www.wpc.ncep.noaa.gov/rgnscr/verify.html), with the only differences involving the cutoff at 105°W and a slight southwestward expansion of the southern plains region ("SPL" on Fig. 1) to encompass Texas's Big Bend. Statistics were also produced for various "metaregions" (Table 5), including a metaregion spanning approximately two-thirds of the CONUS composed of the union of all eight regions in Fig. 1 (the $\text{CONUS}_{2/3}$ metaregion). We primarily focused on next-day (18–36 h) forecasts to avoid the spinup period.

The popular fractions skill score (FSS; Roberts and Lean 2008) was used to evaluate precipitation forecast skill. To compute FSSs, events were defined as whether

TABLE 4. Physical parameterizations for all WRF Model forecasts. No cumulus parameterization was used.

Parameterization	WRF Model option	References
Microphysics	Thompson	Thompson et al. (2008)
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)
Planetary boundary layer	Mellor–Yamada–Janjić (MYJ)	Mellor and Yamada (1982); Janjić (1994, 2002)
Land surface model	Noah	Chen and Dudhia (2001)

TABLE 5. Definitions of metaregions used in forecast evaluation.

Metaregion	Geographic regions from Fig. 1
East	NEC, SEC, APL
Mississippi River basin (MRB)	MDW, LMV, GMC
Plains	NPL, SPL
$CONUS_{2/3}$	NEC, SEC, APL, MDW, LMV, GMC, NPL, SPL

1-h accumulated precipitation met or exceeded a certain threshold, and a neighborhood length scale (r) was selected that determined a local neighborhood about a particular grid point. Then, two precipitation fields on the common ST4 grid, A and B, were transformed into fractional fields at the ith of N_v verification points by dividing the number of events occurring within the neighborhood of the ith point by the total number of points within the neighborhood (e.g., Theis et al. 2005). Typically, A represents a forecast field and B observations, although both A and B can be forecasts (e.g., Dey et al. 2014). Letting A_i^F and B_i^F denote fractional fields at the ith point derived from A and B, respectively, the FSS for a single forecast at a particular time is

$$\mathrm{FSS}(A,B) = 1 - \frac{\sum_{i=1}^{N_v} (A_i^F - B_i^F)^2}{\sum_{i=1}^{N_v} (A_i^F)^2 + \sum_{i=1}^{N_v} (B_i^F)^2}. \quad (1)$$

Letting o, f_3, and f_1 respectively denote ST4 observations, 3-km forecasts, and 1-km forecasts, we define FSS_3, FSS_1, and $FSS_{1–3}$ as FSSs obtained from comparing 3-km forecasts to ST4 observations, 1-km forecasts to ST4 observations, and 3- and 1-km forecasts to each other, respectively:

$$\mathrm{FSS}_3 = \mathrm{FSS}(f_3, o), \quad (2)$$

$$\mathrm{FSS}_1 = \mathrm{FSS}(f_1, o), \quad (3)$$

$$\mathrm{FSS}_{1–3} = \mathrm{FSS}(f_1, f_3). \quad (4)$$

The FSS ranges from 0 to 1. When A represents a forecast and B observations (e.g., FSS_1, FSS_3), perfect forecasts have FSS = 1 and FSS = 0 indicates no skill, while when A and B are both forecasts (e.g., $FSS_{1–3}$), higher FSSs indicate the two fields are more similar. Neighborhood length scales of r = 5, 25, 50, 75, 100, 150, and 200 km were used and defined as radii of a circle.

We computed aggregate FSSs (e.g., Mittermaier 2019) for both individual forecast hours and periods spanning multiple forecast hours; aggregate FSSs over M forecasts and a p–h period were obtained by summing over $i = 1, \ldots, N_v \times p \times M$ grid points in Eq. (1). By aggregating quantities in the numerator and denominator of Eq. (1) before producing final FSSs, more weight was given to larger events and fractional "correct negatives" (points where $A_i^F = B_i^F = 0$) did not impact scores (Mittermaier 2019).

FSSs were calculated for both raw and bias-corrected precipitation fields, where bias correction was performed with a probability matching approach that forced the forecast precipitation distribution over a particular verification region (Fig. 1) to that observed, thus, eliminating bias (Ebert 2001; Clark et al. 2009, 2010a,b; S17; Pyle and Brill 2019). Even though the 3- and 1-km forecasts sometimes exhibited different bias characteristics (section 4), FSSs computed from bias-corrected precipitation fields yielded identical overall conclusions as FSSs based on raw fields, suggesting FSSs computed from raw fields primarily measured spatial errors, not biases. Thus, solely FSSs based on raw 3- and 1-km precipitation forecasts are presented.

A bootstrap resampling technique (e.g., Hamill 1999; Davis et al. 2010; Wolff et al. 2014) with 10 000 resamples was used to assess statistical significance based on differences between pairs of 3- and 1-km forecasts. The forecasts, which were initialized at least 24 h apart, were assumed to be uncorrelated, following Hamill (1999). Thus, when assessing statistical significance for a single forecast hour (i.e., $p = 1$), we assumed all resamples were also uncorrelated.

Conversely, when analyzing statistical significance of FSSs aggregated over several forecast hours (i.e., $p > 1$), we could no longer assume resamples were independent due to potential autocorrelations of errors within a single forecast. Therefore, to assess statistical significance of FSSs aggregated over more than one forecast hour, we used a moving circular block-bootstrapping approach (e.g., Politis and Romano 1992; Wilks 1997; Gilleland et al. 2018) with a block length of 4 h to preserve autocorrelations in the resampling.[2]

In both bootstrapping scenarios, the significance level for differences between the 3- and 1-km forecasts was defined as the percentile where the distribution of resampled differences crossed zero (e.g., Davis et al. 2010). Differences were deemed statistically significant if the significance level was 95% (5%) or higher (lower).

4. Precipitation climatologies and biases

a. Areal coverages

Areal coverages of 1-h accumulated precipitation exceeding various thresholds revealed interesting regional

[2] Using block lengths of 3 and 5 h did not alter conclusions regarding statistical significance.

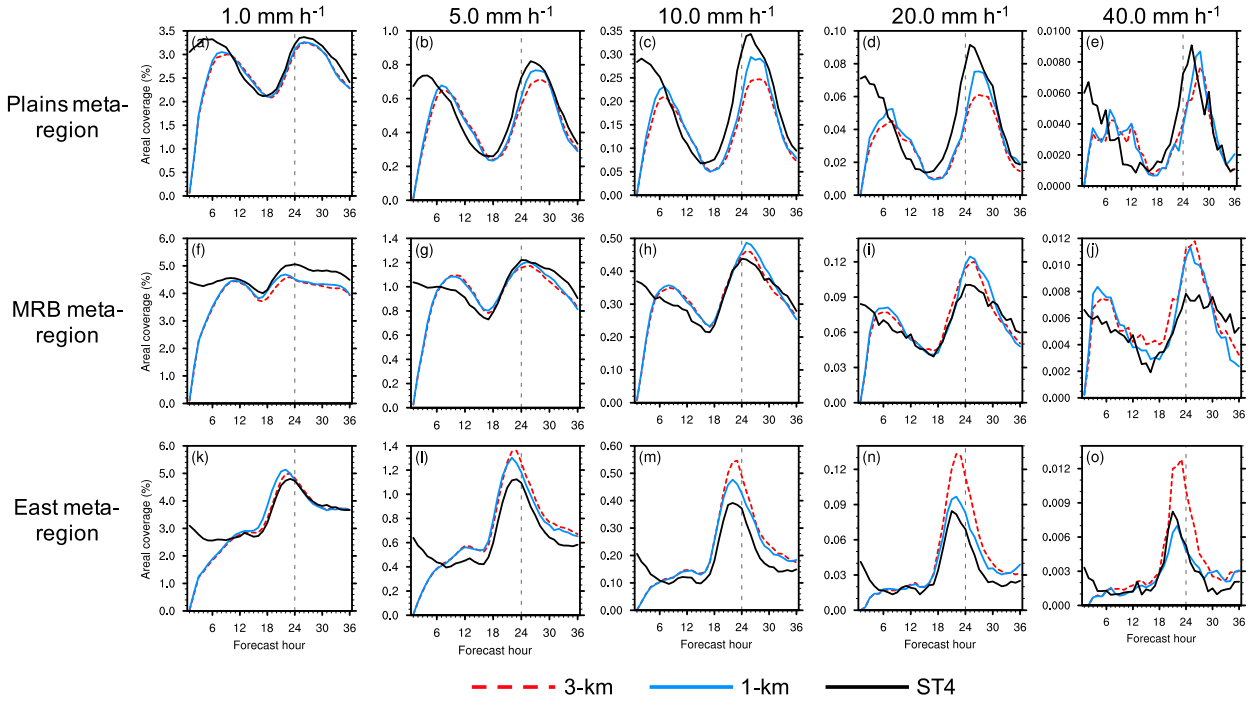

FIG. 2. Fractional areal coverage (%) of 1-h accumulated precipitation meeting or exceeding (a),(f),(k) 1.0, (b),(g),(l) 5.0, (c),(h),(m) 10.0, (d),(i),(n) 20.0, and (e),(j),(o) 40.0 mm h^{-1} over the (a)–(e) Plains, (f)–(j) MRB, and (k)–(o) East metaregions (Fig. 1; Table 5), aggregated over all 279 springtime (15 Mar–14 Jun) forecasts as a function of forecast hour. Values on the *x* axis represent ending hours of 1-h accumulation periods (e.g., an *x*-axis value of 24 is for 1-h accumulated precipitation between 23 and 24 h). Dashed vertical lines at hour 24 are for reference.

and seasonal differences between the 3- and 1-km forecasts. Because comparisons of forecast and observed areal coverages over a specific verification region could be impacted by small spatial errors near the region's boundaries, to lessen these potential impacts, areal coverages were computed over the Plains, MRB, and East metaregions (Fig. 1; Table 5) instead of over the eight individual regions.

1) SPRING AND SUMMER

In both spring and summer, the 3- and 1-km forecasts usually well represented the timing of the diurnal cycle for thresholds ≥5.0 mm h^{-1} (Figs. 2 and 3). However, during springtime for precipitation rates ≥5.0 mm h^{-1}, both the 3- and 1-km forecasts underpredicted coverages compared to observations over the Plains metaregion (Figs. 2b–e) and overpredicted coverages over the East metaregion (Figs. 2l–o); in both metaregions, 1-km coverages were closest to ST4 coverages and 3-km biases were clearly largest. Over the MRB metaregion, springtime 3- and 1-km biases were more similar (Figs. 2g–j), with overprediction for precipitation rates ≥10.0 mm h^{-1}. Summertime coverages for thresholds ≥5.0 mm h^{-1} broadly echoed springtime characteristics over the MRB and East metaregions (Figs. 3g–j,l–o), although differences between 3- and 1-km coverages were larger than in spring over the MRB metaregion. Over the Plains metaregion during summer, underprediction compared to spring was reduced (Figs. 3b–d), with overprediction for precipitation rates ≥40.0 mm h^{-1} (Fig. 3e).

At the 1.0 mm h^{-1} threshold, 3- and 1-km areal coverages were generally similar. Springtime coverages well-matched observations over the Plains and East metaregions (Figs. 2a,k), while coverages were too low over the MRB metaregion during spring (Fig. 2f) and all metaregions in summer (Figs. 3a,f,k). The biggest differences at the 1.0 mm h^{-1} threshold occurred between 18 and 24 h over the MRB and East metaregions, where 1-km coverages were higher and earlier-peaking than 3-km coverages during both spring (Figs. 2f,k) and summer (Figs. 3f,k). S17 documented similar quantitative springtime findings and attributed higher 1-km coverages to spurious light rainfall streaks produced by shallow cumulus clouds. Interestingly, the Plains metaregion, which featured higher lifting condensation levels than the other metaregions, lacked these streaks, as confirmed by visual inspection of individual forecasts and manifested by areal coverages (Figs. 2a and 3a), suggesting they only occurred in areas with relatively shallow, moist boundary layers.

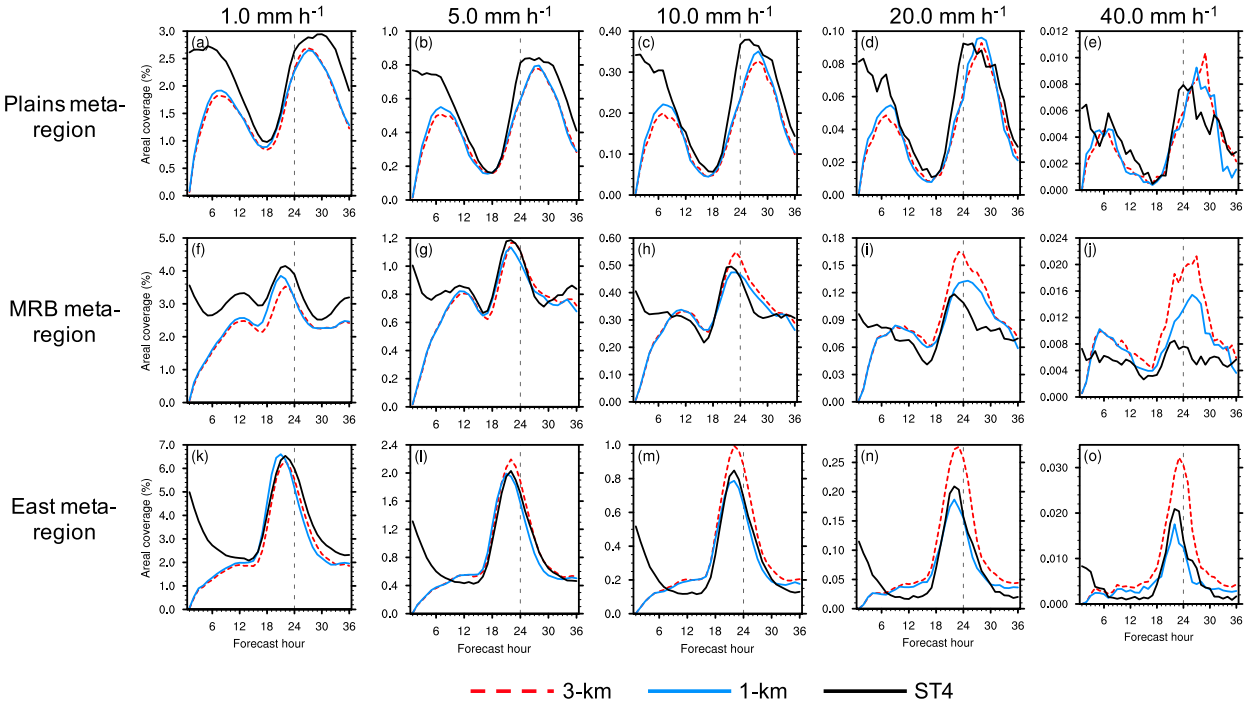

FIG. 3. As in Fig. 2, but for aggregate areal coverages over the 140 summertime (15 Jun–15 Jul) forecasts.

