A Study of Multiscale Initial Condition Perturbation Methods for Convection-Permitting Ensemble Forecasts

AARON JOHNSON

Cooperative Institute for Mesoscale Meteorological Studies, Norman, Oklahoma

XUGUANG WANG

School of Meteorology, University of Oklahoma, Norman, Oklahoma

(Manuscript received and in final form 14 February 2016)

ABSTRACT

The impacts of multiscale flow-dependent initial condition (IC) perturbations for storm-scale ensemble forecasts of midlatitude convection are investigated using perfect-model observing system simulation experiments. Several diverse cases are used to quantitatively and qualitatively understand the impacts of different IC perturbations on ensemble forecast skill. Scale dependence of the results is assessed by evaluating 2-h storm-scale reflectivity forecasts separately from hourly accumulated mesoscale precipitation forecasts.

Forecasts are initialized with different IC ensembles, including an ensemble of multiscale perturbations produced by a multiscale data assimilation system, mesoscale perturbations produced at a coarser resolution, and filtered multiscale perturbations. Mesoscale precipitation forecasts initialized with the multiscale perturbations are more skillful than the forecasts initialized with the mesoscale perturbations at several lead times. This multiscale advantage is due to greater consistency between the IC perturbations and IC uncertainty. This advantage also affects the short-term, smaller-scale forecasts. Reflectivity forecasts on very small scales and very short lead times are more skillful with the multiscale perturbations as a direct result of the smaller-scale IC perturbation energy. The small-scale IC perturbations also contribute to some improvements to the mesoscale precipitation forecasts after the ~5-h lead time. Altogether, these results suggest that the multiscale IC perturbations provided by ensemble data assimilation on the convection-permitting grid can improve storm-scale ensemble forecasts by improving the sampling of IC uncertainty, compared to downscaling of IC perturbations from a coarser-resolution ensemble.

1. Introduction

An open question for storm-scale ensemble forecast (SSEF) design is how to optimally perturb the initial conditions (ICs) so that the differences among ensemble members at the initial time adequately sample the analysis errors contributing to the subsequent forecast errors. Convective-scale precipitation forecasting is an inherently multiscale challenge because of the broad range of spatial scales impacting the initiation and evolution of convective systems (Lorenz 1969; Perkey and Maddox 1985; Zhang et al. 2007; Rotunno and Snyder 2008; Johnson et al. 2014; 2015). Small-scale errors in the initial state can grow upscale and contaminate even the larger scales of the forecast (e.g., Zhang et al. 2006, 2007). Such loss of deterministic predictability motivates the ensemble approach (Ehrendorfer 1997). Successful ensemble forecasts require all significant sources of forecast error to be sampled in the ensemble design (Toth and Kalnay 1997). While IC errors on multiple scales contribute to the forecast error, it is not clear how to optimally design corresponding multiscale IC perturbations for SSEFs that resolve features on synoptic to convective scales.

Similar to regional mesoscale ensembles [i.e., tens of kilometers grid spacing, e.g., Wang et al. (2014)], IC perturbations for convection-permitting (i.e., 1–4-km grid spacing) SSEFs are typically generated by either downscaling perturbations from a coarser-resolution ensemble without small-scale IC perturbations (e.g., Hohenegger et al. 2008; Xue et al. 2010; Zhang et al.

Corresponding author address: Dr. Aaron Johnson, Cooperative Institute for Mesoscale Meteorological Studies, 120 David L. Boren Blvd., Norman, OK 73072.
E-mail: ajohns14@ou.edu

DOI: 10.1175/MWR-D-16-0056.1

© 2016 American Meteorological Society

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

2010; Peralta et al. 2012; Schwartz and Liu 2014; Kühnlein et al. 2014), or generating multiscale IC perturbations directly on the forecast grid using methods such as cycled ensemble-based data assimilation (e.g., Vié et al. 2011; Snook et al. 2011; Yussouf et al. 2013; Harnisch and Keil 2015). Multiscale IC perturbations for regional mesoscale ensembles (i.e., grid spacing on the order of tens of kilometers) have been shown to be more effective than coarser IC perturbations downscaled from a global-scale ensemble (e.g., Wang et al. 2014). However, evaluation of the relative advantages of these IC perturbation methods for SSEFs (i.e., grid spacing on the order of 1–4 km) has been very limited (e.g., Kühnlein et al. 2014; Harnisch and Keil 2015). While it may be computationally expensive to generate an ensemble of IC perturbations at a convection-permitting resolution, IC perturbation methods appropriate for coarser-resolution ensembles may be particularly ill suited for SSEFs (Hohenegger and Schär 2007a,b). It remains unclear what the advantages of generating the IC perturbations at the full model resolution are for SSEFs.

Small-scale IC perturbations can have significant impacts on SSEF spread as a result of rapid propagation and upscale growth (Hohenegger et al. 2006; Hohenegger and Schär 2007a,b; Zhang et al. 2003, 2006; Leoncini et al. 2010; Johnson et al. 2014). However, initial studies have suggested that the added benefit of small-scale IC perturbations may be very limited when larger-scale perturbations are already present (Johnson et al. 2014; Kong et al. 2007). Durran and Gingrich (2014) have even suggested that explicitly added small-scale IC perturbations have no practical importance because of rapid downscale propagation of larger-scale perturbation energy. While Johnson et al. (2014) and Durran and Gingrich (2014) used random homogenous small-scale IC perturbations, the forecast sensitivity can depend on the spatial structure of the small-scale IC perturbations (Hohenegger and Schär 2007a; Johnson et al. 2014). Therefore, the impact of more realistic flow-dependent small-scale IC perturbations that sample the fastest-growing errors remains an open question in the context of multiscale IC perturbations for SSEFs.

The impacts of more realistic flow-dependent multiscale perturbations are a worthy topic of study because such perturbations are generated as a by-product of multiscale ensemble data assimilation. The impacts of such perturbations should therefore be included in any cost–benefit analysis of whether the computational expense of a multiscale ensemble data assimilation system is warranted. It is hypothesized that the flow-dependent multiscale IC perturbations will provide ensemble forecast benefits that were not seen with previous methods of generating multiscale IC perturbations for SSEFs where the flow-dependent nature of small-scale perturbations is neglected.

This study has three main goals aimed at better understanding the impacts of flow-dependent multiscale IC perturbations. The first goal of this study is to understand the impact of the multiscale IC perturbations generated on the convection-permitting grid as a by-product of ensemble data assimilation (hereafter MULTI) in comparison to larger-scale IC perturbations downscaled from a coarser-resolution ensemble (hereafter LARGE). The different IC perturbation methods include not only differences in resolution, but also differences on commonly resolved scales as a result of being generated on different model grids with different data assimilation configurations. Specifically, the MULTI perturbations are generated by assimilating radar observations and cycling on the convection-permitting grid, which allows for both upscale and downscale interactions between the mesoscales and convective scales. The second and third goals of the study are therefore to distinguish the impacts of the differences between MULTI and LARGE on commonly resolved scales and the presence of the smaller-scale IC perturbations in MULTI. Convection-permitting forecasts provide information that is useful for users interested in applications ranging from ~1-h predictions of individual storms and severe weather events (e.g., Stensrud et al. 2009; Yussouf et al. 2013) to mesoscale quantitative precipitation forecasting (e.g., Clark et al. 2009, 2012; Duc et al. 2013). The impacts of the IC perturbations are therefore evaluated in terms of both hourly accumulated precipitation and short-term (2 h) reflectivity forecasts.

While model and physics diversity are also an important part of the ensemble design (e.g., Clark et al. 2008; Johnson et al. 2011a,b), this study focuses only on the IC perturbations. Therefore, perfect-model observing system simulation experiments (OSSEs) are used to isolate the IC error from the model and physics errors, allowing a focused study of multiscale IC perturbation methods. In an OSSE, the same model that is used for the experiment is also used to simulate a proxy for the real atmosphere, leading to the assumption that the effects of model error on the experiment results are negligible. Therefore, the validity and limitations of this assumption are also discussed further in section 4. A total of 11 diverse cases of midlatitude convection in the central United States are used to systematically address the three scientific goals mentioned above. A case study is also used to qualitatively understand the features seen in the systematic evaluation. Section 2 describes the

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

OSSE design, model, and experiment configurations; the IC perturbation methods; and the verification methods. Results are presented in section 3 while section 4 contains a summary and discussion.

2. Methods and experiments

a. GSI-based EnKF and model configuration

The Gridpoint Statistical Interpolation (GSI)-based ensemble Kalman filter (EnKF) data assimilation (DA) system that was implemented for the Global Forecast System (GFS) model at the National Centers for Environmental Prediction (NCEP) as part of the hybrid DA system (Whitaker et al. 2008; Wang et al. 2013; Wang and Lei 2014; Mahajan et al. 2016) has also been extended to directly assimilate radar reflectivity and radial velocity observations for multiscale DA for convective-scale weather forecasts (Johnson et al. 2015). This newly extended GSI-EnKF system was shown to accurately analyze features across multiple scales for convection-permitting forecasts of midlatitude convection (Johnson et al. 2015). Given the multiscale emphasis of this study, the multiscale GSI-based EnKF system is adopted, following the configuration of Johnson et al. (2015).

The GSI-based EnKF system provides an analysis ensemble on both 12- and 4-km grids. The analysis ensembles are used to construct IC perturbations for the experiments in this study. While greater detail can be found in Johnson et al. (2015), a brief overview of the configuration of the multiscale DA and forecast system follows. The Weather Research and Forecasting (WRF) Advanced Research WRF (ARW) Model version 3.2 (Skamarock and Klemp 2007) is used with an inner convection-permitting domain of 346 × 277 grid points at 4-km grid spacing, nested within an outer convection-parameterizing domain of 326 × 259 grid points at 12-km grid spacing (Fig. 1a). Simulated (see section 2b) synoptic and mesoscale observations from surface station; surface mesonet; Aircraft Communication, Addressing, and Reporting System (ACARS); NOAA wind profiler platforms; and radiosondes are assimilated on the outer domain every 3 h during a 24-h DA period (e.g., Fig. 1a). The outer domain analyses provide initial and lateral boundary conditions (LBCs) for the inner domain DA ensemble. Simulated storm-scale NEXRAD radar observations are then assimilated on the inner domain every 5 min during a 3-h period preceding the final analysis time (Fig. 1b). The different cycling intervals for the mesoscale and storm-scale DA are chosen based on the different approximate error growth rates on the different spatial scales. This configuration is therefore expected to result in IC perturbations that are flow dependent and fast growing on multiple scales (Peña and Kalnay 2004).

b. OSSE framework and selection of cases

In the OSSE framework, a model simulation referred to as the nature run represents the "true" atmosphere, the state and dynamics of which are perfectly known. In this study, the nature run is initialized from the NCEP GFS analysis at 0000 UTC (24 h before the analysis time for each case) at 4-km grid spacing over the entire outer domain (Fig. 1a). Observations of wind, temperature, water vapor, and sea level pressure are then simulated by sampling the nature run at observation locations representative of the actual observation networks (e.g., Fig. 1), with representative observation error characteristics. The simulated observations are then assimilated into the experiment forecasts in order to try to recover the true state of the nature run, using the GSI-based multiscale DA system.

An advantage of the OSSE framework is that the observations are perfectly known at the model grid points. For this study of IC perturbations, the OSSE framework has the additional advantage of eliminating model and physics uncertainty as a source of forecast error by using identical model configurations for the nature run and experiment forecasts (i.e., a "perfect model" OSSE). The outer domain analyses do contain model error arising from the coarser resolution and convection parameterization. However, such errors only enter the convection-permitting forecasts through the ICs provided by the inner domain DA and the LBCs from the outer domain.

The same 10 cases used in Johnson et al. (2015) are also adopted for this OSSE study, with the addition of a forecast initialized at 2100 UTC 19 May 2010 for a total of 11 cases. Like the real data cases (Johnson et al. 2015), the nature run simulations of these cases include a variety of forcing mechanisms and levels of convective organization ranging from disorganized cellular convection to supercells to long-lived MCSs. Another advantage of the OSSE framework is that the nature run provides the exact truth values for verification of all forecast variables on the same grid as the forecast variables. For the hourly accumulated precipitation forecasts, rectangular verification domains for each case are chosen to include the areas of active convection at all lead times while excluding large areas where convection is neither observed nor forecast. For the 2-h lead time reflectivity forecasts, smaller rectangular verification domains are used to encompass each subjectively identified mesoscale area of organized convection during the first two forecast hours. Some of the forecast cases contained multiple areas of mesoscale organized