2) COOL SEASON

Cool season areal coverages differed substantially from warm season coverages, with less clearly defined peaks and valleys (Fig. 4), as decreased solar insolation during the cool season limited diurnally-driven precipitation. As forecast length increased, observed coverages generally decreased, remained steady, and increased over the Plains, MRB, and East metaregions, respectively. These behaviors reflect our focus on severe weather events (section 2a); many cool season events produced severe weather over the Plains or the MRB metaregion early in the forecasts before precipitation moved northeastward. Nonetheless, in all metaregions, 3- and 1-km trends well matched those observed.

The 3- and 1-km coverages were similar at the 1.0 and 5.0 mm h^{-1} thresholds (Figs. 4a,b,f,g,k,l), with larger differences at the 10.0 and 20.0 mm h^{-1} thresholds, where 1-km coverages were typically higher than 3-km coverages across all metaregions (Figs. 4c,d,h,i,m,n). Over the Plains and MRB metaregions, 1-km coverages at the 10.0 and 20.0 mm h^{-1} thresholds were usually closest to those observed despite low biases between 18 and 36 h (Figs. 4c,d,h,i), whereas 3-km coverages were closest to ST4 coverages over the East metaregion despite high biases (Figs. 4m,n). For the 40.0 mm h^{-1} threshold, both forecasts underpredicted over the MRB metaregion (Fig. 4j), while coverages over the Plains and East metaregions were noisy due to small sample sizes, but nonetheless broadly consistent with observations (Figs. 4e,o).

b. *Precipitation entity size*

To further understand geographic variations of areal coverage behaviors during spring and summer, using identical methods as K08, we examined sizes of precipitation "entities", where entity size was defined as the area of a collection of contiguous grid points exceeding an accumulation threshold. While both 3- and 1-km entities were too large compared to those observed, 1-km entities were always smaller than 3-km entities and closest to observations (Fig. 5), and there were fewer 3- and 1-km entities than those observed (not shown). These results suggest upscale growth may have occurred too frequently in both forecast sets.

Compared to the Plains and MRB metaregions, East metaregion entities were smaller (Fig. 5), indicative of disorganized precipitation, while larger entities in the other two metaregions suggested organized features like MCSs were common. Moreover, forecast and observed springtime areal coverage peaks between 1800 and 0000 UTC over the East and 0000 and 0600 UTC over the Plains and MRB metaregions (Figs. 2b–e,g–j,l–o) were consistent with diurnally-driven convection over the East metaregion and larger nocturnal systems elsewhere. While summertime entities were smaller than springtime

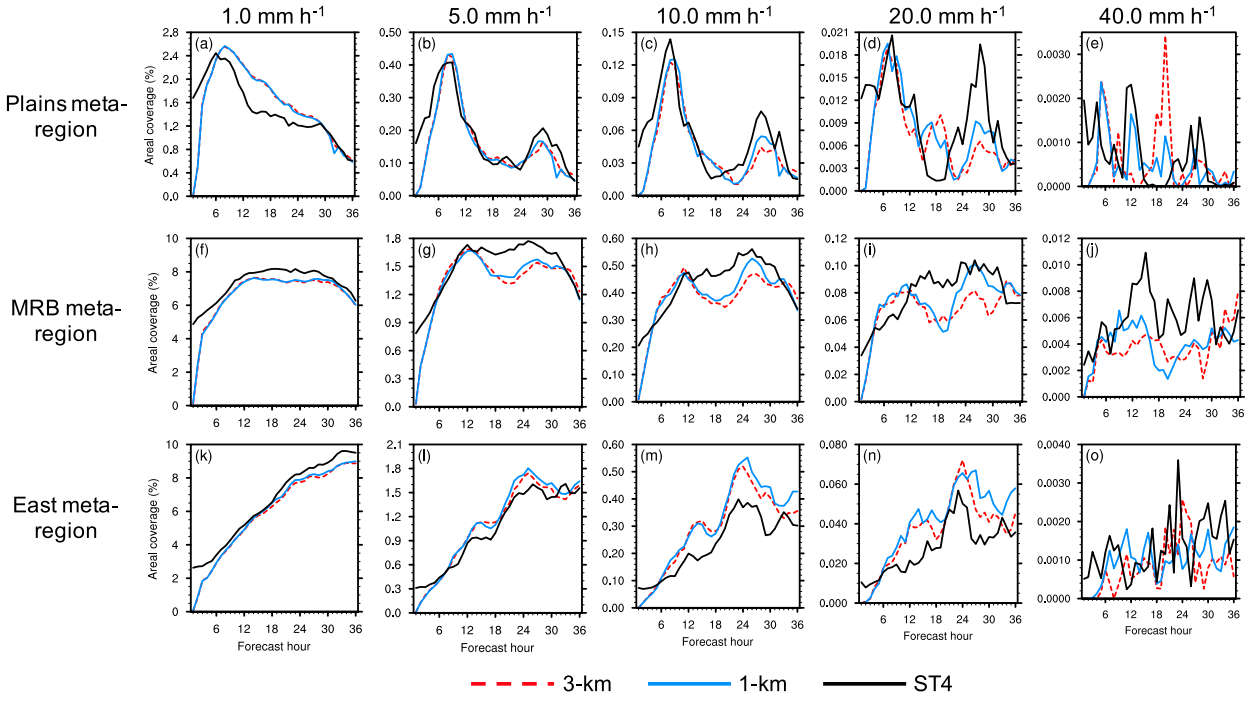

FIG. 4. As in Fig. 2, but for aggregate areal coverages over the 78 cool season (15 Oct–14 Mar) forecasts.

entities across all metaregions (Fig. 5), the decrease was largest over the MRB metaregion, where summertime areal coverages usually peaked earlier than in spring (cf. Figs. 2g–j, 3g–j). These shifts suggest more diurnally-driven precipitation over the MRB metaregion in summer compared to spring, perhaps contributing to the large summertime 3-km overprediction for thresholds $\geq 10.0\,\text{mm}\,\text{h}^{-1}$ (Figs. 3h–j) as over the East metaregion. Generally, both areal coverage and entity size patterns are consistent with observations that MCSs contribute less to warm season

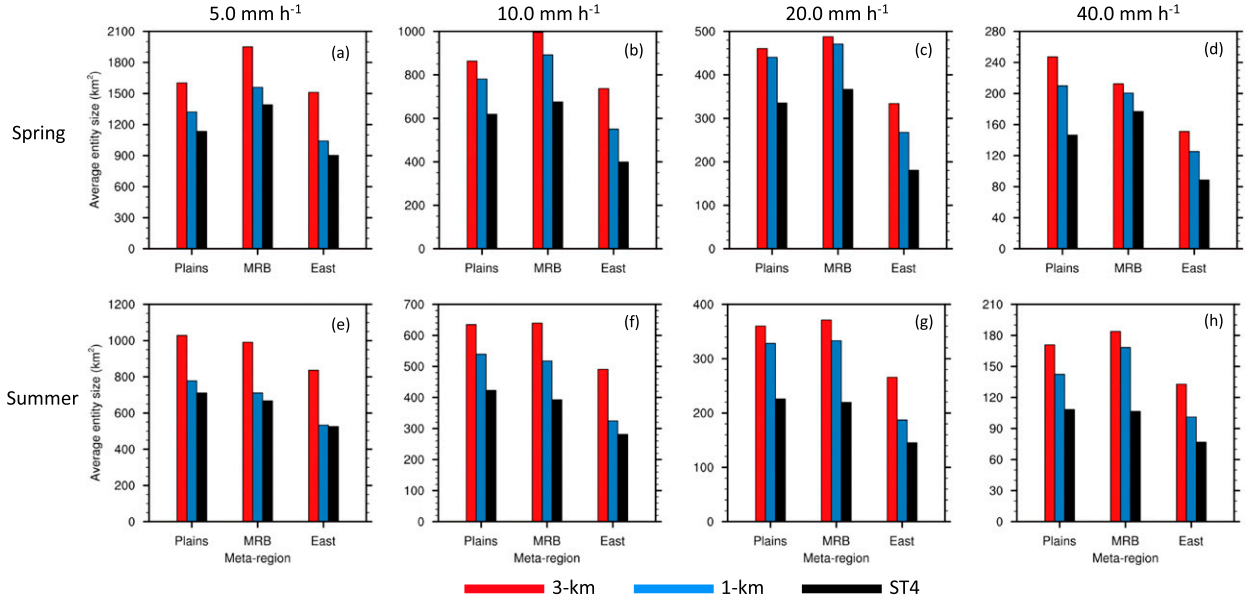

FIG. 5. Average entity size (km^2) over all 279 (a)–(d) springtime (15 Mar–14 Jun) 18–36-h forecasts of 1-h accumulated precipitation, where entities were defined as collections of contiguous grid points exceeding (a) 5.0, (b) 10.0, (c) 20.0, and (d) $40.0\,\text{mm}\,\text{h}^{-1}$ thresholds. (e)–(h) As in (a)–(d), but for average entity size over all 140 summertime (15 Jun–15 Jul) 18–36-h forecasts of 1-h accumulated precipitation.

rainfall over the east coast than other areas east of the Rockies (e.g., Haberlie and Ashley 2019).

c. Synthesis

Overall, 1-km areal coverages were typically closer to those observed than 3-km coverages, especially for thresholds $\geq 5.0\,\text{mm h}^{-1}$, similar to findings from S17. However, behaviors differed spatially and seasonally: while 3- and 1-km precipitation forecasts overpredicted areal coverages for some metaregions and seasons (e.g., East metaregion in spring), underprediction occurred in others (e.g., Plains metaregion in spring), and 3-km coverages were not always higher than 1-km coverages.

The smaller 1-km entities may be more prone to turbulent entrainment within convective updrafts than the larger 3-km entities (e.g., Bryan and Morrison 2012), particularly when updrafts and entities are small, as over the East metaregion in spring and summer (Fig. 5). Therefore, enhanced entrainment into relatively small entities may explain why 1-km coverages at thresholds $\geq 10.0\,\text{mm h}^{-1}$ were closer to those observed and lower than 3-km coverages over the East metaregion during the warm season (Figs. 2m–o, 3m–o).

Conversely, arguments concerning entrainment do not appear to explain why 3-km forecasts had lower, more biased, springtime coverages than 1-km forecasts for relatively large entities over the Plains metaregion (e.g., Figs. 2b–d). Differences between the 3- and 1-km springtime coverages over the Plains metaregion may be due to improved representation of trailing stratiform precipitation regions in 1-km MCSs, although further work is needed to provide insights about resolution dependence of MCS structures.

Finally, these collective results potentially add nuance to prior work that documented convection-allowing NWP models produce excessive rainfall over the central and eastern CONUS (e.g., Weisman et al. 2008; Clark et al. 2009, 2010a; S09; S17; Schwartz et al. 2010, 2015; J13). These previous studies solely examined forecasts over one verification region, and had we also only computed statistics over a single area, the interesting regional characteristics would have been unquantifiable. Thus, future verification efforts should strongly consider geographic heterogeneity of precipitation climatologies.

5. Daily, seasonal, and regional variations of forecast skill

a. Statistics over the entire central–eastern CONUS

1) Aggregate FSSs

Aggregate FSSs over all 497 forecasts and the $\text{CONUS}_{2/3}$ metaregion (Fig. 1; Table 5) indicated 1-km forecasts overall performed best, with statistically significant differences between 3- and 1-km FSSs for thresholds $\leq 10.0\,\text{mm h}^{-1}$ at nearly every forecast hour for all r (Figs. 6a–c). The biggest benefit of 1-km Δx compared to 3-km Δx occurred in spring (Figs. 6d–f), with FSSs resembling those over all 497 forecasts (the 279 springtime forecasts dominated statistics over all 497). Conversely, differences between 3- and 1-km FSSs were typically small and statistically insignificant during summer (Figs. 6g–i), especially for thresholds $\geq 5.0\,\text{mm h}^{-1}$. Like spring, 1-km FSSs were regularly higher than 3-km FSSs in the cool season (Figs. 6j–l), but statistically significant differences were primarily confined to before 30 h; thereafter, 3- and 1-km FSSs were typically similar. This convergence may be related to rapid CAPE decreases after 30 h over southern regions during the cool season (not shown), as sensitivity to Δx was greatest under conditions with moderate–strong forcing and large CAPE (discussed in section 5b).

2) Daily FSSs

On a day-to-day basis, 3- and 1-km FSSs aggregated over 18–36-h forecasts of 1-h accumulated precipitation were strongly correlated over the $\text{CONUS}_{2/3}$ metaregion, with Spearman rank correlation coefficients (ρ) $\geq$ 0.74 (Figs. 7a–e). FSSs were typically highest in the cool season and lowest in summer for thresholds $\leq 10.0\,\text{mm h}^{-1}$, while at heavier precipitation rates grouping by season was less apparent. These results indicate good 3-km forecasts were usually associated with good 1-km forecasts, suggesting common ICs, LBCs, and physics may have constrained how much corresponding 3- and 1-km forecasts diverged.

Probability density functions (PDFs) of daily differences between 1- and 3-km FSSs aggregated over 18–36-h forecasts of 1-h precipitation (FSS_1–FSS_3) were quasi-Gaussian and further revealed different seasonal behaviors (Fig. 7f–j). Cool season PDFs were typically sharper than springtime and summertime PDFs, indicating relatively small cool season differences between 3- and 1-km FSSs were common.

Nonetheless, despite regularly close cool season 3- and 1-km FSSs, cool season PDF peaks were > 0, and 1-km FSSs were greater than or equal to 3-km FSSs for a majority (i.e., >50%) of cool season cases (Figs. 7f–j). Results were similar during springtime. Conversely, for thresholds $\geq 10.0\,\text{mm h}^{-1}$, summertime PDF peaks were <0 and 3-km forecasts had higher FSSs for the majority of cases. Additionally, for the 1.0 and $5.0\,\text{mm h}^{-1}$ thresholds, the proportion

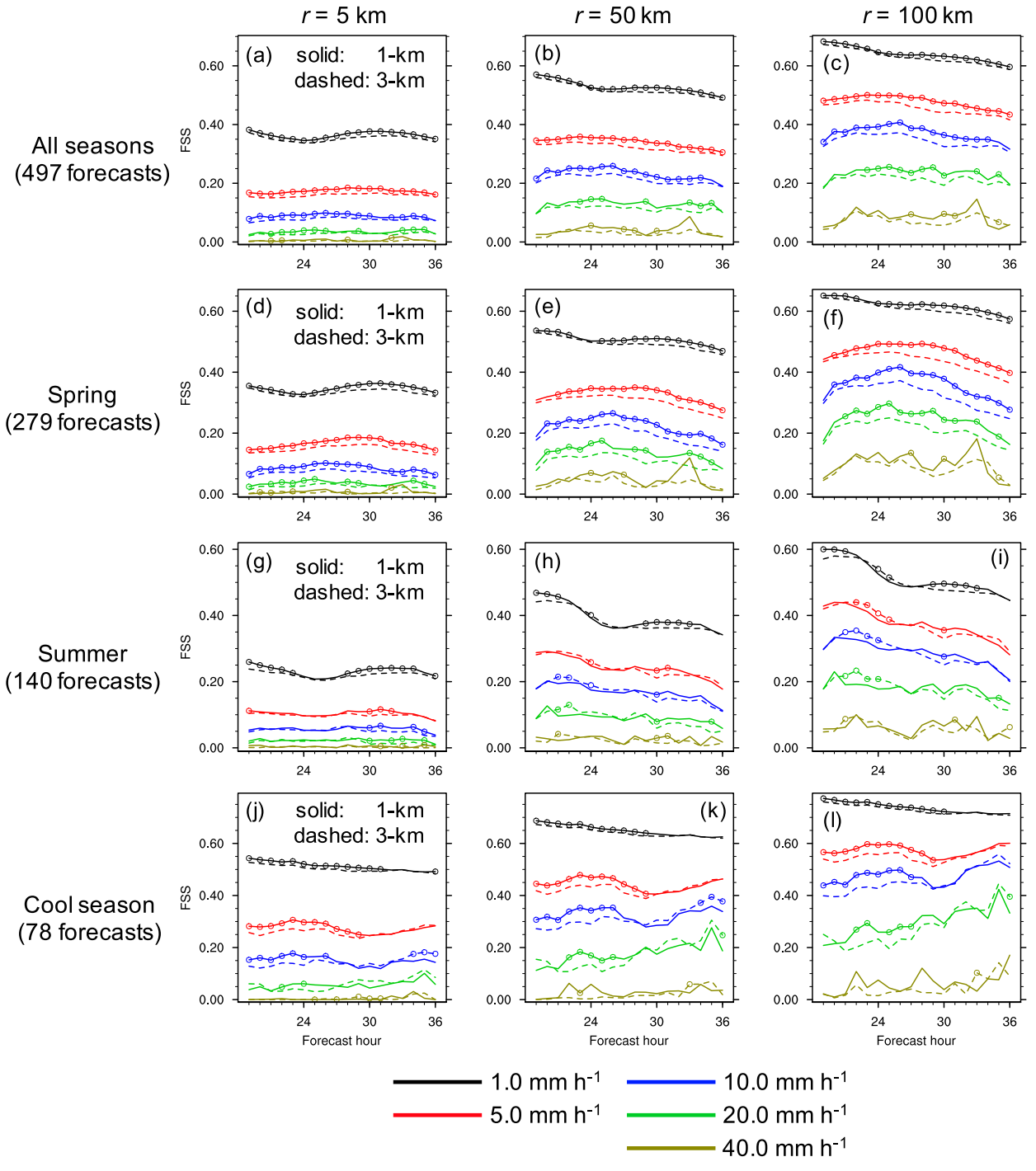

FIG. 6. FSSs over the CONUS$_{2/3}$ metaregion (Fig. 1; Table 5) for (a),(d),(g),(j) 5-, (b),(e),(h),(k) 50-, and (c),(f),(i),(l) 100-km neighborhood length scales aggregated over all (a)–(c) 497 forecasts and (d)–(f) 279 springtime (15 Mar–14 Jun), (g)–(i) 140 summertime (15 Jun–15 Jul), and (j)–(l) 78 cool season (15 Oct–14 Mar) forecasts as a function of forecast hour. Values on the x axis represent ending hours of 1-h accumulation periods and begin at hour 19 (i.e., the first x-axis value is for 1-h accumulated precipitation between 18 and 19 h). FSSs are shown for different event exceedance thresholds (legend), with 3- and 1-km FSSs given by dashed and solid lines, respectively. Circles on the curves denote instances when differences between 3- and 1-km forecasts for a particular threshold were statistically significant at the 95% level, with the circles placed on the curve with the higher FSS. For example, black circles on black solid lines indicate when 1-km forecasts had statistically significantly higher FSSs than 3-km forecasts at the 1.0 mm h^{-1} threshold, while blue circles on dashed blue lines denote when 3-km forecasts had statistically significantly higher FSSs than 1-km forecasts at the 10.0 mm h^{-1} threshold.

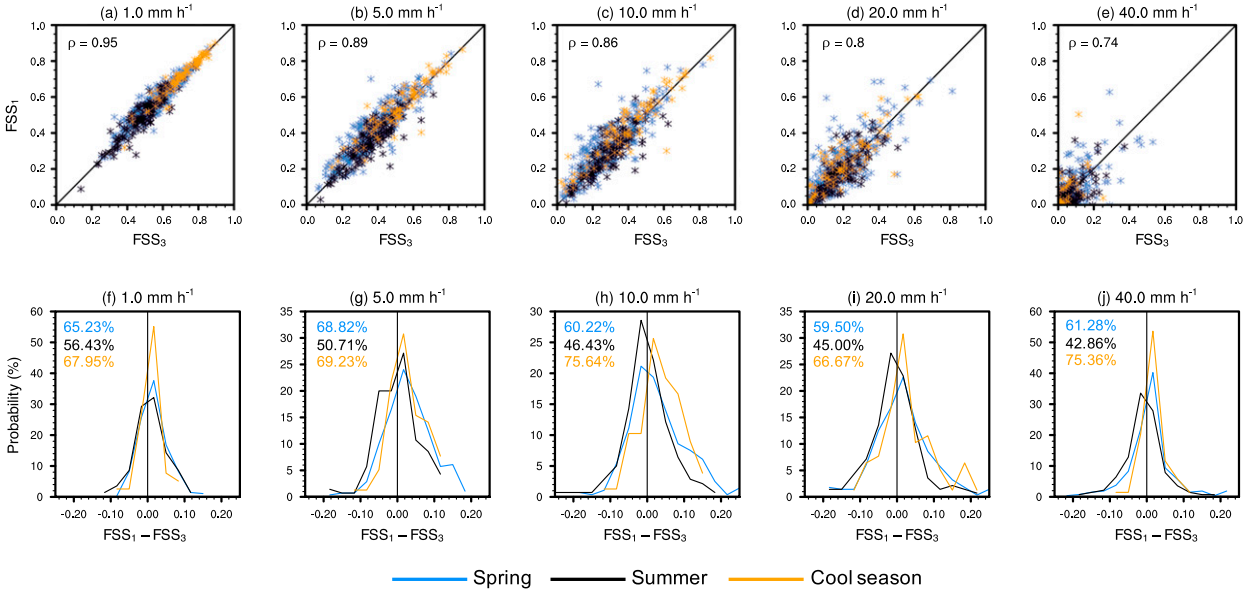

FIG. 7. (a)–(e) Scatterplots comparing FSSs aggregated over daily 3-km (x axis) and 1-km (y axis) 18–36-h forecasts of 1-h accumulated precipitation computed with $r = 100$ km over the $CONUS_{2/3}$ metaregion (Fig. 1; Table 5) for the (a) 1.0, (b) 5.0, (c) 10.0, (d) 20.0, and (e) 40.0 mm h^{-1} thresholds. There are 497 points per panel—one for each forecast—colored according to season, as indicated by the legend. Spearman rank correlation coefficients (ρ) are shown in each panel. (f)–(j) PDFs (%) of the daily differences between 1- and 3-km FSSs (FSS_1–FSS_3) based on the data in (a)–(e) for the spring (blue), summer (black), and cool season (orange) forecasts. Annotated percentages refer to the proportion of forecasts where $FSS_1 \geq FSS_3$ (i.e., area under the curves for $x \geq 0$) and are colored according to season (legend).

of summertime cases with (FSS_1–FSS_3) ≥ 0 was notably lower than in the cool season and spring.

3) Investigating seasonal and daily FSS variations

Results indicated both seasonal and daily variations of 3- and 1-km forecast skill and sensitivity to Δx (Figs. 6, 7). Thus, we attempted to quantify whether these daily and seasonal differences were associated with specific environmental properties (e.g., instability, forcing strength) or characteristics of the precipitation itself (e.g., areal coverage, entity size). Our ultimate goal was to understand those situations when forecasts with $\Delta x = 1$ km provided the largest benefits over forecasts with $\Delta x = 3$ km.

In aggregate, 1-km forecasts were usually better than 3-km forecasts in the spring and cool season but not during summer (Figs. 6 and 7), and within each season, on some days 1-km forecasts outperformed 3-km forecasts while on others they did not (Fig. 7). Unfortunately, we could not find any field that correlated even moderately with daily variations of differences between 1- and 3-km FSSs (i.e., FSS_1–FSS_3); in other words, we were unable to unearth a robust day-to-day statistical indicator associated with superior 1-km forecast skill when looking for meaningful correlations both within a single season or across all three seasons. For example, $|\rho|$ was just ~0.1 when comparing (FSS_1–FSS_3) with mean CAPE over the $CONUS_{2/3}$ metaregion over all 497 18–36-h forecasts.

Although discouraging, perhaps this result should have been unsurprising. Each daily scenario is unique, and the physical processes determining whether a 1-km forecast will be more skillful than a 3-km forecast likely change from case to case, rendering it difficult to uncover a strong daily statistical relationship between (FSS_1–FSS_3) and other fields. Thus, we can only broadly conclude that 1-km forecasts were most likely to outperform 3-km forecasts during the spring and cool season (Figs. 6 and 7).

Despite this disappointment, we did find specific fields that featured robust correlations with *closeness of 3- and 1-km forecasts to each other* [i.e., FSS_{1-3}; Eq. (4)], providing complementary information to Figs. 6 and 7 about sensitivity to Δx. One such field, the convective adjustment time scale (τ_c; Done et al. 2006; Zimmer et al. 2011), quantifies large-scale forcing strength. Following Surcel et al. (2017), τ_c (seconds) was defined as

$$\tau_c = \frac{1}{2}\frac{\text{MUCAPE}}{P} \times 49.58\ \text{mm s}^3\ \text{m}^{-2}\ \text{h}^{-1}, \qquad (5)$$

where MUCAPE (J kg^{-1}) is the most unstable CAPE at time T and P (mm h^{-1}) is 1-h accumulated precipitation ending at T. Small τ_c means the large-scale flow quickly