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

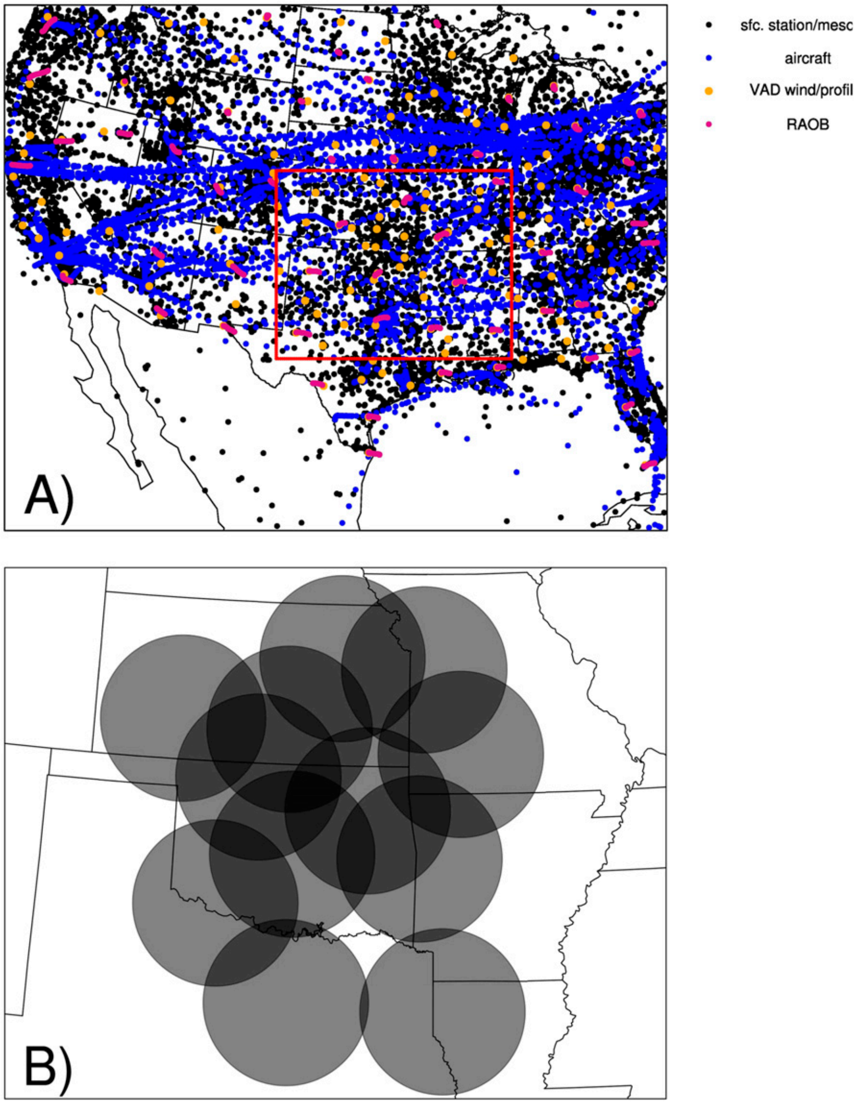

FIG. 1. (a) Outer 12-km domain with the location of the inner 4-km domain enclosed in the red rectangle and a representative distribution of the observation types assimilated on the outer domain. (b) Enlarged inner domain with the locations and coverage of the assimilated NEXRAD radars shaded. Adapted from Johnson et al. (2015).

convection, resulting in a total of 18 unique verification domains for the reflectivity verification. Since different MCSs on the same case occur within the same larger-scale environment, such MCSs are not treated as independent samples for statistical significance testing. Since the verification domain for each case was a slightly different size, the main results were also recalculated using same-sized verification domains and the conclusions were not changed (not shown).

The actual evolution of the 20 May 2010 case study, including upscale growth of initially cellular convection into a long-lived MCS in central Oklahoma (OK), has been described in Johnson et al. (2014, 2015). (The nature run for this case also shows similar upscale growth of convection into a long-lived MCS, as seen in the observation contours in Fig. 6.) This case is chosen for an initial investigation into multiscale IC perturbation methods because of the multiple scales of motion influencing such upscale growing MCSs (e.g., Perkey and Maddox 1985), the sensitivity of this case to IC errors on multiple scales in non-OSSE experiments (Johnson et al. 2014), and the similarity between the nature run and actual evolution for this case. This case is also selected to qualitatively demonstrate the impact of the IC perturbation methods because it is found to be representative of the systematic results in this study.

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

Experiment forecasts are initialized at 0000 UTC 20 May 2010, about half-way through the upscale growth of the MCS, as well as 2100 UTC 19 May 2010 before the storms begin to grow upscale.

c. IC perturbation methods

In all experiments, the ensemble forecasts have the same mean analysis, provided by the ensemble mean analysis of the multiscale GSI-based EnKF DA system. The only difference among the experiments is the IC perturbations added to the ensemble mean to generate the initial ensemble. The MULTI IC perturbations are obtained by directly using the inner domain multiscale analyses to initialize the ensemble forecasts. The LARGE IC perturbations are obtained by adding to the inner domain ensemble mean analysis the difference between each outer domain ensemble member and the outer domain ensemble mean (both interpolated to the inner domain). A third ensemble, MULTI48, is constructed by filtering[1] wavelengths less than 48 km from each MULTI perturbation before adding it back to the inner domain ensemble mean. Since MULTI48 is the same as MULTI except for the absence of perturbations on scales not resolved by LARGE, comparison of MULTI48 with LARGE allows goal two (see section 1) to be investigated while comparison of MULTI48 with MULTI allows goal three (see section 1) to be investigated. For simplicity, wavelengths less than 48 km are therefore referred to as "small scale" IC perturbations in this study while the larger scales are referred to as "mesoscale" IC perturbations in this study. Although there is not such a sharp cutoff in the scales resolved by the LARGE IC perturbations, the difference in perturbation energy between MULTI and LARGE is particularly pronounced at wavelengths smaller than approximately ~50 km (Fig. 2a), motivating the choice of 48 km (i.e., 12 grid points) to separate the small scales and mesoscales in the IC perturbations.

d. Verification methods

SSEFs have proven useful for users interested in convective precipitation forecast applications on space and time scales ranging from very short-term warn-on-forecast applications (e.g., Stensrud et al. 2009; Yussouf et al. 2013) to mesoscale quantitative precipitation forecasting (e.g., Clark et al. 2009, 2012; Duc et al. 2013). Forecasts on such different scales may show different sensitivities to the multiscale IC perturbation methods. To provide a robust understanding of the impacts of IC perturbation methods, the convection forecasts are here evaluated in terms of both instantaneous reflectivity during the first two forecast hours and mesoscale hourly precipitation accumulation out to 9 h. Reflectivity results are shown using model level 12 (~750 hPa). Reflectivity at model level 5 (~900 hPa) was also evaluated and showed very similar results (not shown).

The forecasts are objectively verified using the Brier skill score (BSS; Brier 1950; Murphy 1973; Wilks 2006) of neighborhood ensemble probability (NEP; Theis et al. 2005; Schwartz et al. 2010). The NEP is defined as the percentage of grid points from all ensemble member forecasts within a search radius that exceed the threshold being forecast. The use of the NEP reduces the sensitivity to errors on scales smaller than the search radius (Roberts and Lean 2008). A radius of 48 km is chosen for the mesoscale hourly accumulated precipitation forecasts in order to eliminate the impact of smaller-scale and less predictable details (Johnson and Wang 2012). The observation (hourly accumulated precipitation in the nature run simulation) is not converted to a neighborhood probability, following Johnson and Wang (2012). The reflectivity forecasts are evaluated across a range of different spatial scales (i.e., radii less than 48 km) and verification thresholds (Stratman et al. 2013). The BSS provides a simple way to verify the ensemble probabilistic forecasts that is sensitive to both the reliability and resolution of the forecasts (Murphy 1973). For both precipitation and reflectivity, the reference forecast for calculating BSS is the domain-averaged observed event frequency, as also used in Johnson and Wang (2012). Differences in skill among experiments are not sensitive to the choice of reference forecast since it is the same for all experiments. In addition to the objective verification, subjective verification is also conducted to qualitatively understand the physical processes behind the objective skill metrics for the selected, representative case.

The statistical significance of differences in BSS is determined using permutation resampling of the 11 cases (Hamill 1999). For reflectivity forecasts with multiple MCSs on the same day, results are first aggregated for that day and treated as a single sample since the different MCSs may not be statistically independent. For the hourly accumulated precipitation forecasts, statistical significance is plotted at the 80% confidence level. The relatively low confidence level is chosen because an 11-sample dataset is rather small to expect very high levels of confidence. While this choice does leave a 20% chance of a "significant" result occurring due to random chance, it allows the more robust results to be distinguished from the less robust

[1] The filtering consists of truncation of wavelengths below 48 km in the two-dimensional discrete cosine transform (Denis et al. 2002) of the IC perturbation field.

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

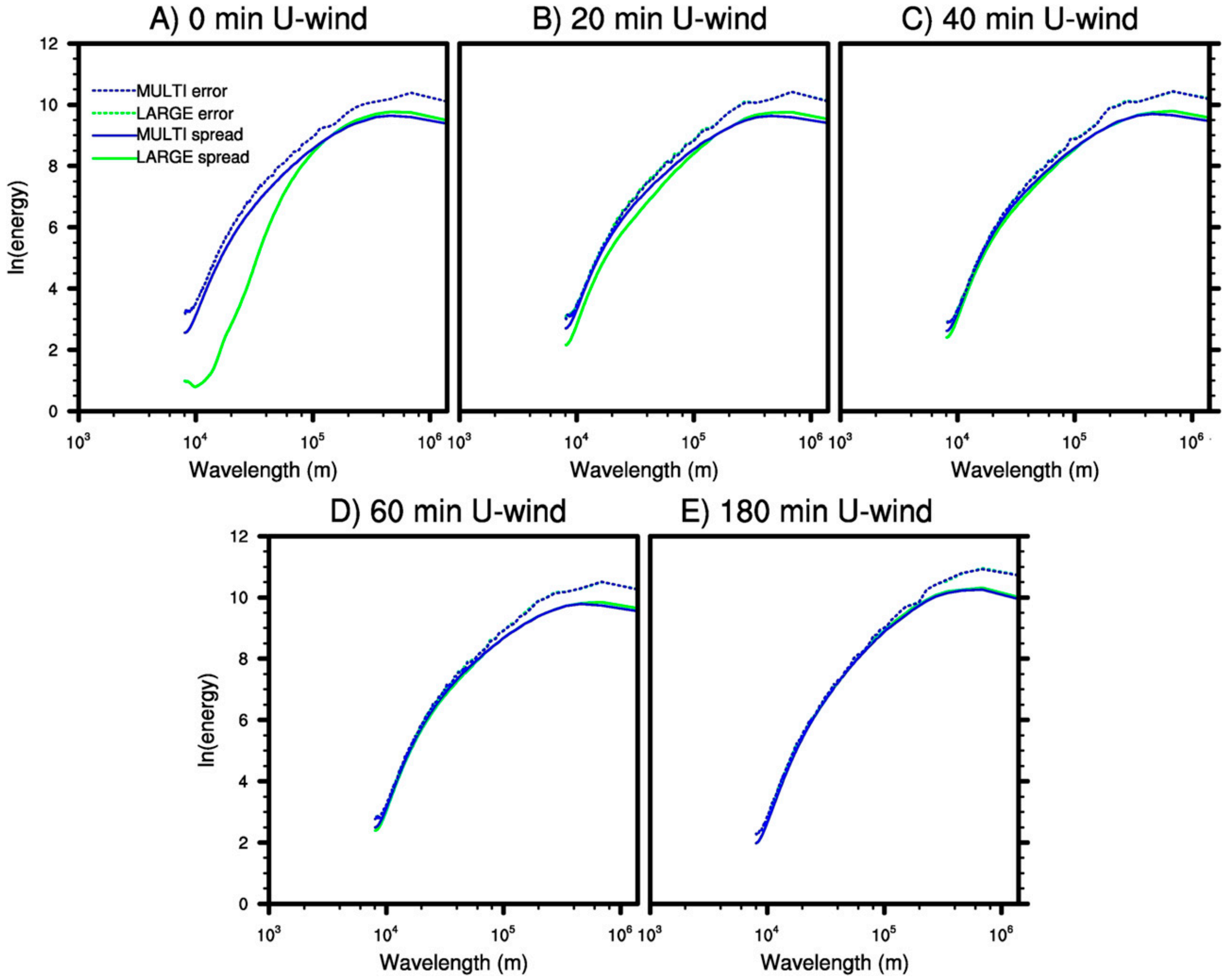

FIG. 2. Fourier spectra decomposition of ensemble perturbations (ensemble member minus ensemble mean, averaged over all members; solid) and ensemble mean error (dashed) for the *u* component of wind at model level 5 (~900 hPa) averaged over all 11 cases at (a) the analysis time, (b) 20-min forecast time, (c) 40-min forecast time, (d) 60-min forecast time, and (e) 180-min forecast time.

results. For the reflectivity forecasts, statistical significance is plotted at the 90% confidence level because the impacts of the IC perturbation methods on reflectivity forecast skill are more consistent from case to case, allowing for greater levels of statistical significance to be established.