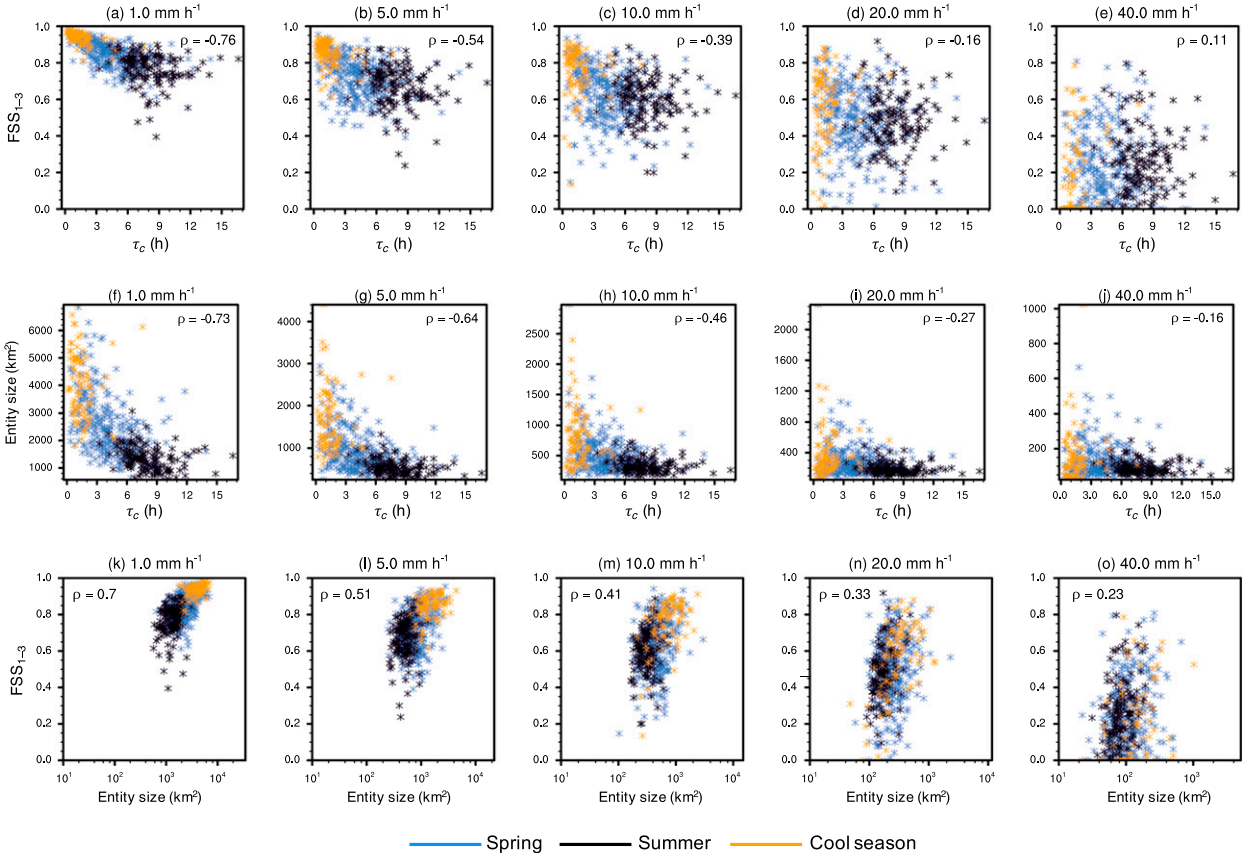

FIG. 8. (a)–(e) Scatterplots comparing convective adjustment time scale (τ_c; x axis, in hours) from 3-km forecasts to FSSs measuring 3- and 1-km forecast closeness (i.e., $FSS_{1–3}$; y axis) computed with $r = 100$ km for the (a) 1.0, (b) 5.0, (c) 10.0, (d) 20.0, and (e) 40.0 mm h^{-1} thresholds. Note that τ_c is insensitive to precipitation threshold but $FSS_{1–3}$ is not. The values were computed by aggregating over daily 18–36-h forecasts of 1-h accumulated precipitation and τ_c, respectively, over the $CONUS_{2/3}$ metaregion (Fig. 1; Table 5). There are 497 points per panel—one for each forecast—colored according to season, as indicated by the legend. Spearman rank correlation coefficients (ρ) are shown in each panel. (f)–(j),(k)–(o) As in (a)–(e), but (f)–(j) y-axis and (k)–(o) x-axis values are observed (i.e., ST4) entity size (km^2) aggregated over daily 18–36-h forecasts of 1-h accumulated precipitation.

removes convective instability, indicating strong synoptic-scale forcing scenarios, while large τ_c indicates weak synoptic forcing situations where convective instability festers (Done et al. 2006).

Equation (5) was applied to each point with $P \geq 0.25\,\text{mm}\,\text{h}^{-1}$, similar to Flack et al. (2016, 2018), and mean τ_c values were obtained by averaging over all points within the $CONUS_{2/3}$ metaregion. While many studies explicitly defined strong and weak forcing regimes based on τ_c thresholds (e.g., Molini et al. 2011; Keil et al. 2014; Kober et al. 2014; Flack et al. 2016, 2018), we did not make categorical distinctions and instead examined how metrics related to forecast skill (i.e., FSSs) correlated with τ_c, similar to Surcel et al. (2017). Moreover, because we were only interested in relative τ_c values, the importance of subjective choices like minimum precipitation threshold and whether to smooth MUCAPE and P before applying Eq. (5) was diminished.

Magnitudes of τ_c computed from 3-km forecasts over the $CONUS_{2/3}$ metaregion were broadly consistent with Surcel et al. (2016, 2017) and clearly delineated the seasons: τ_c was smallest during the cool season and largest in summer, with springtime τ_c in the middle (Figs. 8a–j). These findings reflect flow patterns over the CONUS that typically feature relatively strong synoptic-scale forcing during the cool season compared to summer. Moreover, for thresholds $\leq 10.0\,\text{mm}\,\text{h}^{-1}$, there were moderate to strong correlations between 3- and 1-km forecast similarity and forcing strength[3] (Figs. 8a–c), meaning smaller cool season and larger

[3] Increasing the minimum precipitation threshold to compute τ_c from 0.25 to 1.0 and 5.0 mm h^{-1} progressively decreased domain-average τ_c but did not impact seasonal delineations or correlations with $FSS_{1–3}$.

summertime τ_c corresponded to more similar and disparate 3- and 1-km forecasts (i.e., FSS_{1-3}), respectively. However, for thresholds $\geq$20.0 mm h^{-1}, these associations regarding forcing and 3- and 1-km forecast similarity did not hold (Figs. 8d,e), possibly due to smaller sample sizes and because τ_c does not effectively measure small-scale processes often responsible for locally heavy precipitation. Values of τ_c also indicate forecast quality was related to forcing strength, with higher FSSs in more strongly forced regimes (e.g., Figs. 6 and 7), consistent with several studies indicating convection-allowing NWP models perform best under strong forcing (e.g., Duda and Gallus 2013; Keil et al. 2014; Sobash and Kain 2017).

Forcing strength also had meaningful correlations with precipitation entity size, with larger cool season and smaller summertime ST4 entities associated with stronger and weaker forcing, respectively (Figs. 8f–j). Moreover, similar to τ_c, ST4 entity size was moderately to strongly correlated with 3- and 1-km forecast closeness (Figs. 8k–o) and clearly demarcated the seasons, particularly for thresholds $\leq$10.0 mm h^{-1} where sample sizes were largest.

Although correlations have limitations for establishing attribution, the relationship between entity size and τ_c (Figs. 8f–j) was likely causal; large precipitation areas are often driven by strong forcing. Also, while MUCAPE explained τ_c variations more than P, and thus had similar correlations with FSS_{1-3} as τ_c (not shown), CAPE alone is insufficient to produce precipitation. Thus, from a physical perspective, it seems more sensible to discuss variations in forcing (i.e., τ_c), rather than CAPE, as being associated with seasonal variations of precipitation characteristics and forecast skill.

4) Forecast skill as a function of entity size

Given that entity size had meaningful relationships with forcing strength and 3- and 1-km forecast closeness (Figs. 8f–o), we further examined forecast skill as a function of entity size during spring, when benefits of 1-km Δx were greatest (Fig. 6). Methodologically, the size of each entity within corresponding forecast and observation fields was first determined. Then, if a particular entity's area fell outside a preselected range, that entity was discarded, ultimately yielding forecast and observation grids consisting of solely those entities meeting prescribed size criteria, and FSSs were computed from these size-selected fields. This process was repeated for several size bins.

Although 1-km forecasts had statistically significantly higher FSSs than 3-km forecasts for all entity sizes, 3- and 1-km FSSs were relatively similar for entities with areas $<$10 000 km^2, with generally larger FSS differences for entities with areas $\geq$10 000 km^2 (Figs. 9a–c). These larger entities corresponded to MCSs, that, while relatively infrequent (Figs. 9d–f), produced a disproportionate share of rainfall; for example, just ~3% of ST4 entities at the 5.0 mm h^{-1} threshold over the $CONUS_{2/3}$ metaregion had areas $\geq$10 000 km^2 but produced ~50% of all ST4 precipitation with rates $\geq$5.0 mm h^{-1}. Therefore, larger entities likely dominated all-size FSSs (e.g., Figs. 6d–f), and it appears that much springtime 1-km improvement over 3-km forecasts was related to better forecasts of large precipitation systems, consistent with S17.

b. *Regional sensitivities to* Δx

1) Aggregate FSSs

During the weakly forced summer, there were few regional differences regarding sensitivity to Δx; across all regions (Fig. 1), 1-km FSSs were not systematically higher than 3-km FSSs (not shown). However, there were more regional differences in spring (Fig. 10). Over the GMC and LMV regions, 1-km forecasts were consistently better than 3-km forecasts (Figs. 10b,c), and most differences for thresholds $\leq$10.0 mm h^{-1} were statistically significant. The 1-km FSSs were also usually highest over the SPL, SEC, and MDW regions (Figs. 10a,d,f), but with fewer instances of statistically significant differences than over the GMC and LMV regions. Conversely, over the NPL, APL, and NEC regions (Figs. 10e,g,h), 3- and 1-km FSS differences were usually small.

Cool season regional differences before 30 h resembled those in spring, except there were fewer statistically significant differences and there were no signals 1-km FSSs were highest over the SPL region (Fig. 11). After 30 h, statistically significant differences between 3- and 1-km forecasts were uncommon across all regions, reflecting FSSs over the $CONUS_{2/3}$ metaregion (Figs. 6j–l).

2) Investigating springtime regional differences

Because the biggest regional differences occurred in spring, we focus on springtime environments and precipitation properties to understand regional variations of sensitivity to Δx. Regarding forcing strength, springtime τ_c variations across the regions were generally small compared to seasonal changes and did not indicate meaningfully different interregion forcing scenarios (not shown). However, for thresholds $\leq$10.0 mm h^{-1}, MUCAPE strongly correlated with regional differences between aggregate 1- and 3-km FSSs (i.e., FSS_1–FSS_3), considering all 279 18–36-h springtime forecasts

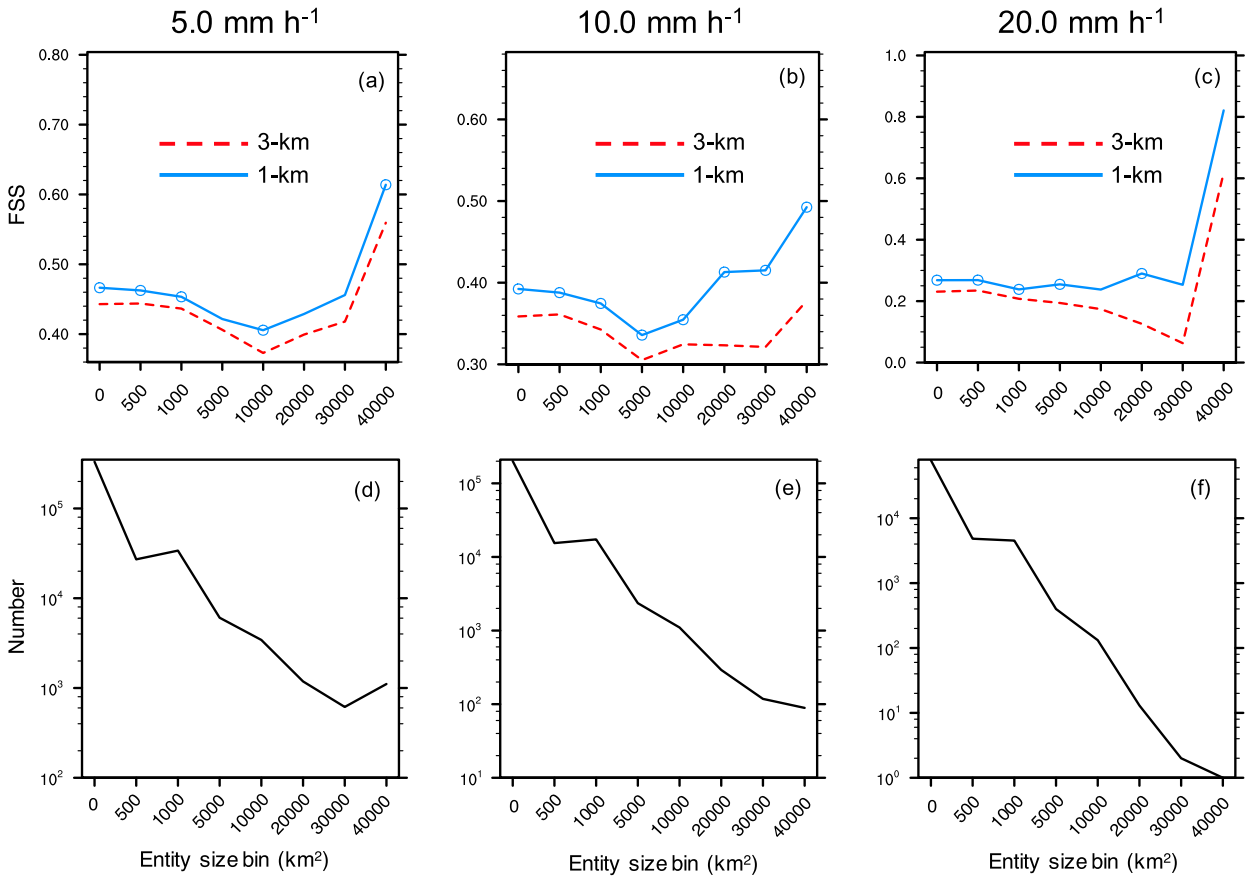

FIG. 9. FSSs for $r = 100$ km as a function of entity size (km^2) aggregated over all 279 springtime (15 Mar–14 Jun) 18–36-h forecasts of 1-h accumulated precipitation for the CONUS$_{2/3}$ metaregion (Fig. 1; Table 5) and (a) 5.0, (b) 10.0, and (c) 20.0 mm h^{-1} thresholds. Values on the x axis denote beginning bounds of a particular size bin (i.e., the leftmost bin encompasses entities with areas <500 km^2). Circles on the curves denote instances when differences between 3- and 1-km forecasts were statistically significant at the 95% level using a block-bootstrap resampling technique (section 3), with the circles placed on the curve with the higher FSS. (d)–(f) Number of observed (i.e., ST4) precipitation entities falling into each size bin for accumulation thresholds of (d) 5.0, (e) 10.0, and (f) 20.0 mm h^{-1}.