3. Results

a. Nonprecipitation forecasts

Since the nonprecipitation variables are the directly perturbed IC variables, results for wind, temperature, and water vapor are considered first. One-dimensional detrended Fourier spectra for these variables are calculated along east–west grid lines, and averaged over all possible such grid lines (Skamarock 2004). Ensemble mean error is also calculated, using the nature run as truth (i.e., not the simulated observations, which contain observation error). The spectra of the ensemble mean error and the ensemble member perturbations, averaged over the 40 ensemble members, are compared for the *u* component of wind at model level 5 (~900 hPa) in Fig. 2. Results for this variable are similar to wind at model level 12 (~750 hPa) as well as temperature and water vapor (not shown). The ensemble mean error spectra are very similar for the MULTI and LARGE ensembles at the lead times shown in Fig. 2, with the green and blue dashed lines nearly on top of each other. The LARGE ensemble perturbations are markedly underdispersive, compared to the ensemble mean error, at scales less than ~50 km at the initial time (Fig. 2a). The lack of initial small-scale spread is a result of the coarser resolution of the outer domain ensemble used to generate the LARGE

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

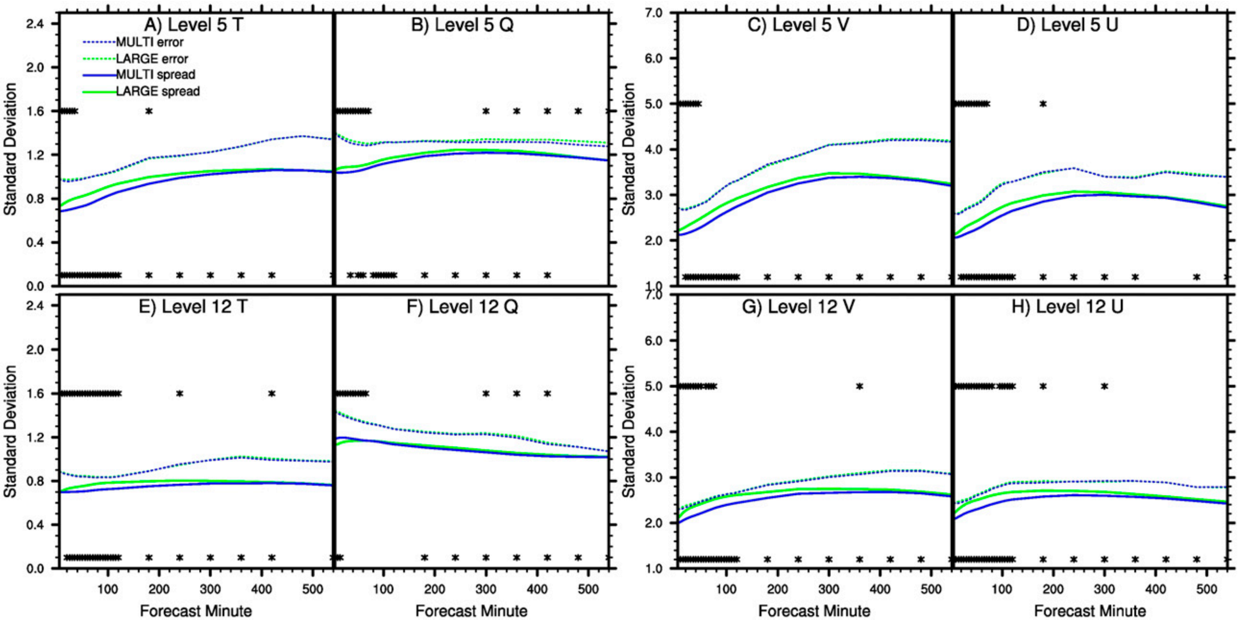

FIG. 3. Ensemble spread (i.e., standard deviation) and ensemble mean RMSE as a function of forecast lead time averaged over all cases at model level 5 (~900 hPa) for (a) temperature (K), (b) water vapor mixing ratio (g kg^{-1}), (c) v-wind component (m s^{-1}), and (d) u-wind component (m s^{-1}). (e)–(h) As in (a)–(d), but at model level 12 (~750 hPa). Statistical significance at the 90% confidence level is indicated by asterisks above (below) the curves for a significant difference between the two error (spread) curves.

perturbations. On average, the MULTI and LARGE spread at small scales are nearly indistinguishable after ~40–60 min (Fig. 2).

Although the small scales are initially very underdispersive for LARGE, downscale energy propagation results in rapid perturbation growth on such scales, consistent with the results of Durran and Gingrich (2014). The small-scale energy for LARGE catches up to that for MULTI within about an hour (Fig. 2). This confirms that explicitly including small-scale IC perturbations has little impact on the ensemble spread of the directly perturbed variables on such scales for lead times beyond ~1 h. However, it is not clear what impacts the small-scale IC perturbations during the first hour have on the convective precipitation forecasts both during and after the first hour and on larger scales.

The ensemble spread (i.e., standard deviation) of wind, temperature, and moisture at model levels 5 and 12 (~900 and 750 hPa, respectively) are also evaluated and compared to the ensemble mean root-mean-square error (RMSE) in Fig. 3. Most variables and lead times show much less ensemble spread than ensemble mean error for both MULTI and LARGE (Fig. 3). The systematic underdispersion cannot be attributed to insufficient sampling of model errors in the ensemble design because of the perfect-model OSSE framework. It also likely is not attributable to the LBC perturbations, generated on the outer domain, which does contain model error, since the underdispersion is present from the beginning of the forecasts. However, the LBC perturbations may contribute to limiting the spread growth after ~5–6 h (Fig. 3). The systematic results here are representative of the results for the 20 May case study, and therefore not shown again for the case study.

b. *Convective precipitation forecasts*

The following subsections evaluate the differences between MULTI and LARGE, MULTI48 and LARGE, and MULTI and MULTI48 in order to address the three goals of this study (section 1).

1) IMPACT OF OVERALL IC PERTURBATION METHOD (MULTI VS LARGE)

(i) Mesoscale hourly accumulated precipitation

The first goal of this study is to understand the overall systematic impacts of the flow-dependent multiscale IC perturbations by comparing MULTI and LARGE. For the mesoscale precipitation forecasts, MULTI is more skillful than LARGE during the first hour and, for the 2.54 and 6.35 mm h^{-1} thresholds, after 4 h (Figs. 4a–c). MULTI is slightly less skillful than LARGE at 8–9 h for the 12.7 mm h^{-1} threshold but this difference is not statistically significant (Figs. 4a–c). The MULTI skill advantages are significant at the 1-h lead time for all

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

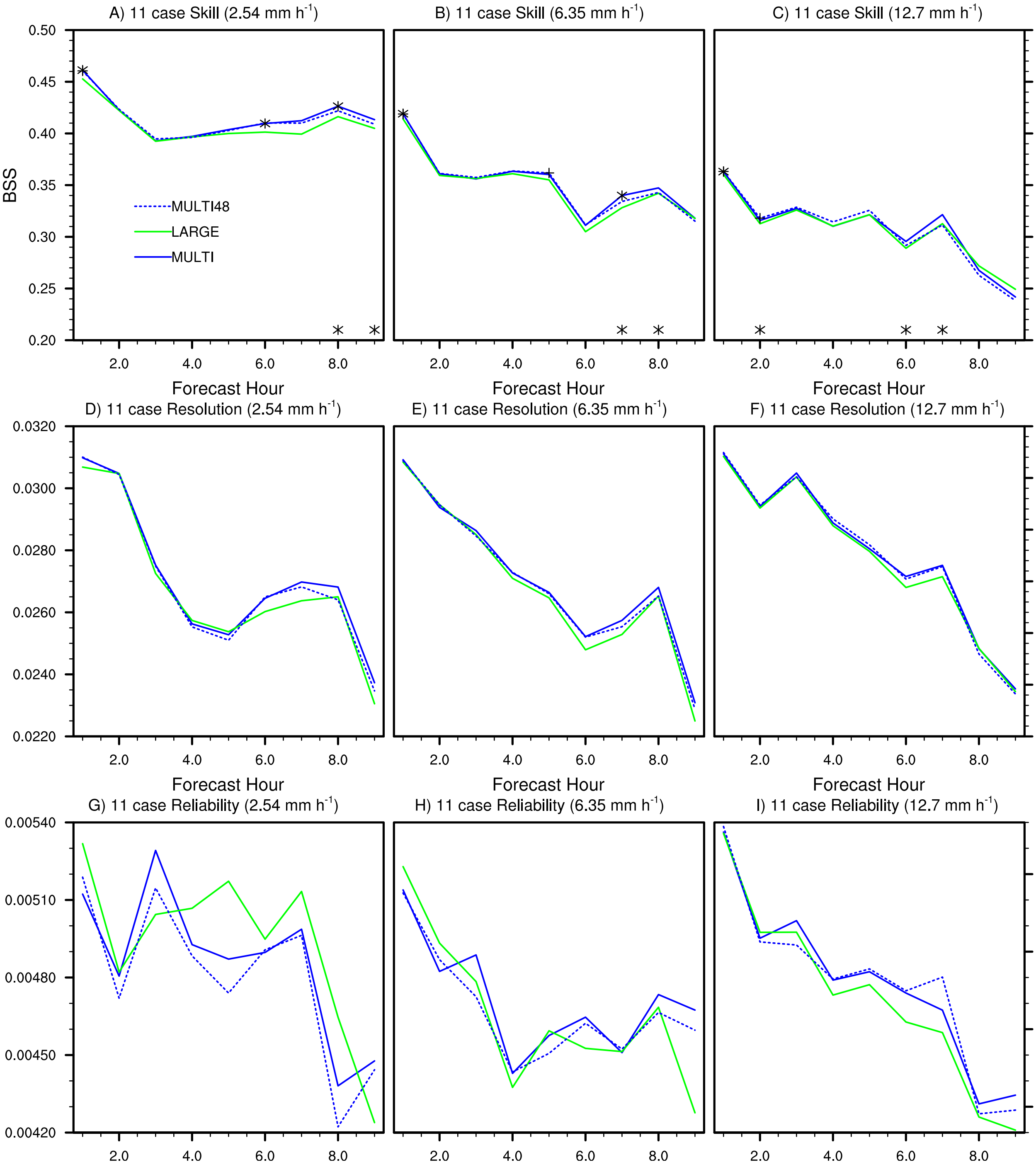

FIG. 4. Brier skill score (BSS) of the neighborhood ensemble probability (NEP) forecasts from all 11 cases for hourly accumulated precipitation thresholds of (a) 2.54, (b) 6.35, and (c) 12.7 mm h^{-1}. Statistical significance is plotted at the 80% confidence level, with significant differences between MULTI and LARGE, MULTI48 and LARGE, or MULTI and MULTI48 indicated by asterisks on the MULTI line, plus signs on the MULTI48 line, or asterisks along the horizontal axis, respectively. Also shown are the resolution component of the Brier score at the (d) 2.54, (e) 6.35, and (f) 12.7 mm h^{-1} thresholds and the reliability component of the Brier score at the (g) 2.54, (h) 6.35, and (i) 12.7 mm h^{-1} thresholds.

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

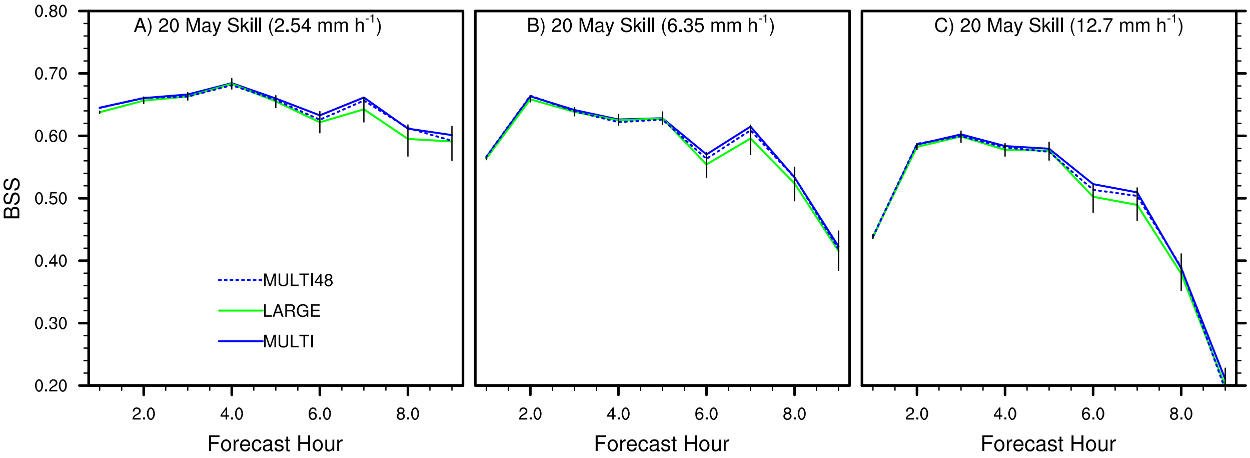

FIG. 5. As in Figs. 4a–c, but for the 20 May 2010 case and instead of a statistical significance test, confidence intervals are placed on the LARGE skill line. The 90% confidence intervals are calculated using 5000 bootstrap resamples with replacement of the 40 ensemble members for each BSS value. The purpose is to estimate the uncertainty in the verification statistic resulting from sampling errors due to the finite ensemble size.

thresholds, the 7-h lead for the 6.35 mm h^{-1} threshold, and the 6- and 8-h lead times for the 2.54 mm h^{-1} threshold. The 7- and 9-h lead times are nearly significant at the 2.54 mm h^{-1} threshold, with p values of 0.2024 and 0.236, respectively (not shown). The differences between MULTI and LARGE BSS correspond more closely to the differences between MULTI and LARGE resolution than to the differences in reliability (Figs. 4d–i).[2] The reliability of MULTI is actually worse (i.e., larger) than that of LARGE at the higher threshold (Fig. 4i), unlike the BSS and resolution that are generally better for MULTI than LARGE at this threshold before the 8-h lead time (Figs. 4c,f). The correspondence between the resolution differences and the BSS differences suggest that the same impact of the different IC perturbations cannot be achieved by simple calibration of the forecasts to improve their reliability.

The systematic results are generally representative of the 20 May case study since at most lead times MULTI is more skillful than LARGE for the case study (Fig. 5). Strictly speaking, given only one realization (i.e., one case) two nonzero/non-100% probabilities should not be compared as better or worse. However, subjective evaluation of this case study provides physical understanding of the causes of the more systematic differences in forecast skill noted above. The LARGE probabilistic forecast subjectively corresponds well to the "observed" (i.e., nature run) precipitation (Fig. 6). However, there are subtle errors such as a slight westward displacement of the axis of maximum NEP, relative to the observed MCS in the nature run, at the southern end of the MCS at later lead times (Fig. 6), similar to the forecast in the real data case in Johnson et al. (2015). The LARGE ensemble also predicts some spurious cells in the MCS cold pool resulting in nonzero probability northwest of the observed MCS during the first ~2 h (Fig. 6a; blue circle). The figures discussed below are plotted as differences from the LARGE (or MULTI48) NEP in order to emphasize such subtle forecast differences.