(Figs. 12a–e). For example, at most thresholds, (FSS$_1$–FSS$_3$) was largest over regions with relatively high MUCAPE (the southernmost SPL, GMC, LMV, and SEC regions) compared to those with smaller MUCAPE (the four northern regions).

There were also meaningful relationships between (FSS$_1$–FSS$_3$) and entity size, as regions with bigger entities tended to have larger (FSS$_1$–FSS$_3$), although the correspondence was not 1:1 across all thresholds (Figs. 12f–j). Entities were biggest over the GMC, LMV, and SPL regions, where MCSs contribute substantially to springtime rainfall (e.g., Haberlie and Ashley 2019) and MUCAPE was relatively high.

Thus, it appears 1-km skill was maximized relative to 3-km skill during the moderately–strongly forced springtime over southern regions, where instability was greatest and entities were largest. Because instability and forcing are primarily modulated by synoptic-scale flow, these regional results, coupled with the seasonal differences regarding sensitivity to Δx (Figs. 6–8), suggest the potential importance of large scales and climatology for fostering conditions when 1-km forecasts are most likely to yield benefits compared to 3-km forecasts.

6. Discussion and auxiliary experiments

Our results strongly contrast K08, S09, and J13, who all objectively assessed springtime precipitation forecast sensitivity to Δx over the central and eastern CONUS (Table 1). In attempt to reconcile our results with theirs, we performed several additional analyses and experiments, which included revisiting previous work.

a. Sample size

To examine whether disparities between K08's, S09's, and J13's results and ours were due to vastly different

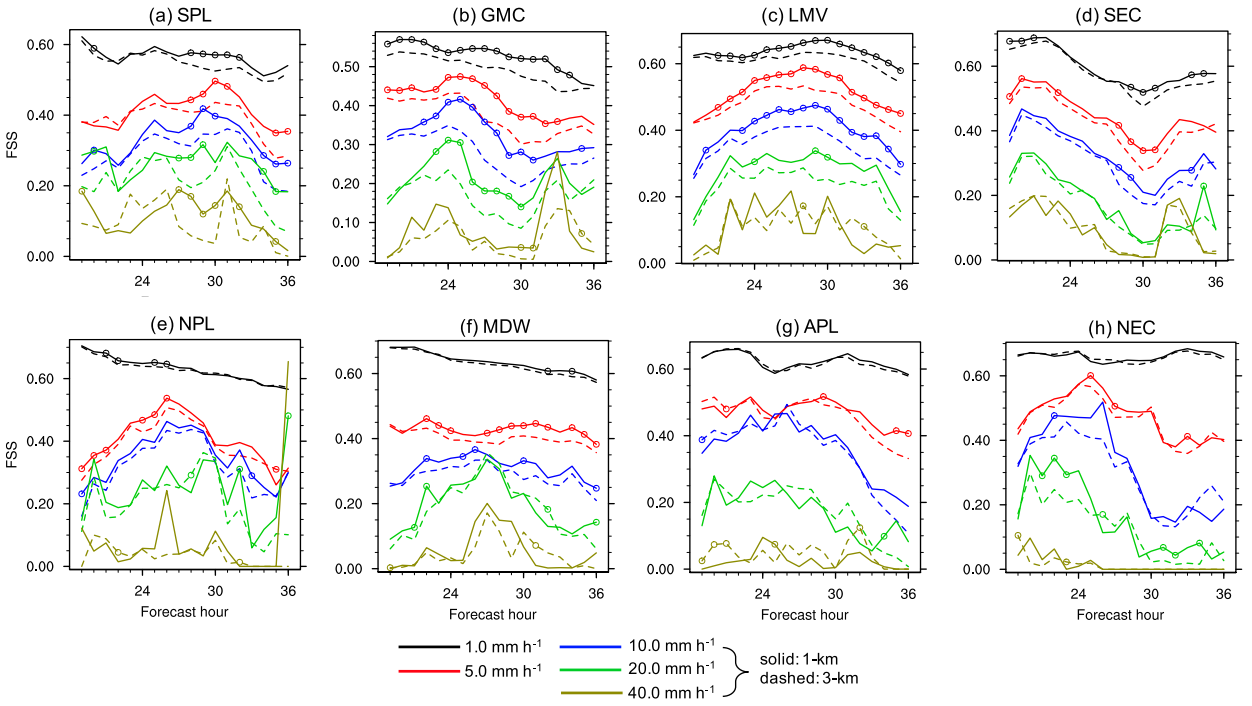

FIG. 10. FSSs over the (a) SPL, (b) GMC, (c) LMV, (d) SEC, (e) NPL, (f) MDW, (g) APL, and (h) NEC verification regions (Fig. 1) for a 100-km neighborhood length scale, aggregated over all 279 springtime (15 Mar–14 Jun) forecasts as a function of forecast hour. Values on the *x* axis represent ending hours of 1-h accumulation periods and begin at hour 19 (i.e., the first *x*-axis value is for 1-h accumulated precipitation between 18 and 19 h). FSSs are shown for different event exceedance thresholds (legend), with 3- and 1-km FSSs given by dashed and solid lines, respectively. Circles on the curves denote instances when differences between 3- and 1-km forecasts for a particular threshold were statistically significant at the 95% level, with the circles placed on the curve with the higher FSS. For example, black circles on black solid lines indicate when 1-km forecasts had statistically significantly higher FSSs than 3-km forecasts at the 1.0 mm h^{-1} threshold, while blue circles on dashed blue lines denote when 3-km forecasts had statistically significantly higher FSSs than 1-km forecasts at the 10.0 mm h^{-1} threshold.

sample sizes (e.g., Table 1), aggregate FSSs over 18–36-h forecasts of 1-h accumulated precipitation were computed for each possible consecutive 35-forecast window across all 497 forecasts; a 35-forecast window was chosen to match the smallest sample size among K08, S09, and J13. Across the 462 possible 35-forecast samples, aggregate 1-km FSSs were usually higher than 3-km FSSs (Figs. 13a–e) and the differences were regularly statistically significant in favor of the 1-km forecasts (Figs. 13f–j). Moreover, instances when 3-km FSSs were higher than 1-km FSSs occurred nearly exclusively for samples composed of mostly summertime forecasts. Therefore, these findings indicate it is possible, but ultimately unlikely, that differences concerning sample size explain our different results relative to K08, S09, and J13.

b. Model configurations and upgrades

As sampling probably cannot explain disparities with previous results, we examined whether changes to the WRF Model over the past decade could be responsible. Thus, we revisited the 33-h 2- and 4-km forecasts analyzed by S09, which were produced in spring 2007 over three-fourths of the CONUS.

First, using the exact configurations as S09 and version 2.2.1 of the WRF Model, we reproduced the 2100 UTC-initialized 2- and 4-km forecasts described in S09 on a modern supercomputer.[4] Then, we produced a second set of 2- and 4-km forecasts using identical ICs and LBCs as the first, but employed the WRF Model version (3.6.1) and configurations described in section 2b; these configurations, in particular, the physical parameterizations, were very different from those used in S09. We reproduced S09's forecasts instead of processing archived output to preclude the possibility that varied computing architectures (e.g., Hong et al. 2013) could impact comparisons between the two forecast sets.

Both forecast sets were verified with S09's methods, which included using Stage II (Lin and Mitchell 2005),

[4] The precise source code used in S09, based on WRF version 2.2, was unavailable.

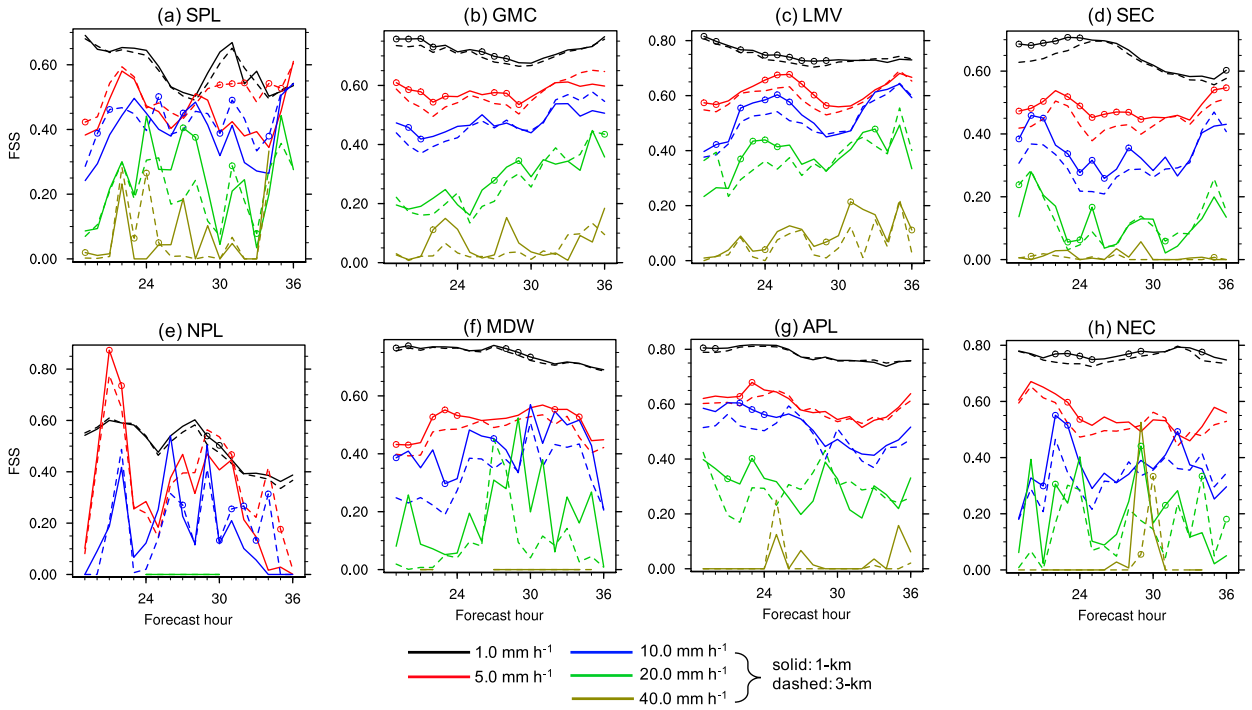

FIG. 11. As in Fig. 10, but for aggregate FSSs over the 78 cool season (15 Oct–14 Mar) forecasts.

rather than ST4 precipitation observations. While forecasts using version 3.6.1 of the WRF Model had markedly better domain-total precipitation than forecasts using version 2.2.1 (Fig. 14a), within both sets, relative differences between 2- and 4-km forecasts in terms of both domain-total precipitation and skill (Figs. 14b–f) indicated no benefits of 2-km over 4-km forecasts, corroborating S09.

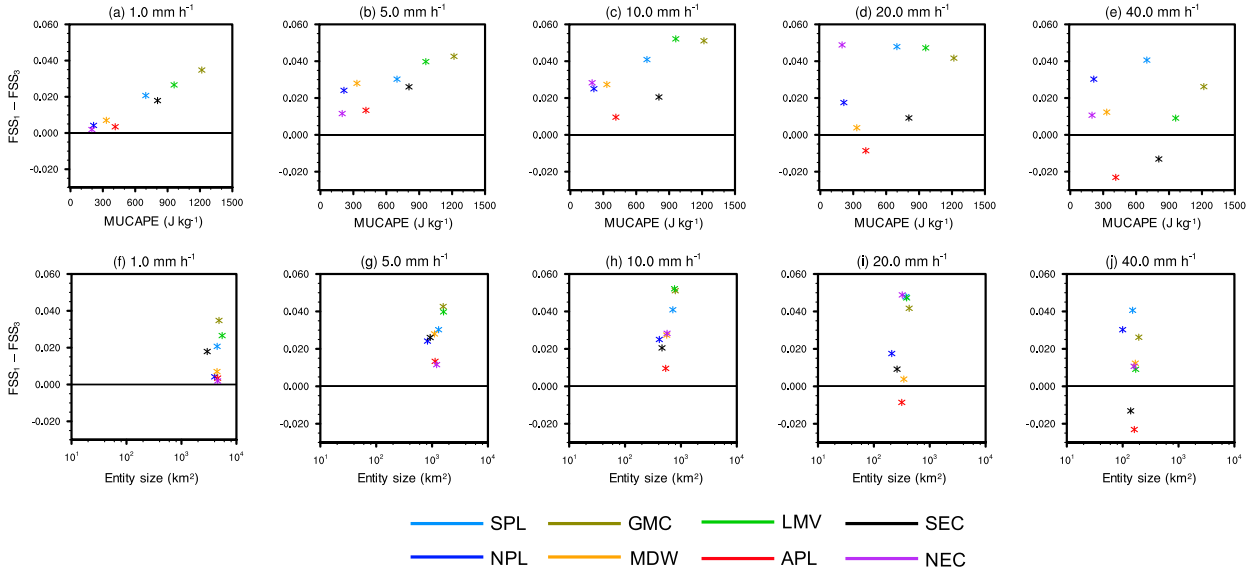

FIG. 12. (a)–(e) Scatterplots comparing MUCAPE (x axis; J kg^{-1}) from 3-km forecasts to differences between aggregate 1- and 3-km FSSs (FSS_1–FSS_3) with r = 100 km (y axis) for the (a) 1.0, (b) 5.0, (c) 10.0, (d) 20.0, and (e) 40.0 mm h^{-1} thresholds and different verification regions (legend; Fig. 1) over spring. MUCAPE values were obtained by averaging over all 279 springtime 18–36-h forecasts, while FSS values were obtained by aggregating over all 279 springtime 18–36-h forecasts of 1-h accumulated precipitation. Note that MUCAPE is insensitive to precipitation threshold but (FSS_1–FSS_3) is not. (f)–(j) As in (a)–(e), but x-axis values are observed (i.e., ST4) entity size (km^2) aggregated over all 279 springtime 18–36-h forecasts of 1-h accumulated precipitation.