The generally greater skill for MULTI than LARGE on 20 May (Fig. 5) is consistent with subjective evaluation (Figs. 7a–i). Initially, MULTI shows reduced probability, compared to LARGE, in the cold pool region northwest of the MCS and increased probability farther east. The MULTI advantage of reducing the NEP in the cold pool region persists for ~3–4 h (Figs. 7a–d). Starting at ~0500 UTC, MULTI has higher probability along the eastern edge of the MCS and another area of reduced probability west of the southern end of the MCS (Figs. 7e–i). The westward displacement of the maximum NEP at later lead times for LARGE at the southern end of the MCS is thus partly corrected in the MULTI forecast.

(ii) *Storm-scale reflectivity*

The reflectivity forecasts, verified on smaller scales than the accumulated precipitation forecasts, are not considered beyond the 2-h forecast range because of the intrinsic lack of predictability at longer lead times for storm-scale features (Cintineo and Stensrud 2013). For the storm-scale reflectivity forecasts, MULTI is

[2] Note that a larger resolution component or a smaller reliability component both contribute to a smaller BS, and therefore a larger BSS (Brier 1950; Murphy 1973; Wilks 2006).

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

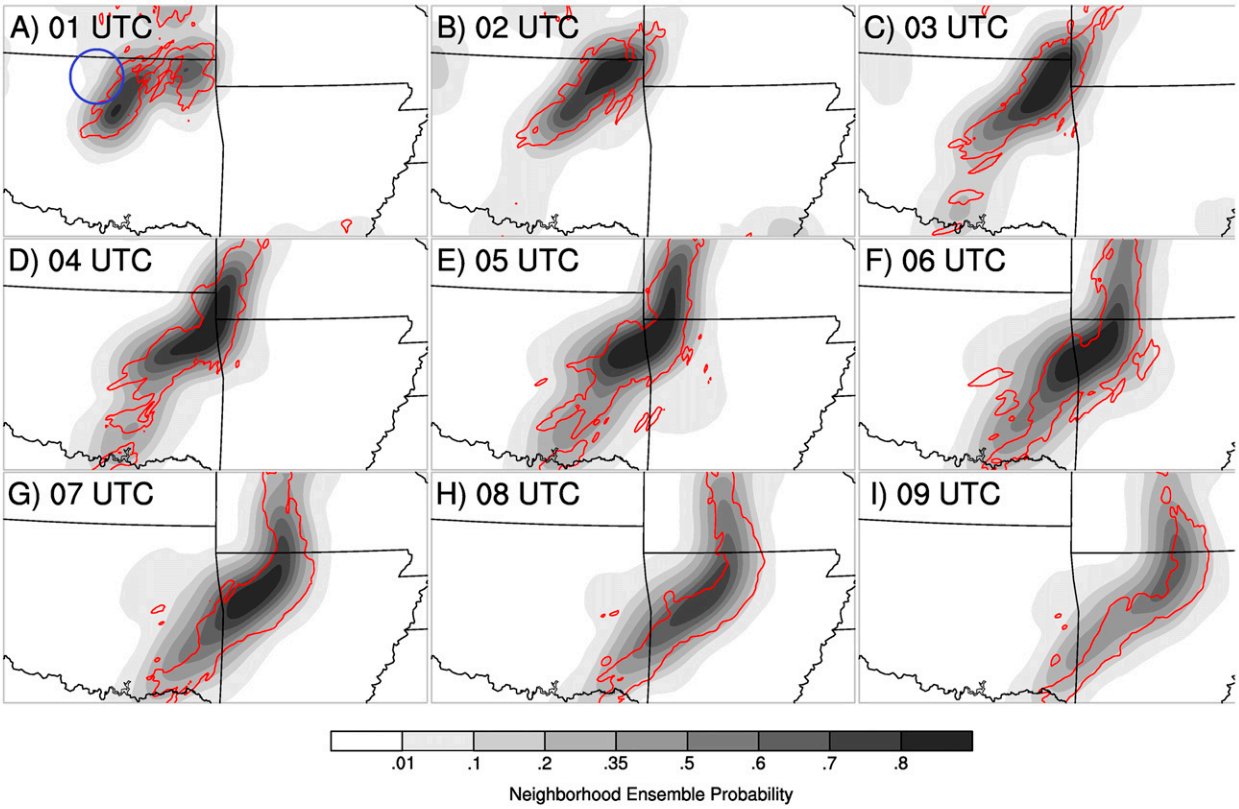

FIG. 6. (a)–(i) NEP (shaded) and observation contour (red line) for the LARGE ensemble forecast of hourly accumulated precipitation exceeding 6.35 mm h^{-1}, initialized at 0000 UTC 20 May 2010. The blue circle in (a) highlights the subtle area of nonzero forecast probability that is reduced in Fig. 7a. The corresponding observation of hourly accumulated precipitation is contoured in black at the same threshold.

again more skillful than LARGE where there are statistically significant differences (Fig. 8). The statistically significant MULTI advantages last for about 65 min at the lower thresholds (e.g., 20–25 dBZ; Fig. 8) and about 45 min at the higher thresholds (e.g., 40 dBZ; Fig. 8). Like the hourly accumulated precipitation forecasts, the differences in reflectivity forecast skill also correspond to the differences in the resolution component, rather than the reliability component (Figs. 9 and 10). Whereas the resolution component of the Brier score is generally better (i.e., larger) for MULTI than LARGE (Fig. 9), the reliability component of the Brier score is generally worse (i.e., larger) for MULTI than LARGE after ~30–45 min (Fig. 10). Since the differences between MULTI and LARGE include both smaller-scale perturbations in MULTI and different methods of generating the mesoscale perturbations, the impacts of these two factors are distinguished in the following subsections.

The reflectivity forecast skill differences on the 20 May case study are again representative of the systematic results (Fig. 11). For the case study, MULTI is generally the more skillful ensemble where the BSS difference exceeds a magnitude of 0.01 (color shading in Fig. 11). The MULTI advantage is most pronounced at ~30–40 min (Figs. 11f–h), while slight LARGE advantages begin to appear during the last ~20 min (Figs. 11s,t).

Subjectively, there are two competing factors that qualitatively explain the differences between MULTI and LARGE reflectivity forecast skill, as illustrated with the representative forecasts for the 30-dBZ threshold (Figs. 12a–h). First, MULTI provides a sharper forecast of the MCS with greater resolution than LARGE during the first forecast hour, since MULTI has lower probability outside of the observed MCS and higher probability within the observed MCS (Figs. 12a–f). In particular, there is reduced MULTI probability outside and to the west of the observed MCS and a corresponding increase in MULTI probability inside the northern end of the observed MCS contour (Figs. 12a–f). This difference is most pronounced at ~35–45 min, consistent with the greatest MULTI skill advantage in Fig. 11. Second, MULTI

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

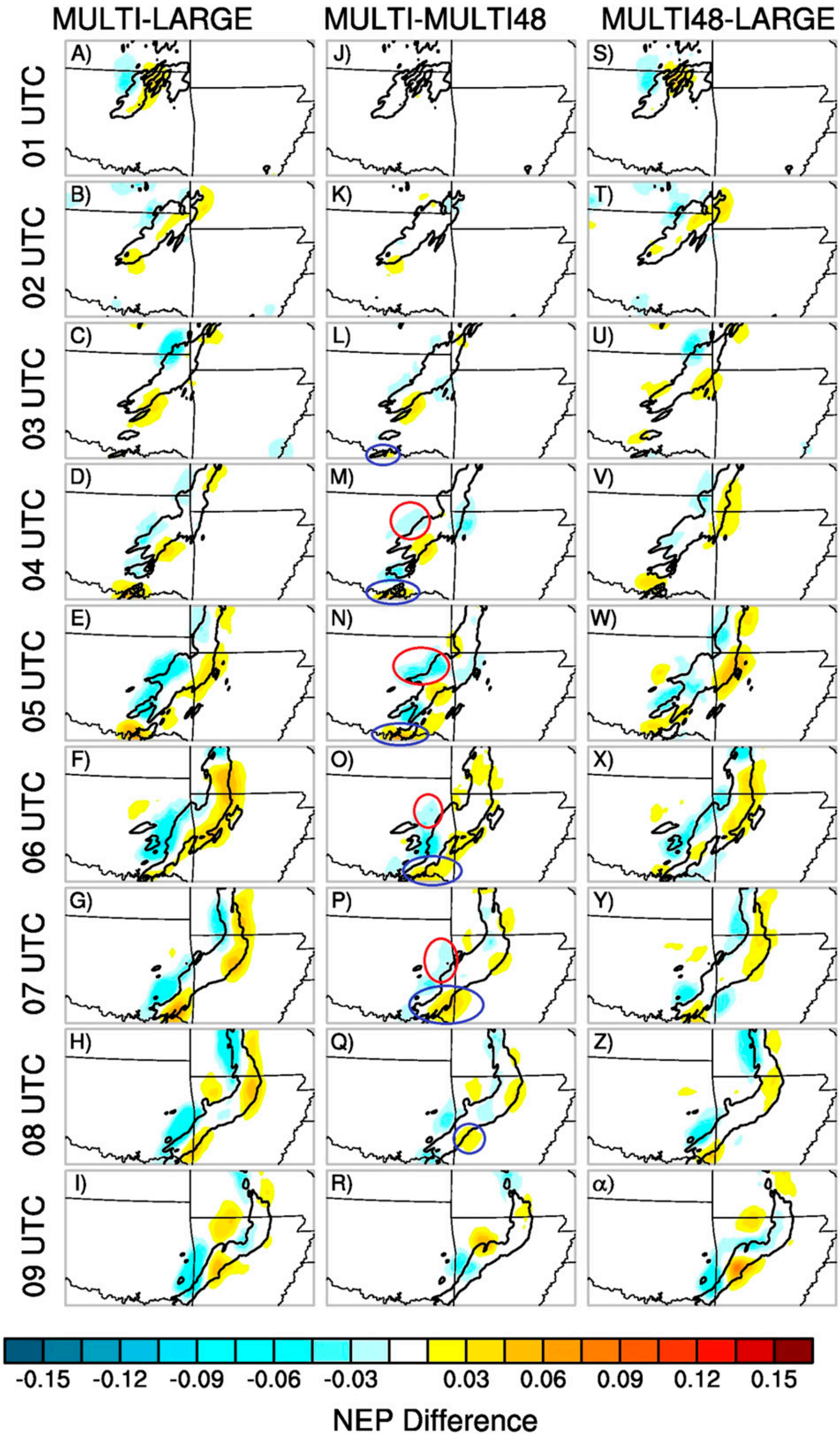

FIG. 7. Difference in NEP between (left) MULTI and LARGE, (middle) MULTI and MULTI48, and (right) MULTI48 and LARGE, for hourly accumulated precipitation forecasts initialized at 0000 UTC 20 May 2010, for the 6.35 mm h^{-1} threshold. Blue and red circles highlight areas referred to in the text.

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

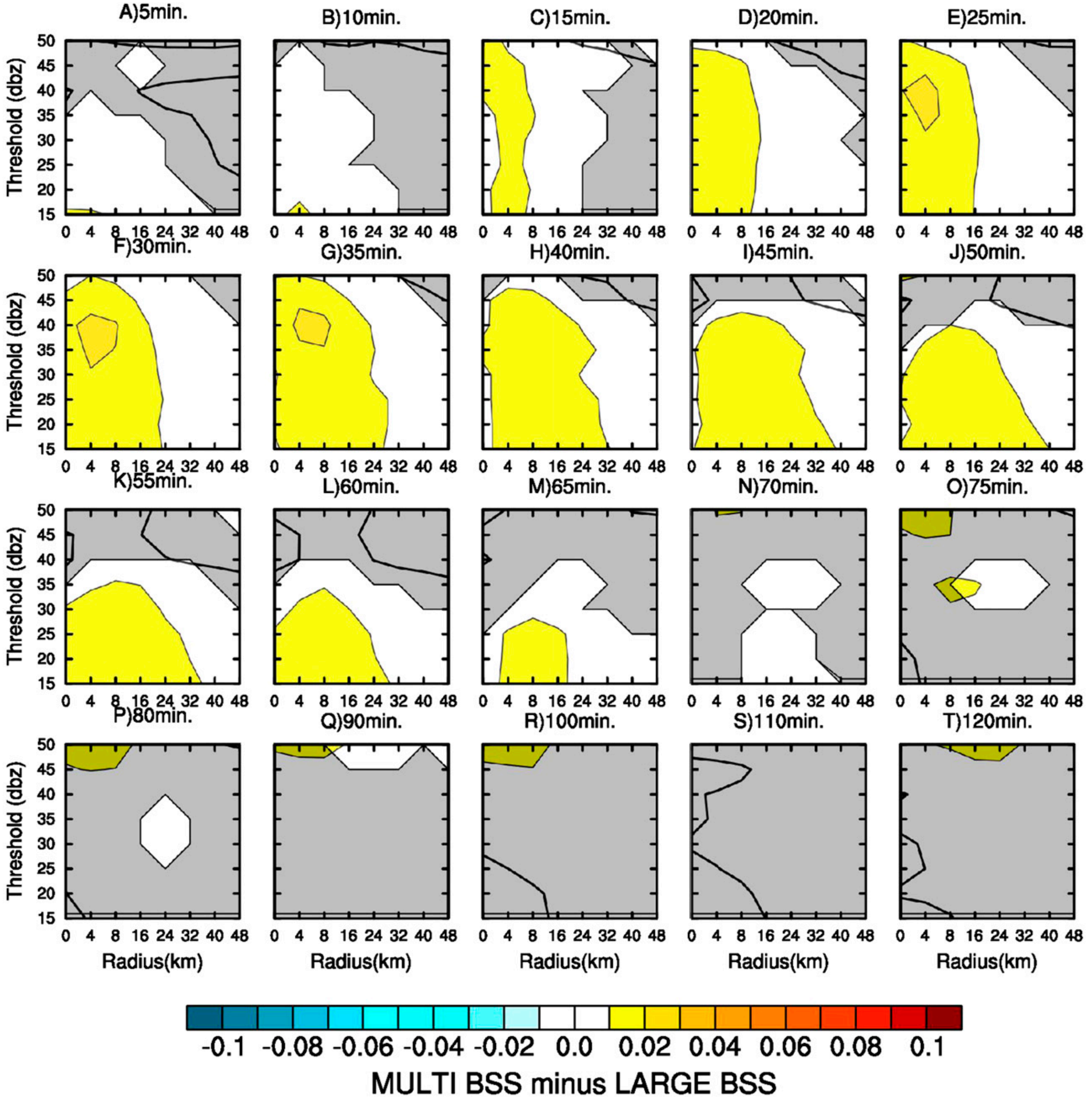

FIG. 8. Difference in BSS between the MULTI and LARGE ensembles, averaged over all 18 MCS cases, for reflectivity at model level 12 at 5-min intervals during the first 80 min and at 10-min intervals between 80 and 120 min. The vertical axis on each panel is the reflectivity threshold (dBZ) and the horizontal axis is the neighborhood radius (km). Values that are not statistically significant at the 90% level are covered by shading. The unshaded values are statistically significant at the 90% level.

enhances the forecast probability outside of the observation contour at the southern and eastern edges of the observed MCS, negatively impacting the forecast skill. This difference becomes more pronounced at the later lead times, explaining the decreasing MULTI skill advantage and eventual slight LARGE skill advantage in Fig. 11. The causes of these qualitative differences are discussed further in the following subsections.

2) IMPACT OF MESOSCALE COMPONENT OF IC PERTURBATION METHOD (MULTI48 VS LARGE)

(i) *Mesoscale hourly accumulated precipitation*

The second goal of this study is to understand the systematic impacts of the differences between MULTI and LARGE on commonly resolved scales (i.e., mesoscales),

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

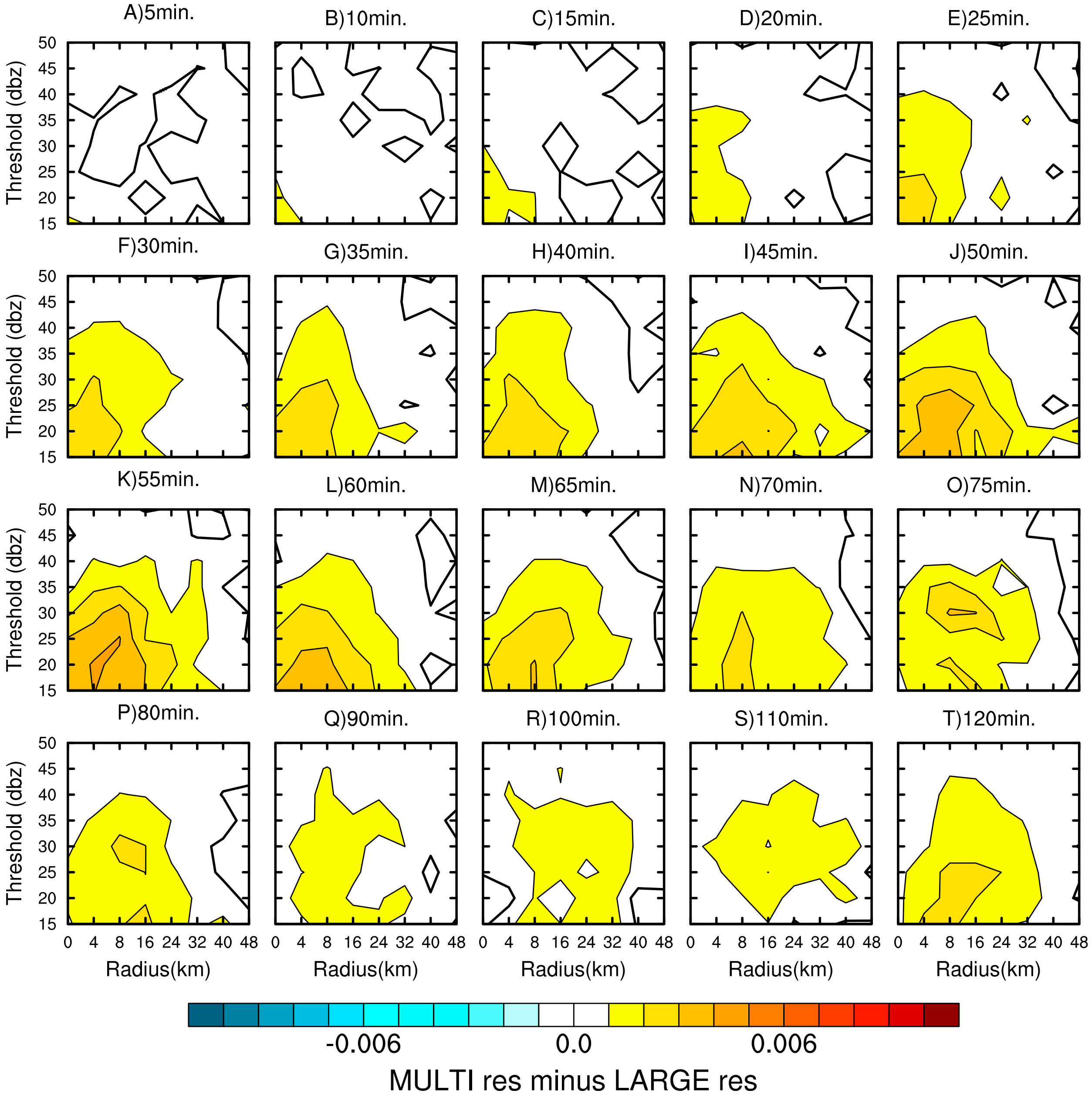

FIG. 9. As in Fig. 8, but for the resolution component of the Brier score. Significance tests were not repeated separately for the resolution and reliability components.

as opposed to the smaller scales resolved only by MULTI. MULTI48 is therefore compared to LARGE in order to focus only on the mesoscale IC perturbations (Fig. 4). The differences in mesoscale precipitation forecast skill between MULTI48 and LARGE are similar in many ways to the differences between MULTI and LARGE. Specifically, MULTI48 is more skillful than LARGE after ~4 h, except for the last few hours at the 12.7 mm h^{-1} threshold and the last couple of hours at the 6.35 mm h^{-1} threshold. Unlike the differences between MULTI and LARGE, the differences between MULTI48 and LARGE are generally not statistically significant, except at the 1-h lead time (Fig. 4). This result shows that the small-scale IC perturbations, omitted from MULTI48, also play an important role as further discussed in section 3b(3). Also, as in the differences between MULTI and LARGE, the differences in BSS between MULTI48 and LARGE correspond to differences in resolution, not just reliability (Fig. 4).

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

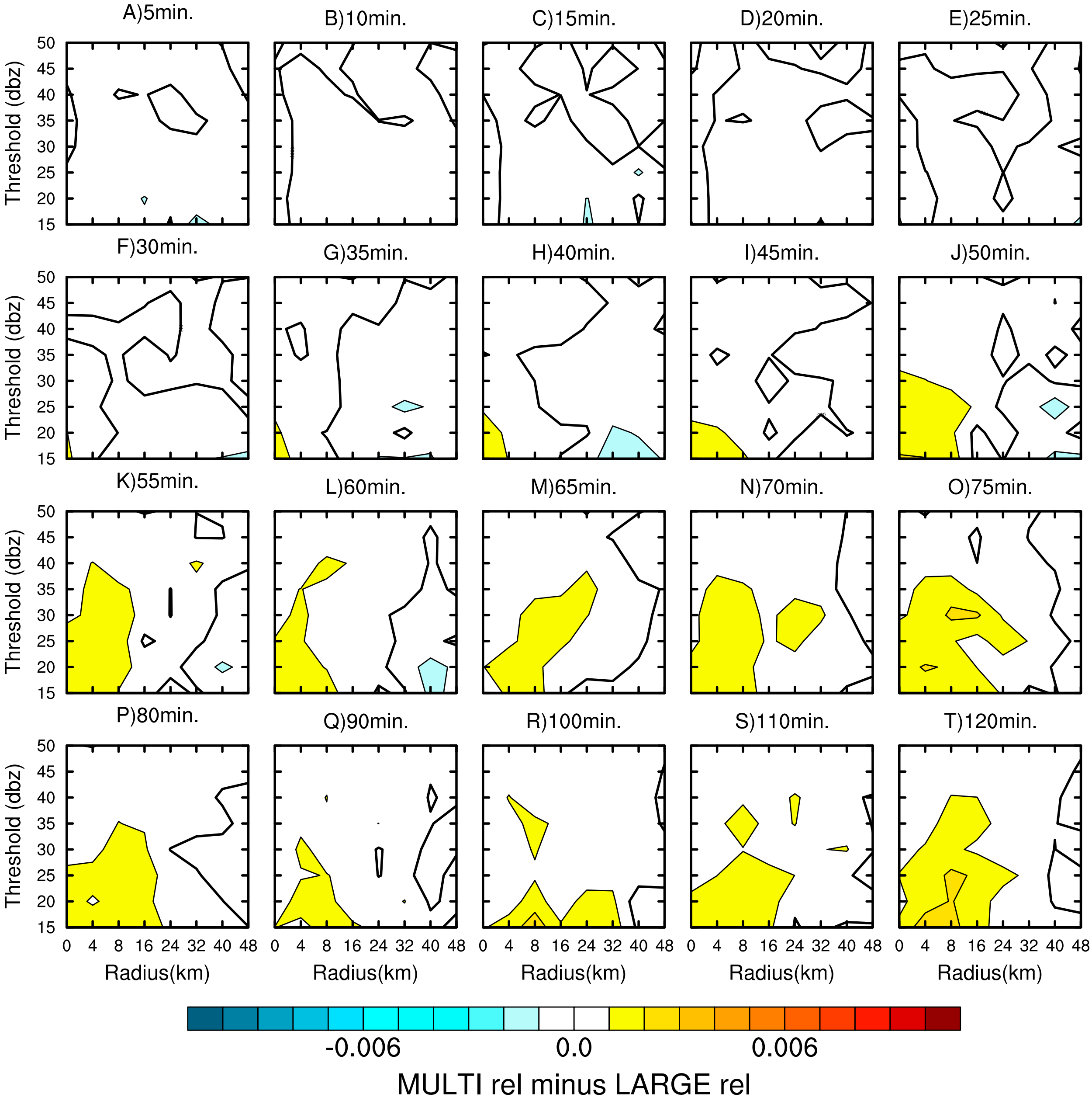

FIG. 10. As in Fig. 8, but for the reliability component of the Brier score. Significance tests were not repeated separately for the resolution and reliability components.

For the 20 May case study, the mesoscale precipitation forecast differences between MULTI and LARGE are primarily determined by the differences between MULTI48 and LARGE (i.e., subjective similarity between the left and right columns of Fig. 6 and between the blue lines in Fig. 5). In particular, the reduction of spurious precipitation behind the MCS for MULTI, and the subsequent differences from LARGE in the northern and eastern parts of the MCS at later times are also present in the differences between MULTI48 and LARGE (Figs. 6s–α). The differences from LARGE in the southern part of the MCS are also more strongly impacted by the mesoscale IC perturbation differences than the small-scale IC perturbations since the left and right columns of Fig. 6 are more similar in this area than the left and middle columns at most lead times. Since the mesoscale differences in the IC perturbation methods have similar qualitative impacts on precipitation and reflectivity forecasts, the qualitative explanation of these

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

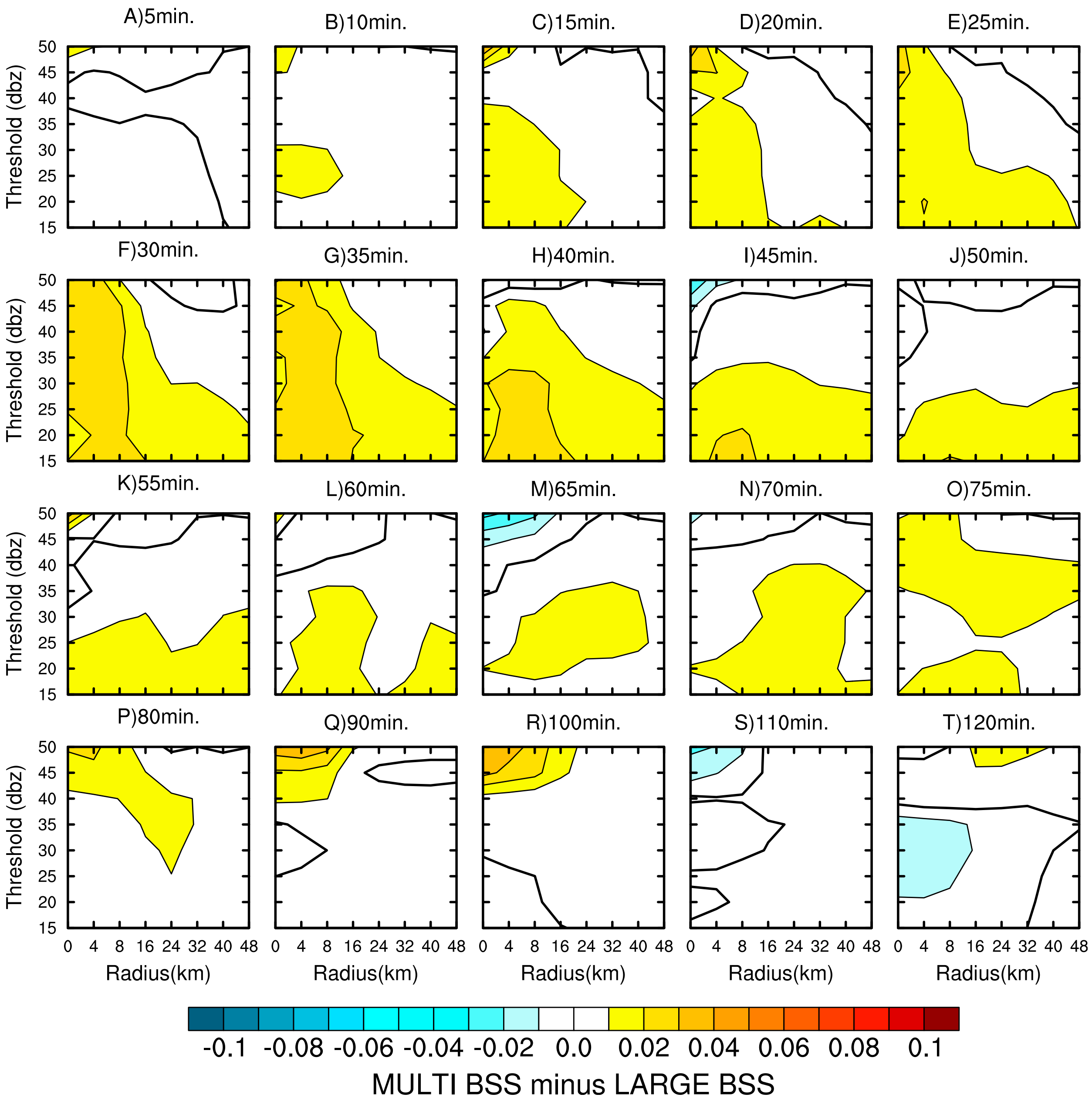

FIG. 11. Difference in BSS in the 0000 UTC 20 May case study for reflectivity at model level 12 between the MULTI and LARGE ensemble at 5-min intervals during the first 80 min and at 10-min intervals between 80 and 120 min. The vertical axis on each panel is the reflectivity threshold (dB*Z*) and the horizontal axis is the neighborhood radius (km).

forecast differences are explained in the following subsection.