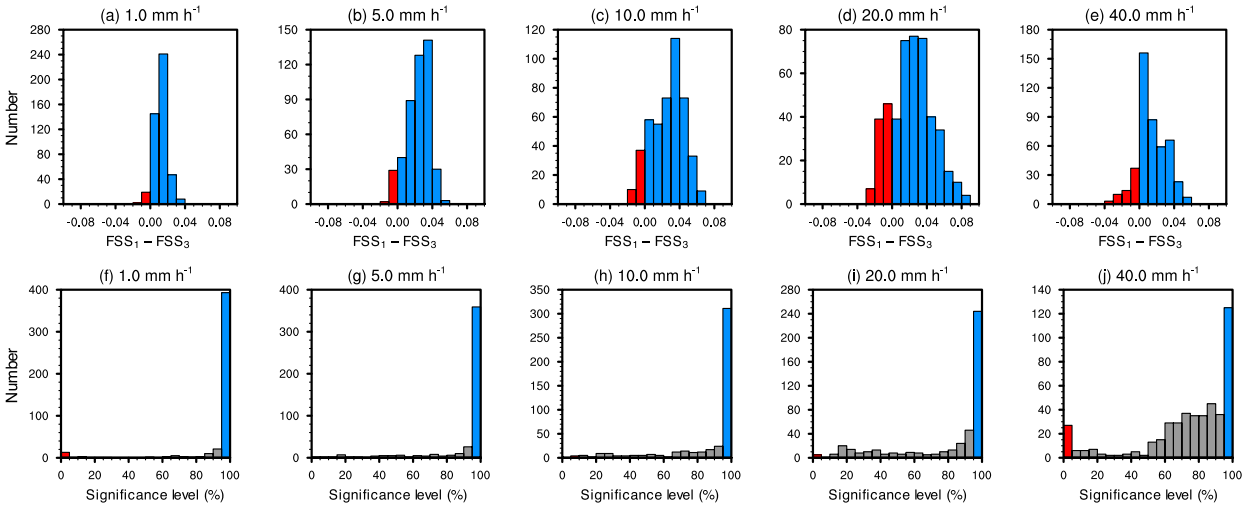

FIG. 13. (a)–(e) Histogram of differences between aggregate 1- and 3-km FSSs (FSS_1–FSS_3) for the (a) 1.0, (b) 5.0, (c) 10.0, (d) 20.0, and (e) 40.0 mm h^{-1} thresholds computed with $r = 100$ km, where FSSs were obtained by aggregating over 18–36-h forecasts of 1-h accumulated precipitation for each 35-forecast window over the $CONUS_{2/3}$ metaregion (Fig. 1; Table 5). Values > 0 (blue) indicate 1-km FSSs were larger than 3-km FSSs, while values < 0 (red) indicate the opposite. (f)–(j) As in (a)–(e), but for the significance level (%) of the difference between 1- and 3-km FSSs for each 35-forecast window, as determined by a block-bootstrap resampling technique (section 3). Significance levels >95% (blue) and <5% (red) indicate differences deemed statistically significant in favor of the 1- and 3-km forecasts, respectively.

Therefore, changes to WRF Model version and physics probably do not explain differences between our results and K08, S09, and J13. Moreover, these findings (i.e., Fig. 14) were insensitive to whether forecasts were verified over all forecasts or solely those corresponding to archived SPC severe weather events (e.g., section 2a), suggesting our case selection strategy did not introduce biases into our 3- versus 1-km results.

c. *Initial condition quality*

Forecasts examined by K08, S09, and J13 were initialized from analyses produced by three-dimensional variational (3DVAR) data assimilation (DA) systems. However, after 2010, most operational centers began transitioning to ''hybrid'' variational-ensemble DA systems that, unlike 3DVAR systems, incorporate ensemble-based flow-dependent background error covariances to improve use of assimilated observations. Thus, hybrid analyses are typically better than 3DVAR analyses and initialize superior forecasts than 3DVAR-based ICs (e.g., Hamill et al. 2011; Wang et al. 2013; Zhang et al. 2013; Schwartz and Liu 2014; Schwartz 2016).

NCEP's operational GFS model transitioned from 3DVAR to hybrid DA on 22 May 2012 (e.g., Wang et al. 2013), meaning 360 of our 497 3- and 1-km forecasts were initialized from hybrid-based ICs. Therefore, we wondered if our use of presumably better-quality ICs than K08, S09, and J13 was related to differences between our and their results regarding sensitivity to Δx. Although there were no indications relationships between our 3- and 1-km forecasts differed for pre– and post–22 May 2012 forecasts, to properly assess whether IC quality impacts sensitivity to Δx, a set of demonstrably ''good'' and ''bad'' ICs over a common period is required.

So, we revisited the analysis systems from Schwartz and Liu (2014, hereafter SL14), who produced continuously cycling 20-km hybrid and 3DVAR analyses for a consecutive 44-day period spanning May–June 2011 over the CONUS and unequivocally showed downscaled 20-km hybrid analyses initialized better 4-km precipitation forecasts than downscaled 20-km 3DVAR analyses. Thus, we considered SL14's 20-km hybrid and 3DVAR analyses as ''good'' and ''bad'' ICs, respectively.

The 20-km hybrid and 3DVAR ICs from SL14 were available, which we used in conjunction with the model version and configurations described in section 2b to initialize 36-h 2- and 4-km forecasts from 0000 UTC analyses that were identical except for Δx and time step (2- and 4-km ICs were produced by downscaling 20-km analyses). FSSs comparing the forecasts to ST4 observations indicated differences between 2- and 4-km hybrid forecasts (gray lines in Fig. 15) were typically larger than those between 2- and 4-km 3DVAR forecasts (orange lines in Fig. 15), and identical results were obtained when verifying over all 44 forecasts and solely those forecasts corresponding to SPC severe weather events (section 2a), again suggesting our findings about

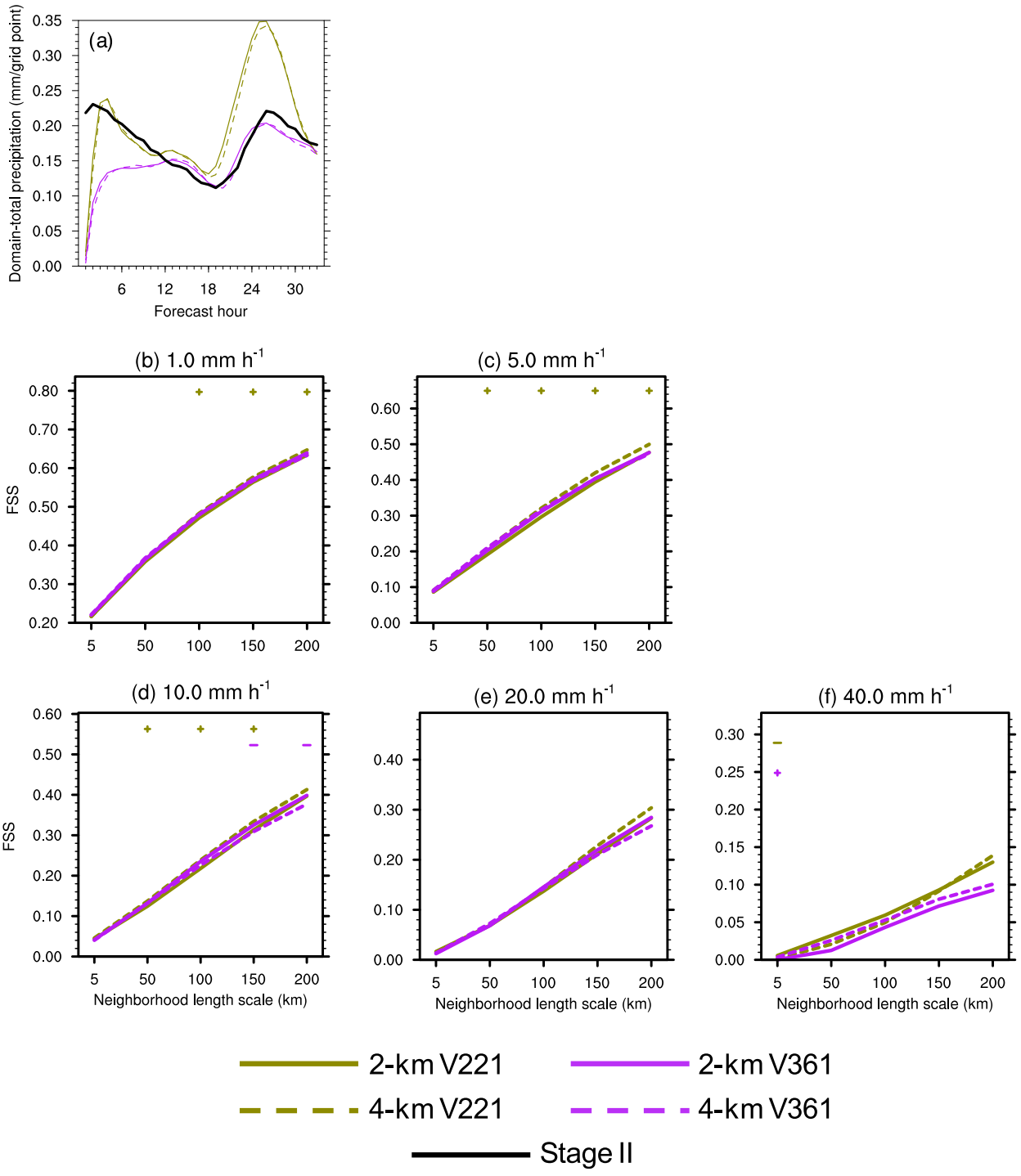

FIG. 14. (a) Aggregate 1-h accumulated precipitation (mm) per grid point over 37 corresponding 2- and 4-km forecasts produced with version 3.6.1 ("2-km V361," "4-km V361") and 2.2.1 ("2-km V221," "4-km V221") of the WRF Model as a function of forecast hour over a region spanning the central CONUS, as in S09 (stippled region in Fig. 1). Values on the *x* axis represent ending hours of 1-h accumulation periods (e.g., an *x*-axis value of 24 is for 1-h accumulated precipitation between 23 and 24 h). (b)–(f) FSSs as a function of neighborhood length scale (km) aggregated over 37 21–33-h forecasts of 1-h accumulated precipitation for the (b) 1.0, (c) 5.0, (d) 10.0, (e) 20.0, and (f) 40.0 mm h^{-1} thresholds over the stippled region in Fig. 1. Symbols along the top axis indicate differences between 2- and 4-km forecasts employing the same model version were statistically significant at the 95% level using a block-bootstrap resampling technique, with "−" and "+" symbols indicating better 2- and 4-km performance, respectively.