Systematically, the differences in mesoscale IC perturbations result in skill advantages for MULTI48 and MULTI, compared to LARGE, at both early (1 h; i.e., valid at 0100 UTC) and later (~5–9 h; i.e., valid at 0500–0900 UTC) lead times (Figs. 4a,b). However, the skill is generally statistically indistinguishable among the forecasts at ~2–4-h lead times (Fig. 4). One possible explanation is that this is a result of the diurnal cycle of convective precipitation. Many of the cases show more convection over larger areas during the evening hours (i.e., the first ~4 h) than the overnight hours when only the better organized systems tend to be maintained (after ~0300–0400 UTC; not shown). It is hypothesized that the advantage of MULTI48 is greatest for the more organized long-lived MCSs, allowing the advantage to be objectively more pronounced after ~4 h when most of the precipitation is associated with such systems.

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

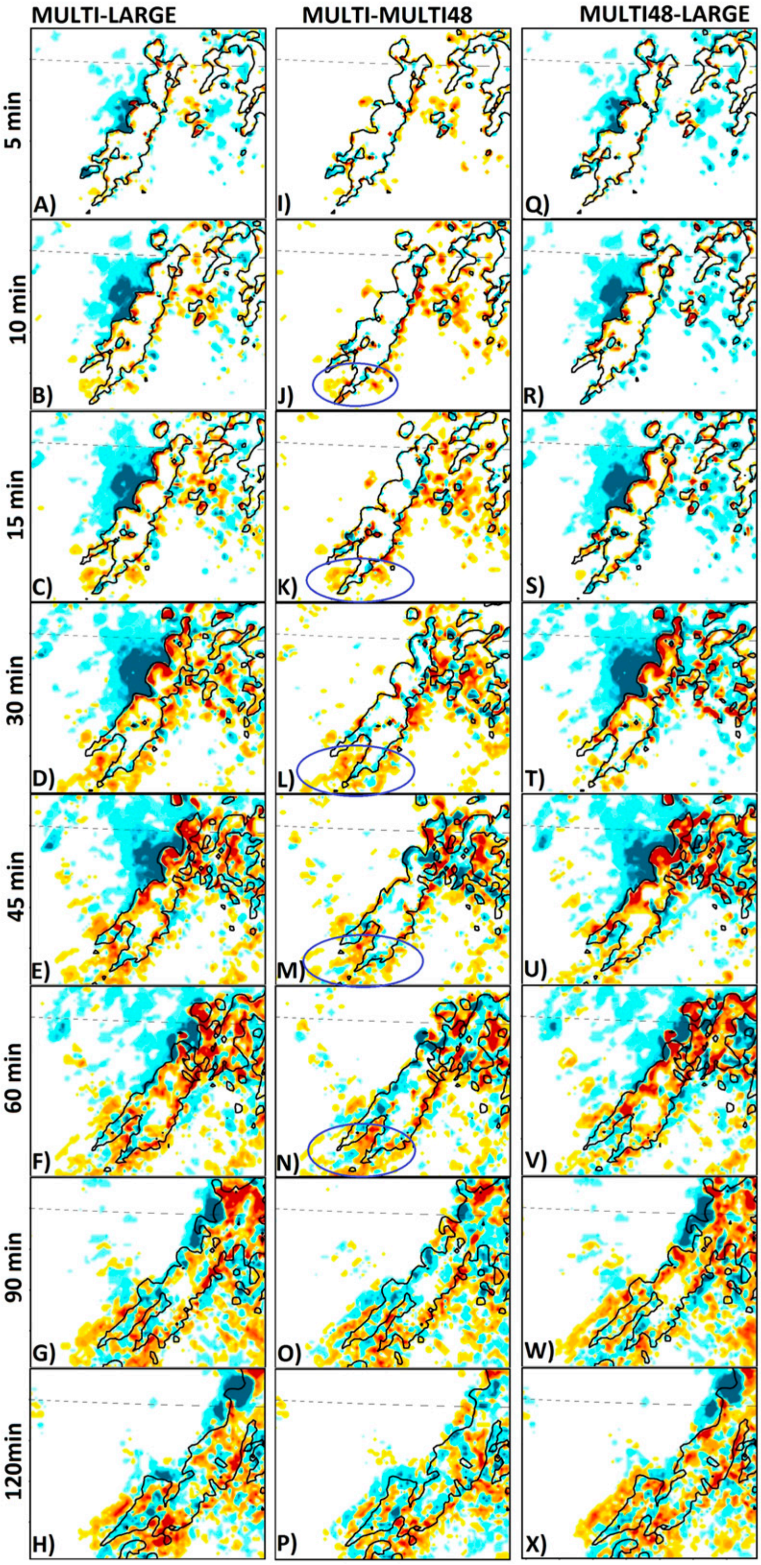

FIG. 12. As in Fig. 7, but for forecasts of reflectivity exceeding 30 dBZ in the verification domain focused on the MCS of interest. The color scale is the same as in Fig. 7.

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

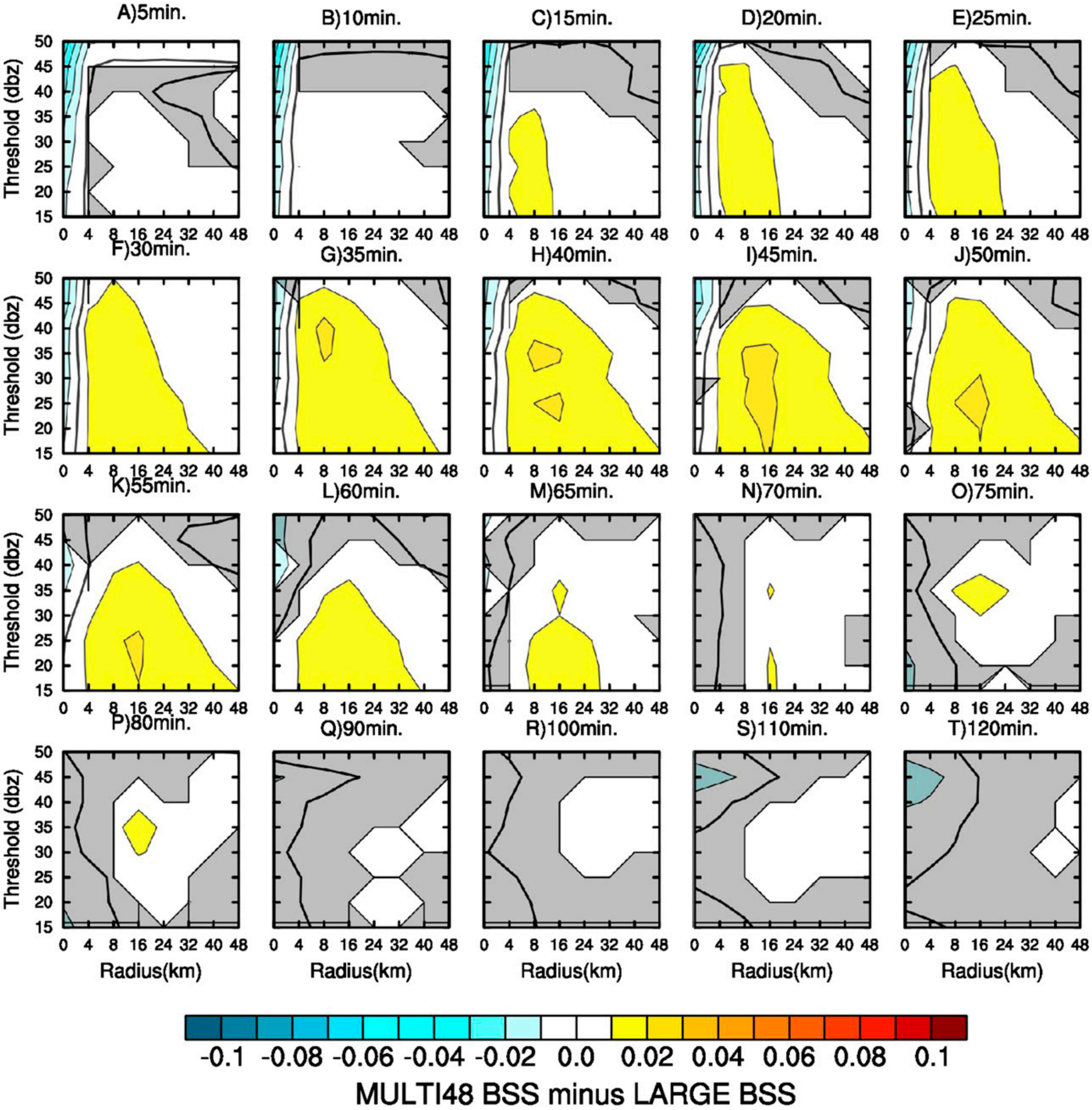

FIG. 13. As in Fig. 8, but for the BSS difference between MULTI48 and LARGE.

(ii) *Storm-scale reflectivity*

For the reflectivity forecasts, the differences in skill between MULTI and LARGE are also dominated by the differences between MULTI48 and LARGE, with the exception of the 0–4-km radius neighborhoods during approximately the first hour (Fig. 13). Statistically significant MULTI48 advantages for some radii/thresholds persist throughout the 2-h forecast period, although the differences become small by the end of the period. MULTI48 has lower skill than LARGE at 0–4 km for early lead times (Figs. 13a–j) because the small scales of MULTI48 are sharply truncated while the small-scale energy in LARGE approaches zero more gradually. The differences in reflectivity BSS between MULTI48 and LARGE also correspond primarily to differences in the resolution component (not shown).

The MULTI48 advantages over LARGE, for both the reflectivity and precipitation forecasts, are attributed to greater consistency of the mesoscale IC perturbations with the analysis errors in the vicinity of the analyzed MCS for MULTI48. For example, member 6 from the

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

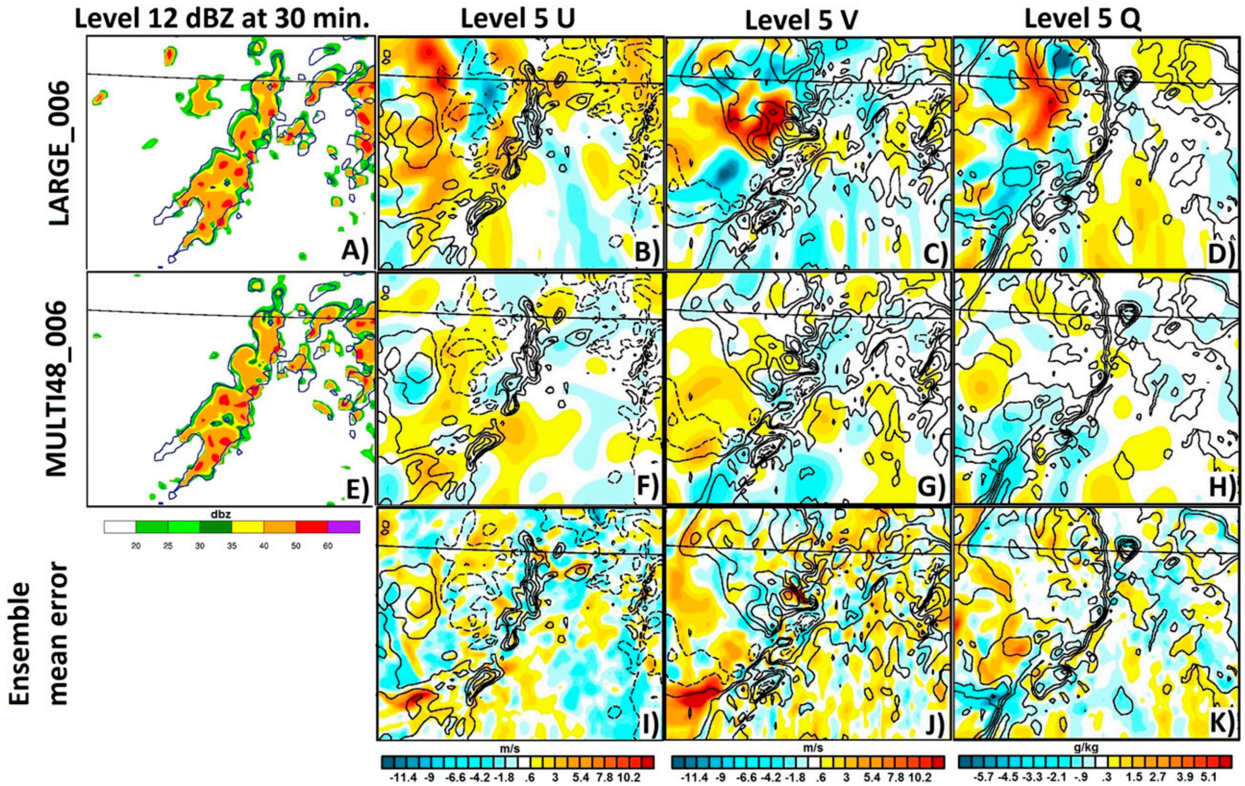

FIG. 14. Comparison of the 20 May case initial perturbations of member 006 from the LARGE and MULTI48 ensembles with the corresponding ensemble mean error. The 30-min reflectivity forecast at model level 12 for (a) LARGE_006 and (e) MULTI48_006. The LARGE_006 IC perturbation from the ensemble mean for the (b) *u* component of wind, (c) *v* component of wind, and (d) water vapor at model level 5. (f)–(h) As in (b)–(d), but for the MULTI48_006 IC perturbation. (i)–(k) The corresponding ensemble mean error (ensemble mean minus truth; note the IC ensemble mean is identical for all ensembles). Black contour overlays are the ensemble mean fields with contour intervals of 5 m s^{-1} for wind (negative values dashed) and 2 g kg^{-1} for water vapor.

LARGE ensemble in the 20 May case study (hereafter LARGE_006) shows several spurious cells west of the MCS in the cold pool region at early lead times (Fig. 14a). The corresponding member 6 of the MULTI48 ensemble (hereinafter MULTI48_006) does not show these spurious cells (Fig. 14e). The spurious cells result from strong convergence and moisture IC perturbations in the vicinity of the analyzed MCS cold pool for LARGE_006 that are not present for MULTI48_006 (Figs. 14b,c,d,f,g,h). Such perturbations may be consistent with the poorly resolved and poorly analyzed cold pools in the outer domain analysis. However, they are inconsistent with the errors of the inner domain analysis of this feature after radar DA (Figs. 14i–k). Therefore the improved consistency between the mesoscale IC perturbations and analysis errors near the analyzed MCS for MULTI48, compared to LARGE, explains the reduction in spurious probability in the cold pool region at early lead times for this case. The excessively large magnitude mesoscale perturbations in and near the initial MCS for LARGE also result in the less sharp probabilistic forecast of the MCS for LARGE, consistent with the poorer resolution component for the LARGE forecast noted above. The smaller magnitude mesoscale IC perturbations for MULTI48 also explain the initially lower ensemble spread of nonprecipitation variables for MULTI, compared to LARGE (Fig. 3). The above discussion was also found to apply to the other cases in this study as well (not shown). Generating just the mesoscale part of the IC perturbations while assimilating radar observations on the convection-permitting grid, and allowing upscale and downscale interactions with the convective scales, is therefore advantageous for storm-scale reflectivity forecasts, in addition to the mesoscale precipitation forecasts.

The impact of the smaller magnitude mesoscale perturbations in MULTI48 than LARGE on ensemble spread and accuracy is quantified with the dispersion and error fractions skill score (dFSS and eFSS, respectively; Dey et al. 2014). As described in greater detail in Dey et al. (2014), the fractions score (FS) is the mean square difference between the forecast and observed neighborhood probability (NP) field. The observed NP is calculated the same way as the forecast NP,

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

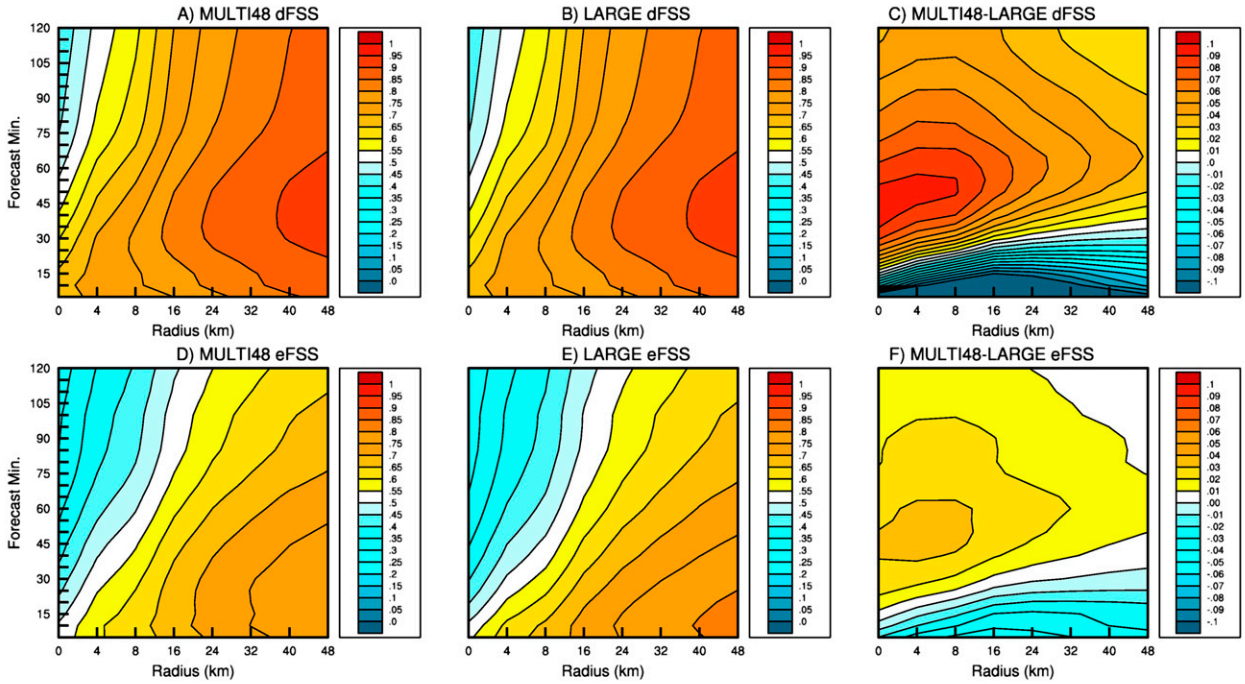

FIG. 15. Dispersion fractions skill score (dFSS) for the (a) MULTI48 ensemble, (b) LARGE ensemble, and (c) difference between the MULTI48 and LARGE ensembles; and error fractions skill score (eFSS) for the (d) MULTI48 ensemble, (e) LARGE ensemble, and (f) difference between the MULTI48 and LARGE ensembles.

instead of using a binary verification field as in the other NEP skill scores. The FSS is then calculated as 1 − FS/FS$_{\text{ref}}$, where FS$_{\text{ref}}$ is the FS that would be obtained if there were no overlap between the forecast and observed NP fields (i.e., the sum of the mean square forecast and observed NP fields; Dey et al. 2014). The dFSS is calculated as the average FSS between all possible member-member pairs as a measure of ensemble spread. Smaller values of dFSS indicate greater spread. The eFSS is calculated as the average FSS between all member-observation pairs and is a measure of the deterministic forecast accuracy of the ensemble members. Smaller values of eFSS indicate greater error of the individual deterministic forecasts comprising the ensemble. An advantage of this method is that it can be calculated over a range of radii to understand the scale dependence of the ensemble characteristics.

The dFSS of reflectivity forecasts is systematically larger for MULTI48 than LARGE after ~20 min, indicating less ensemble spread for MULTI48 (Fig. 15c). This is a result of the smaller magnitude mesoscale IC perturbations in MULTI48. The larger spread for MULTI48 than LARGE during the first ~20 min is likely due to the fact that significant hydrometeor perturbations are not present in the initial LARGE ensemble downscaled from the convection-parameterizing outer domain. It therefore takes some time for the directly perturbed variables to generate reflectivity spread. After ~15–30 min, depending on spatial scale, the smaller spread for MULTI48 also corresponds to larger eFSS values, indicating less error for the MULTI48 members than for the LARGE members (Fig. 15f). Therefore, Fig. 15 shows that the MULTI48 members are systematically both closer to each other and closer to the observations than the LARGE members, consistent with the generally more skillful forecasts for MULTI48 (Fig. 13) and the sharper reflectivity forecasts. Even during the first ~15–30 min, when the individual members have larger errors for MULTI than for LARGE (Fig. 15f), the ensemble probability forecasts are more skillful for MULTI than LARGE (Fig. 8).

3) IMPACT OF SMALL-SCALE COMPONENT OF IC PERTURBATION METHOD (MULTI VS MULTI48)

(i) *Mesoscale hourly accumulated precipitation*

The third goal of this study is to understand the systematic impacts of the small-scale IC perturbations which are resolved by MULTI but not LARGE. MULTI is therefore compared to MULTI48 which does not contain such small-scale IC perturbations. The systematic verification reveals that the small-scale IC perturbations increase the forecast skill starting at ~6 h,

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

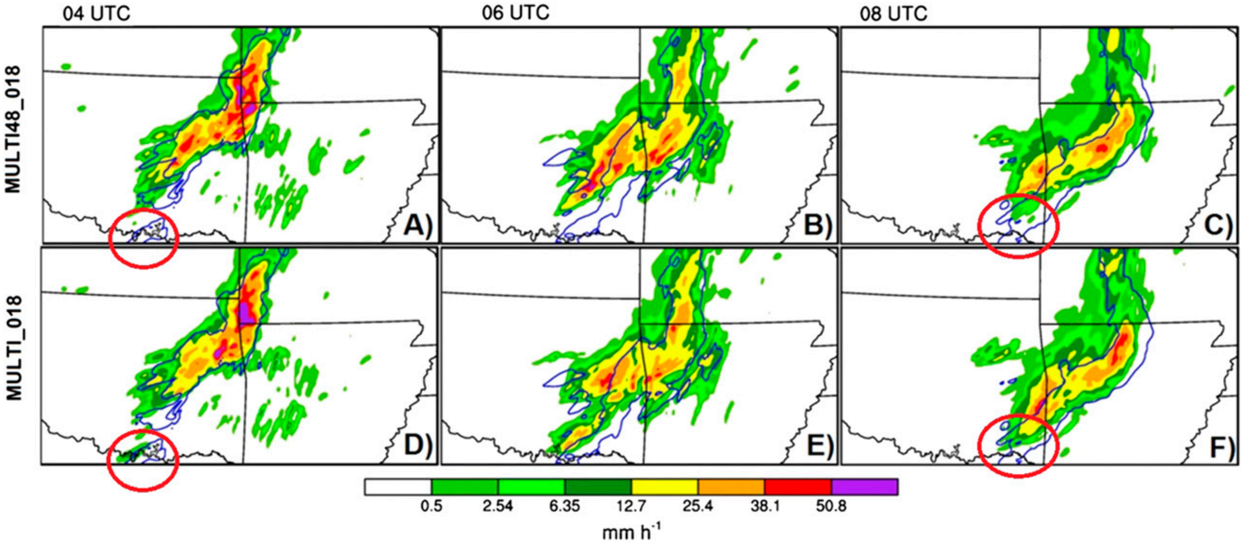

FIG. 16. Forecasts of hourly accumulated precipitation initialized at 0000 UTC 20 May and valid at (a),(d) 0400; (b),(e) 0600; and (c),(f) 0800 UTC for member 18 of the (a)–(c) MULTI48 and (d)–(f) MULTI ensemble. The observation contour at the 6.35 mm h^{-1} level is overlaid in blue.

although the impact on skill at earlier lead times is small or insignificant (Fig. 4). MULTI is more skillful than MULTI48 at these later lead times for all thresholds, with statistical significance at 8–9, 7–8, and 6–7 h at 2.54, 6.35, and 12.7 mm h^{-1} thresholds, respectively (Fig. 4). This result shows that it is important to explicitly include such perturbations in the IC perturbation design, rather than rely on the downscale propagation of perturbation energy indicated by Fig. 1 and Durran and Gingrich (2014). Figure 4 also shows that at forecast hours 2–5, MULTI is slightly less skillful than MULTI48 at all thresholds, although the difference is only significant at the 2-h lead time for the 12.7 mm h^{-1} threshold (Fig. 4). This negative impact of the small-scale IC perturbations may be related to an initial enhancement of disorganized weak convection surrounding the observed convective systems (Fig. 12).

For the 20 May case study, the small-scale IC perturbations contribute to the overall NEP difference at some locations, especially at later lead times. For example, the small-scale IC perturbations increase the probability of precipitation where a storm is observed along the Oklahoma–Texas border at ~0300–0500 UTC (Figs. 7l–n; blue circles). This leads to a corresponding increase in probability along the southeast edge of the MCS at ~0600–0800 UTC (Figs. 7o–q; blue circles). The small-scale IC perturbations also contribute to the decrease in forecast probability to the west of the southern half of the MCS at later lead times, especially at ~0400–0700 UTC (Figs. 7m–p; red circles). The impact of the small-scale IC perturbations in this localized area (although not over the entire domain) is nearly as large, and at some times and places larger than, the impact of the differences in mesoscale IC perturbations. Therefore, while the mesoscale component of the IC perturbations dominates the ensemble forecast skill for this case, the small-scale IC perturbations are not entirely unimportant for the mesoscale hourly accumulated precipitation forecasts.