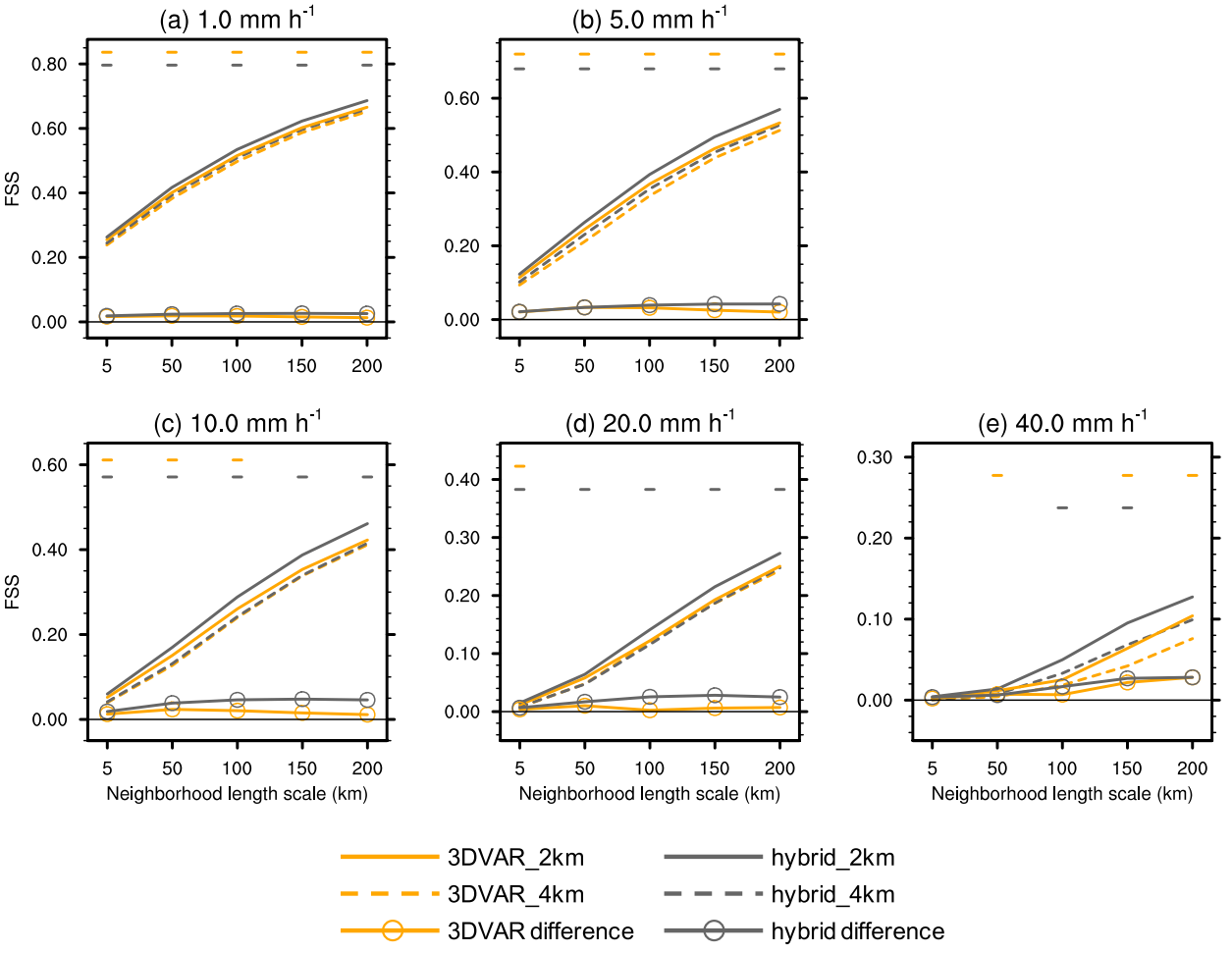

FIG. 15. FSSs as a function of neighborhood length scale (km) aggregated over 44 18–36-h forecasts of 1-h accumulated precipitation for the (a) 1.0, (b) 5.0, (c) 10.0, (d) 20.0, and (e) 40.0 mm h^{-1} thresholds over the stippled region in Fig. 1 for corresponding 2- and 4-km forecasts initialized from hybrid ("hybrid_2km," "hybrid_4km") and 3DVAR ("3DVAR_2km," "3DVAR_4km") analyses. Symbols along the top axis indicate differences between 2- and 4-km forecasts using the same analysis method (i.e., 3DVAR or hybrid) were statistically significant at the 95% level using a block-bootstrap resampling technique, with "−" and "+" symbols indicating better 2- and 4-km performance, respectively. Lines with circular markers near FSS = 0 are differences between the 2- and 4-km FSSs that used a common analysis method (2 km minus 4 km).

3- and 1-km forecast quality were unaffected by our case selection approach.

Therefore, when better (hybrid) ICs were used, there were greater benefits of 2-km Δx compared to 4-km Δx than when poorer (3DVAR) ICs were employed. These findings suggest that perhaps improved analysis quality can indeed translate into benefits of finer Δx at convection-allowing scales and may explain discrepancies between our findings regarding sensitivity to Δx and those from K08, S09, and J13. Exactly why hybrid-based analyses permitted greater benefits of smaller Δx is unclear, but in general, if improved DA reduces large-scale IC errors that in turn lessen large-scale errors at next-day lead times, the smaller large-scale errors may foster relatively uncontaminated environments in which convection can develop and allow intrinsic benefits of finer Δx to be realized. Clearly, much more work in both real-data and idealized scenarios is needed to confirm and elucidate any relationship between analysis quality and sensitivity to Δx in convection-allowing models.

7. Summary and conclusions

This study examined 497 corresponding 3- and 1-km forecasts over the CONUS east of the Rockies on days with severe weather reports. Model climatologies of precipitation revealed seasonally and geographically varying biases and behaviors, although 1-km precipitation distributions were usually closer to those observed than 3-km distributions. Furthermore, precipitation entity

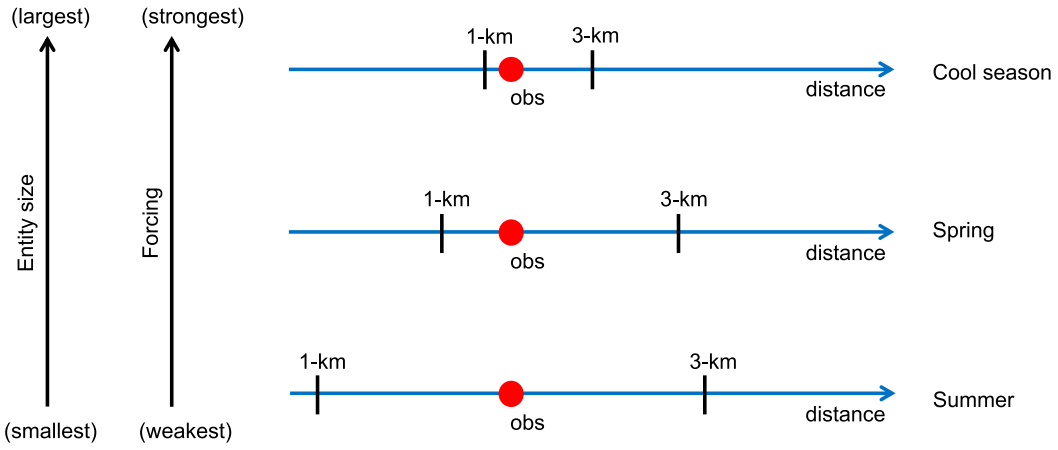

FIG. 16. Schematic diagram summarizing relationships of 3- and 1-km precipitation forecasts to each other and observations during the springtime, summertime, and cool season, considering the entire CONUS east of the Rockies. Blue lines represent a generic measure of distance, red circles labeled "obs" refer to ST4 observations, and black vertical lines denote locations of 3- and 1-km forecasts with respect to each other and observations. Relative forcing strength and entity size in each season are shown by arrows on the left.

sizes in 1-km forecasts more closely matched observed entity sizes than 3-km forecasts.

Regarding spatial skill, cool season 3- and 1-km precipitation forecasts had the highest FSSs and looked most like each other, with 1-km forecasts typically closer to observations than 3-km forecasts (Fig. 16). Benefits of 1-km Δx were maximized in spring, where, compared to the cool season, FSSs were lower and 3- and 1-km forecasts looked less alike, but 1-km forecasts were substantially closer to observations than 3-km forecasts. Finally, 3- and 1-km summertime forecasts produced the lowest FSSs and looked least like each other but were equidistant from observations. These seasonal differences concerning sensitivity to Δx were associated with forcing strength, with 1-km forecasts most likely to outperform 3-km forecasts when forcing was stronger (the spring and cool season) and attendant precipitation entities were larger (i.e., MCSs). During springtime, benefits of 1-km forecasts were largest over southern regions, where instability was greatest. Springtime and cool season improvements from decreasing Δx from 3 to 1 km may represent a lower bound, given that many settings, like number of vertical levels, were not optimized for 1-km Δx.

Why were benefits of 1-km Δx primarily confined to larger springtime and cool season precipitation systems 3- and 1-km forecasts likely similarly resolve? One possibility for this counterintuitive finding regards predictability limits, which are constrained for localized phenomena (e.g., Lorenz 1969). Moreover, Surcel et al. (2017) suggested convection-allowing forecasts over the central–eastern CONUS lose predictability for scales $< \sim$200 km after 18 h. Thus, if very localized—even stochastic—processes on inherently unpredictable scales govern placement of small precipitation features, finer Δx should not be expected to yield improvements, despite presumably more realistic representations of atmospheric processes. Conversely, for larger, more predictable systems, like MCSs, improved representation of physical processes afforded by finer Δx may ultimately translate into forecast improvements, consistent with the greatest springtime benefits of 1-km forecasts over the GMC and LMV regions, where entities were relatively big. We note that while Sobash et al. (2019) found these same 1-km forecasts produced better next-day tornado guidance than 3-km forecasts, there was no evidence better 1-km tornado forecasts were attributable to more accurate placement of severe convection; rather, improved physical representation of low-level rotation in the 1-km forecasts likely was key to yielding better tornado guidance than 3-km forecasts (e.g., Potvin and Flora 2015).

Our results differ from previous work collectively finding springtime forecasts with 4-km Δx had similar skill as forecasts with 1- or 2-km Δx over the central–eastern CONUS (e.g., K08; S09; Clark et al. 2012; J13; Loken et al. 2017). Initial experimentation suggests this disparity may be related to differences regarding IC quality between previous studies and our work. Thus, we encourage further research to examine whether better-quality ICs enable greater benefits of decreasing Δx toward 1 km.

We caution that our findings might be very different for topographically diverse regions, where forecasts with

$\Delta x = 1$ km could still be beneficial compared to forecasts with 3-km Δx under weak synoptic forcing. Moreover, although Potvin et al. (2017) showed 2-h convection-allowing forecasts were fairly insensitive to analysis resolution for a few cases, it is unclear how our 18–36-h results might change if 3- and 1-km forecasts are initialized from 3- and 1-km ICs, respectively, rather than from coarse 0.5° GFS analyses. Furthermore, sensitivity to Δx could itself potentially be sensitive to physics choices, although our results based on reproducing S09's forecasts provide some hope this may not be the case.

The 1-km forecasts were 27 times more expensive than the 3-km forecasts, and ultimately, individual users must decide whether higher resolution warrants the extra cost. As benefits of 1-km Δx were most pronounced in unstable, strongly forced environments—which are usually predictable days in advance—perhaps our results could be leveraged to temporarily increase convection-allowing model resolution in real-time systems when certain conditions are forecast.

Finally, 3-km ensembles will almost certainly outperform deterministic 1-km forecasts (e.g., Hagelin et al. 2017; Loken et al. 2017; Mittermaier and Csima 2017; S17), unless 3-km Δx is fundamentally too coarse to capture phenomena of interest. Thus, the trade-off between finer Δx and explicit probabilistic information provided by ensembles should be carefully considered in future research and operational NWP models.

Acknowledgments. This work was partially funded by NCAR's Short-term Explicit Prediction (STEP) program and NOAA/OAR Office of Weather and Air Quality Grant NA17OAR4590182. All forecasts were produced on NCAR's Cheyenne supercomputer (Computational and Information Systems Laboratory 2017). Thanks to Fanyou Kong (OU/CAPS) for providing archived files needed to reproduce forecasts from S09, Morris Weisman and James Done (NCAR/MMM) for internal reviews, and three anonymous reviewers for their suggestions. This project was pre-registered with the Center for Open Science. NCAR is sponsored by the National Science Foundation.

REFERENCES

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Wea. Forecasting*, **18**, 918–932, https://doi.org/10.1175/1520-0434(2003)018<0918:SOPFSS>2.0.CO;2.

Baldauf, M., A. Seifert, J. Förstner, D. Majewski, M. Raschendorfer, and T. Reinhardt, 2011: Operational convective-scale numerical weather prediction with the COSMO model: Description and sensitivities. *Mon. Wea. Rev.*, **139**, 3887–3905, https://doi.org/10.1175/MWR-D-10-05013.1.

Barthlott, C., B. Mühr, and C. Hoose, 2017: Sensitivity of the 2014 Pentecost storms over Germany to different model grids and microphysics schemes. *Quart. J. Roy. Meteor. Soc.*, **143**, 1485–1503, https://doi.org/10.1002/qj.3019.

Bartsotas, N. S., E. I. Nikolopoulos, E. N. Anagnostou, S. Solomos, and G. Kallos, 2017: Moving toward subkilometer modeling grid spacings: Impacts on atmospheric and hydrological simulations of extreme flash flood–inducing storms. *J. Hydrometeor.*, **18**, 209–226, https://doi.org/10.1175/JHM-D-16-0092.1.

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.

Blake, B. T., J. R. Carley, T. I. Alcott, I. Jankov, M. E. Pyle, S. E. Perfater, and B. Albright, 2018: An adaptive approach for the calculation of ensemble gridpoint probabilities. *Wea. Forecasting*, **33**, 1063–1080, https://doi.org/10.1175/WAF-D-18-0035.1.

Brousseau, P., Y. Seity, D. Ricard, and J. Léger, 2016: Improvement of the forecast of convective activity from the AROME-France system. *Quart. J. Roy. Meteor. Soc.*, **142**, 2231–2243, https://doi.org/10.1002/qj.2822.

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.

Buzzi, A., S. Davolio, P. Malguzzi, O. Drofa, and D. Mastrangelo, 2014: Heavy rainfall episodes over Liguria of autumn 2011: Numerical forecasting experiments. *Nat. Hazards Earth Syst. Sci.*, **14**, 1325–1340, https://doi.org/10.5194/nhess-14-1325-2014.

Chen, F., and J. Dudhia, 2001: Coupling an advanced land-surface–hydrology model with the Penn State–NCAR MM5 modeling system. Part I: Model description and implementation. *Mon. Wea. Rev.*, **129**, 569–585, 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 M. L. Weisman, 2010a: Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF Model simulations and the operational NAM. *Wea. Forecasting*, **25**, 1495–1509, https://doi.org/10.1175/2010WAF2222404.1.

——, ——, M. Xue, and F. Kong, 2010b: Growth of spread in convection-allowing and convection-parameterizing ensembles. *Wea. Forecasting*, **25**, 594–612, https://doi.org/10.1175/2009WAF2222318.1.

——, 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.

Clark, P., N. Roberts, H. Lean, S. P. Ballard, and C. Charlton-Perez, 2016: Convection-permitting models: A step-change in rainfall forecasting. *Meteor. Appl.*, **23**, 165–181, https://doi.org/10.1002/met.1538.

Colle, B. A., and C. F. Mass, 2000: The 5–9 February 1996 flooding event over the Pacific Northwest: Sensitivity studies and evaluation of the MM5 precipitation forecasts. *Mon. Wea. Rev.*, **128**, 593–618, https://doi.org/10.1175/1520-0493(2000)128<0593:TFFEOT>2.0.CO;2.

——, J. B. Wolfe, W. J. Steenburgh, D. E. Kingsmill, J. A. W. Cox, and J. C. Shafer, 2005: High-resolution simulations and microphysical validation of an orographic precipitation event over the Wasatch Mountains during IPEX IOP3. *Mon. Wea. Rev.*, **133**, 2947–2971, https://doi.org/10.1175/MWR3017.1.

Computational and Information Systems Laboratory, 2017: Cheyenne: HPE/SGI ICE XA System (NCAR Community Computing). National Center for Atmospheric Research, https://doi.org/10.5065/D6RX99HX.

Davis, C., W. Wang, J. Dudhia, and R. Torn, 2010: Does increased horizontal resolution improve hurricane wind forecasts? *Wea. Forecasting*, **25**, 1826–1841, https://doi.org/10.1175/2010WAF2222423.1.

Dey, S. R., G. Leoncini, N. M. Roberts, R. S. Plant, and S. Migliorini, 2014: A spatial view of ensemble spread in convection permitting ensembles. *Mon. Wea. Rev.*, **142**, 4091–4107, https://doi.org/10.1175/MWR-D-14-00172.1.

Done, J. M., G. C. Craig, S. L. Gray, P. A. Clark, and M. E. B. Gray, 2006: Mesoscale simulations of organized convection: Importance of convective equilibrium. *Quart. J. Roy. Meteor. Soc.*, **132**, 737–756, https://doi.org/10.1256/qj.04.84.

Duda, J. D., and W. A. Gallus Jr., 2013: The impact of large-scale forcing on skill of simulated convective initiation and upscale evolution with convection-allowing grid spacings in the WRF. *Wea. Forecasting*, **28**, 994–1018, https://doi.org/10.1175/WAF-D-13-00005.1.

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

Flack, D. L. A., R. S. Plant, S. L. Gray, H. W. Lean, C. Keil, and G. C. Craig, 2016: Characterisation of convective regimes over the British Isles. *Quart. J. Roy. Meteor. Soc.*, **142**, 1541–1553, https://doi.org/10.1002/qj.2758.

——, S. L. Gray, and R. S. Plant, 2018: Convective-scale perturbation growth across the spectrum of convective regimes. *Mon. Wea. Rev.*, **146**, 387–405, https://doi.org/10.1175/MWR-D-17-0024.1.

Garvert, M. F., B. A. Colle, and C. F. Mass, 2005: The 13–14 December 2001 IMPROVE-2 event. Part I: Synoptic and mesoscale evolution and comparison with a mesoscale model simulation. *J. Atmos. Sci.*, **62**, 3474–3492, https://doi.org/10.1175/JAS3549.1.

Gilleland, E., A. S. Hering, T. L. Fowler, and B. G. Brown, 2018: Testing the tests: What are the impacts of incorrect assumptions when applying confidence intervals or hypothesis tests to compare competing forecasts? *Mon. Wea. Rev.*, **146**, 1685–1703, https://doi.org/10.1175/MWR-D-17-0295.1.

Haberlie, A. M., and W. S. Ashley, 2019: A radar-based climatology of mesoscale convective systems in the United States. *J. Climate*, **32**, 1591–1606, https://doi.org/10.1175/JCLI-D-18-0559.1.

Hagelin, S., J. Son, R. Swinbank, A. McCabe, N. Roberts, and W. Tennant, 2017: The Met Office convective-scale ensemble, MOGREPS-UK. *Quart. J. Roy. Meteor. Soc.*, **143**, 2846–2861, https://doi.org/10.1002/qj.3135.

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.

——, J. S. Whitaker, D. T. Kleist, M. Fiorino, and S. G. Benjamin, 2011: Predictions of 2010's tropical cyclones using the GFS and ensemble-based data assimilation methods. *Mon. Wea. Rev.*, **139**, 3243–3247, https://doi.org/10.1175/MWR-D-11-00079.1.

Hirahara, Y., J. Ishida, and T. Ishimizu, 2011: Trial operation of the local forecast model at JMA. Research activities in atmospheric and oceanic modelling. CAS/JSC Working Group on Numerical Experimentation Rep. 41, WMO/TD-1578, 5.11–5.12, http://www.wcrp-climate.org/WGNE/BlueBook/2011/individual-articles/05_Hirahara_Youichi_WGNE_LFM.pdf.

Hong, S., M. Koo, J. Jang, J. E. Kim, H. Park, M. Joh, J. Kang, and T. Oh, 2013: An evaluation of the software system dependency of a global atmospheric model. *Mon. Wea. Rev.*, **141**, 4165–4172, https://doi.org/10.1175/MWR-D-12-00352.1.

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.

Ito, J., S. Hayashi, A. Hashimoto, H. Ohtake, F. Uno, H. Yoshimura, T. Kato, and Y. Yamada, 2017: Stalled improvement in a numerical weather prediction model as horizontal resolution increases to the sub-kilometer scale. *SOLA*, **13**, 151–156, https://doi.org/10.2151/sola.2017-028.

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.

——, 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.

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., 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.

Keil, C., F. Heinlein, and G. C. Craig, 2014: The convective adjustment time-scale as indicator of predictability of convective precipitation. *Quart. J. Roy. Meteor. Soc.*, **140**, 480–490, https://doi.org/10.1002/qj.2143.

Klasa, C., M. Arpagaus, A. Walser, and H. Wernli, 2018: An evaluation of the convection-permitting ensemble COSMO-E for three contrasting precipitation events in Switzerland. *Quart. J. Roy. Meteor. Soc.*, **144**, 744–764, https://doi.org/10.1002/qj.3245.

Kober, K., G. C. Craig, and C. Keil, 2014: Aspects of short-term probabilistic blending in different weather regimes. *Quart. J. Roy. Meteor. Soc.*, **140**, 1179–1188, https://doi.org/10.1002/qj.2220.

Kühnlein, C., C. Keil, G. C. Craig, and C. Gebhardt, 2014: The impact of downscaled initial condition perturbations on convective-scale ensemble forecasts of precipitation. *Quart. J. Roy. Meteor. Soc.*, **140**, 1552–1562, https://doi.org/10.1002/qj.2238.

Lean, H. W., P. A. Clark, M. Dixon, N. M. Roberts, A. Fitch, R. Forbes, and C. Halliwell, 2008: Characteristics of high-resolution versions of the Met Office Unified Model for forecasting convection over the United Kingdom. *Mon. Wea. Rev.*, **136**, 3408–3424, https://doi.org/10.1175/2008MWR2332.1.

Lin, Y., and K. E. Mitchell, 2005: The NCEP stage II/IV hourly precipitation analyses: Development and applications. *19th Conf. on Hydrology*, San Diego, CA, Amer. Meteor. Soc., 1.2, http://ams.confex.com/ams/pdfpapers/83847.pdf.

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., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289–307, https://doi.org/10.3402/tellusa.v21i3.10086.

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

MeteoSwiss, 2019: COSMO forecasting system. Accessed 5 March 2019, https://www.meteoswiss.admin.ch/home/measurement-and-forecasting-systems/warning-and-forecasting-systems/cosmo-forecasting-system/cosmo-1-high-resolution-forecasts-for-the-alpine-region.html.

Mittermaier, M., 2019: A ''meta'' analysis of the fractions skill score: The limiting case and implications for aggregation. *Mon. Wea. Rev.*, https://doi.org/10.1175/MWR-D-18-0106.1, in press.

——, and G. Csima, 2017: Ensemble versus deterministic performance at the kilometer scale. *Wea. Forecasting*, **32**, 1697–1709, https://doi.org/10.1175/WAF-D-16-0164.1.

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.

Molini, L., A. Parodi, N. Rebora, and G. Craig, 2011: Classifying severe rainfall events over Italy by hydrometeorological and dynamical criteria. *Quart. J. Roy. Meteor. Soc.*, **137**, 148–154, https://doi.org/10.1002/qj.741.

Nelson, B. R., O. P. Prat, D.-J. Seo, and E. Habib, 2016: Assessment and implications of NCEP Stage IV quantitative precipitation estimates for product intercomparisons. *Wea. Forecasting*, **31**, 371–394, https://doi.org/10.1175/WAF-D-14-00112.1.

Politis, D. N., and J. P. Romano, 1992: A circular block-resampling procedure for stationary data. *Exploring the Limits of Bootstrap*, R. LePage and L. Billard, Eds., John Wiley and Sons, 263–270.

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.

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.

Pyle, M. E., and K. F. Brill, 2019: A comparison of two methods for bias correcting precipitation skill scores. *Wea. Forecasting*, **34**, 3–13, https://doi.org/10.1175/WAF-D-18-0109.1.

Raynaud, L., and F. Bouttier, 2017: The impact of horizontal resolution and ensemble size for convective-scale probabilistic forecasts. *Quart. J. Roy. Meteor. Soc.*, **143**, 3037–3047, https://doi.org/10.1002/qj.3159.

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.

Rogers, E., and Coauthors, 2017: Mesoscale modeling development at the National Centers for Environmental Prediction: Version 4 of the NAM forecast system and scenarios for the evolution to a high-resolution ensemble forecast system. *28th Conf. on Weather and Forecasting/24th Conf. on Numerical Weather Prediction*, Seattle, WA, Amer. Meteor. Soc., 3B.4, https://ams.confex.com/ams/97Annual/webprogram/Paper311212.html.

Schumacher, R. S., 2015: Resolution dependence of initiation and upscale growth of deep convection in convection-allowing forecasts of the 31 May–1 June 2013 supercell and MCS. *Mon. Wea. Rev.*, **143**, 4331–4354, https://doi.org/10.1175/MWR-D-15-0179.1.

Schwartz, C. S., 2014: Reproducing the September 2013 record-breaking rainfall over the Colorado Front Range with high-resolution WRF forecasts. *Wea. Forecasting*, **29**, 393–402, https://doi.org/10.1175/WAF-D-13-00136.1.

——, 2016: Improving large-domain convection-allowing forecasts with high-resolution analyses and ensemble data assimilation. *Mon. Wea. Rev.*, **144**, 1777–1803, https://doi.org/10.1175/MWR-D-15-0286.1.

——, and Z. Liu, 2014: Convection-permitting forecasts initialized with continuously cycling limited-area 3DVAR, ensemble Kalman filter, and ''hybrid'' variational-ensemble data assimilation systems. *Mon. Wea. Rev.*, **142**, 716–738, https://doi.org/10.1175/MWR-D-13-00100.1.

——, 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.

——, 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, M. L. Weisman, R. A. Sobash, K. R. Fossell, K. W. Manning, and S. B. Trier, 2015: A real-time convection-allowing ensemble prediction system initialized by mesoscale ensemble Kalman filter analyses. *Wea. Forecasting*, **30**, 1158–1181, https://doi.org/10.1175/WAF-D-15-0013.1.

——, ——, 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.

Seity, Y., P. Brousseau, S. Malardel, G. Hello, P. Bénard, F. Bouttier, C. Lac, and V. Masson, 2011: The AROME-France convective-scale operational model. *Mon. Wea. Rev.*, **139**, 976–991, https://doi.org/10.1175/2010MWR3425.1.

Skamarock, W. C., and M. L. Weisman, 2009: The impact of positive-definite moisture transport on NWP precipitation forecasts. *Mon. Wea. Rev.*, **137**, 488–494, https://doi.org/10.1175/2008MWR2583.1.

——, 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.

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.

——, 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.

Stein, T. H., and Coauthors, 2019: An evaluation of clouds and precipitation in convection-permitting forecasts for South Africa. *Wea. Forecasting*, **34**, 233–254, https://doi.org/10.1175/WAF-D-18-0080.1.

Surcel, M., I. Zawadzki, and M. K. Yau, 2016: The case-to-case variability of the predictability of precipitation by a storm-scale ensemble forecasting system. *Mon. Wea. Rev.*, **144**, 193–212, https://doi.org/10.1175/MWR-D-15-0232.1.

——, ——, ——, M. Xue, and F. Kong, 2017: More on the scale dependence of the predictability of precipitation patterns: Extension to the 2009–13 CAPS Spring Experiment ensemble forecasts. *Mon. Wea. Rev.*, **145**, 3625–3646, https://doi.org/10.1175/MWR-D-16-0362.1.

Tang, Y., H. W. Lean, and J. Bornemann, 2013: The benefits of the Met Office variable resolution NWP model for forecasting convection. *Meteor. Appl.*, **20**, 417–426, https://doi.org/10.1002/met.1300.

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.

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.

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

Wang, X., D. F. Parrish, D. T. Kleist, and J. S. Whitaker, 2013: GSI 3DVAR-based ensemble–variational hybrid data assimilation for NCEP Global Forecast System: Single-resolution experiments. *Mon. Wea. Rev.*, **141**, 4098–4117, https://doi.org/10.1175/MWR-D-12-00141.1.

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.

——, 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.

Wilks, D. S., 1997: Resampling hypothesis tests for autocorrelated fields. *J. Climate*, **10**, 65–82, https://doi.org/10.1175/1520-0442(1997)010<0065:RHTFAF>2.0.CO;2.

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.

Xue, M., F. Kong, K. W. Thomas, J. Gao, Y. Wang, K. Brewster, and K. K. Droegemeier, 2013: Prediction of convective storms at convection-resolving 1-km resolution over continental United States with radar data assimilation: An example case of 26 May 2008 and precipitation forecasts from spring 2009. *Adv. Meteor.*, **2013**, 259052, https://doi.org/10.1155/2013/259052.

Zhang, F., M. Zhang, and J. Poterjoy, 2013: E3DVar: Coupling an ensemble Kalman filter with three-dimensional variational data assimilation in a limited-area weather prediction model and comparison to E4DVar. *Mon. Wea. Rev.*, **141**, 900–917, https://doi.org/10.1175/MWR-D-12-00075.1.

Zimmer, M., G. C. Craig, C. Keil, and H. Wernli, 2011: Classification of precipitation events with a convective response timescale and their forecasting characteristics. *Geophys. Res. Lett.*, **38**, L05802, https://doi.org/10.1029/2010GL046199.