Subjective evaluation of the differences between individual members of the MULTI and MULTI48 ensembles for the 20 May case study demonstrates how the small-scale IC perturbations can directly affect the development of new convection during the early forecast hours (e.g., Fig. 16). Given the smaller spatial scale of newly developing convection, it is not surprising that it is particularly sensitive to the small-scale IC perturbations. Such convection can then grow upscale during the forecast period, influencing the mesoscale precipitation forecast at later lead times. The continued development of new convection during the early forecast period thus provides a mechanism for the small-scale IC perturbations to impact the mesoscale precipitation forecasts at later lead times. The small-scale perturbation energy that rapidly develops through downscale energy propagation (i.e., Fig. 2) may not have as much impact on the newly developing convection. An example of this mechanism is demonstrated by ensemble member 18 in Fig. 16. The MULTI48 member does not forecast convection along the Texas–Oklahoma border at 0400 UTC (Fig. 16a; red circle). However, the corresponding MULTI member, which also includes the small-scale component of the IC perturbation, does forecast such a convective cell (Fig. 16d; red circle). The location of the

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

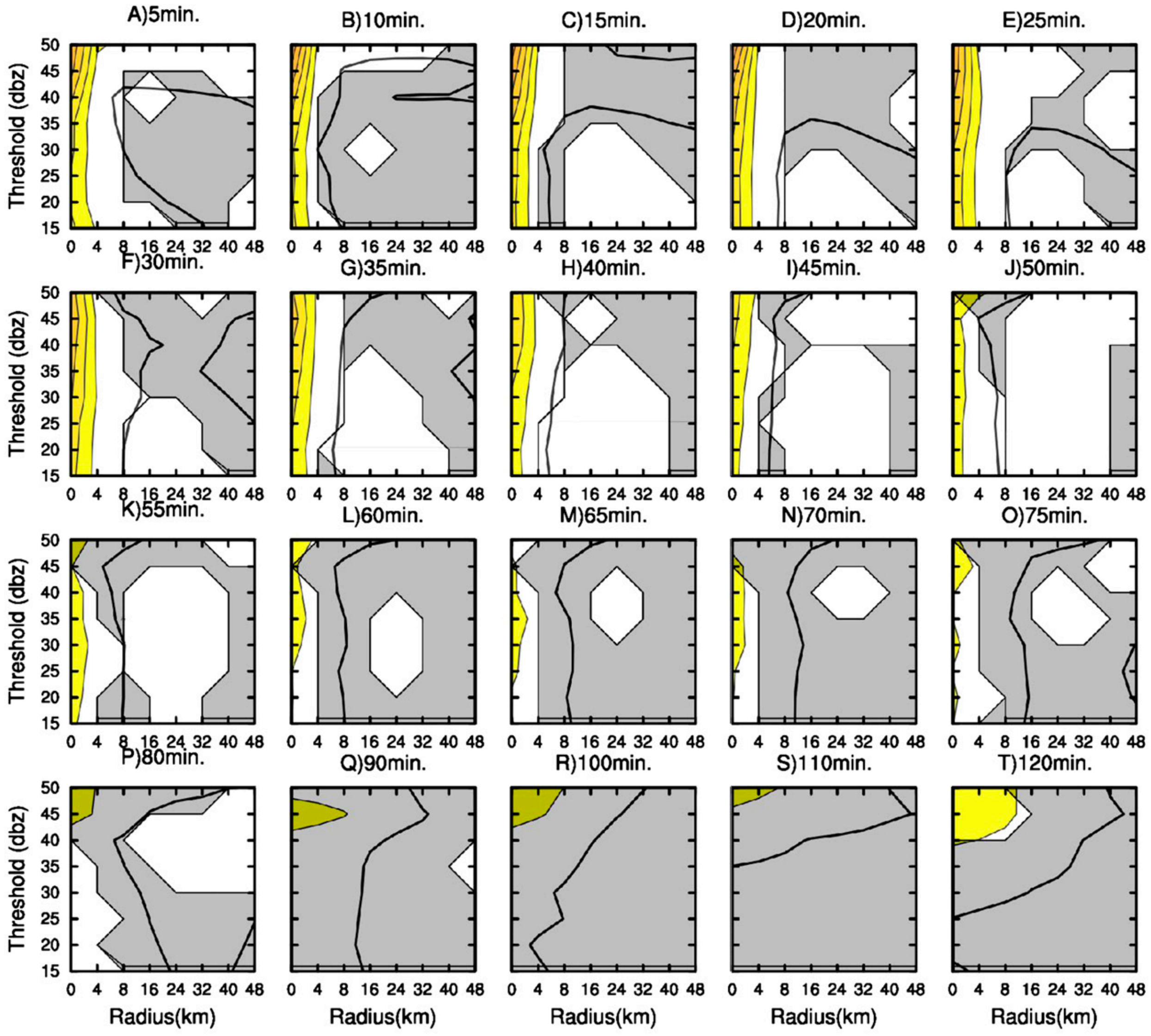

FIG. 17. As in Fig. 8, but for the BSS difference between MULTI and MULTI48.

cell in MULTI is slightly west of and weaker than the observed convection at 0400 UTC. However, for this case the upscale growth of the cell results in the southern end of the MCS being farther southeast and closer to the observed MCS by 0800 UTC for MULTI (Fig. 16f; red circle) than for MULTI48 (Fig. 16c; red circle).

(ii) *Storm-scale reflectivity*

The small-scale IC perturbations in MULTI also systematically improve the reflectivity forecasts over MULTI48 on small forecast scales (i.e., no neighborhood radius; Fig. 17). Although small-scale IC perturbations lead to some small (but statistically significant) disadvantages (e.g., Figs. 17g–j at scales >8 km), there are more pronounced advantages at small scales during approximately the first hour (Figs. 17a–o at scales <8 km). As demonstrated in the following paragraph, the small-scale early advantage corresponds to subjectively smoother probability gradients where grid-scale details of the observation contour at a particular threshold are not well forecast. This contrasts with the mesoscale precipitation forecasts that are improved by upscale growth

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

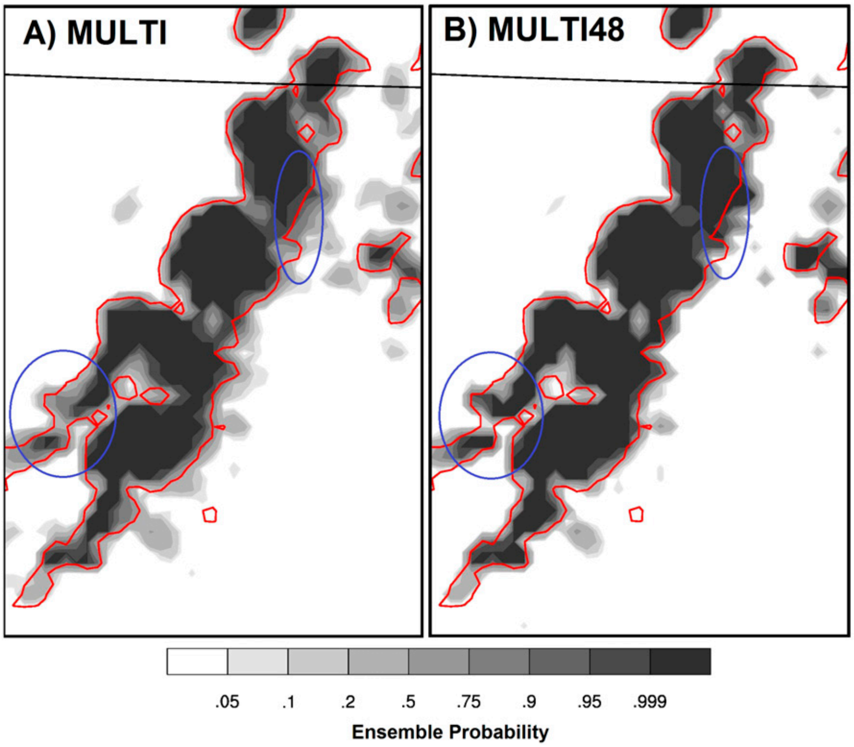

FIG. 18. NEP forecast (with 0-km neighborhood radius) of reflectivity exceeding 30 dBZ at the 15-min lead time (shaded) and observation contour for the 20 May case for (a) MULTI and (b) MULTI48.

of the explicitly added flow-dependent small-scale IC perturbations. For storm-scale reflectivity forecasts, the impact of the small-scale IC perturbations is limited to the very small scales and the time period before downscale propagation generates sufficient perturbation energy on such scales (Fig. 2). The slight degradation of skill at some lead times likely results from an increase in weak spurious convection away from the convective system caused by the small-scale perturbations.

Subjectively, there are two clear impacts of the small-scale IC perturbations in the 20 May case study. First, the MULTI48 NEP forecasts at short lead times show small-scale features of the MCS with strong probability gradients that do not necessarily line up with the observation contour (e.g., Fig. 18b; blue circles). The small-scale IC perturbations in MULTI smooth out the NEP gradient in such cases, making the probabilistic forecasts more consistent with the uncertainty of such features (e.g., Fig. 18a; blue circles). This explains the better BSS for MULTI than MULTI48 for zero or small neighborhood radii during the first ~1 h (e.g., Fig. 17; also seen in the 20 May case, not shown). Second, the small-scale IC perturbations generally increase the probability in several areas where no precipitation is observed (e.g., Figs. 12j–n; blue circles). This is a result of large areas of weak convection resulting from the small-scale IC perturbations and is most pronounced at lower reflectivity thresholds (not shown).

The systematic impact of the small-scale IC perturbations on ensemble spread and accuracy is also quantified with the dFSS and eFSS (Fig. 19). Compared to MULTI48, MULTI initially has greater spread at the grid scale, which grows to slightly larger scales during the first ~45–60 min (Fig. 19c). This difference in spread, resulting from the small-scale IC perturbations in MULTI, remains maximized in neighborhoods of 0–4 km, consistent with the impact on ensemble forecast skill occurring on such scales (Fig. 17). The greater spread for MULTI corresponds to more error of the individual ensemble members (Fig. 19f). This is expected because the ensemble mean is expected to be centered on the most likely observation so as spread increases the accuracy of any given member is expected to be less. However, in the ensemble context this is advantageous since the NEP skill is greater for MULTI than MULTI48 at similar times and scales because the ensemble better reflects the forecast uncertainty for MULTI than MULTI48 (Fig. 17). The impact of the small-scale IC perturbations on both ensemble spread and accuracy begins to

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

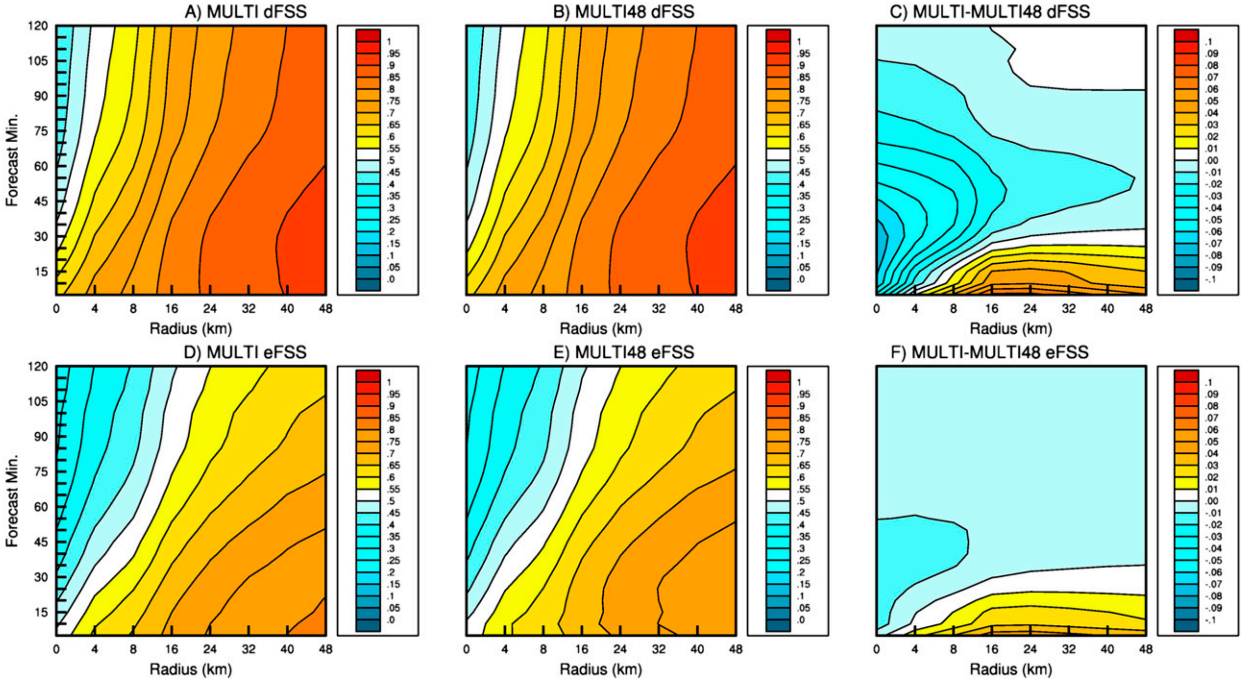

FIG. 19. As in Fig. 15, but for the MULTI and MULTI48 ensembles.

diminish after ~60 min (Figs. 19c,f), also consistent with Fig. 17.

4. Summary and discussion

This study tests the hypothesis that the flow-dependent multiscale IC perturbations resulting from multiscale ensemble data assimilation can provide ensemble forecast advantages that were not found with more simplified methods of generating multiscale IC perturbations for SSEFs. This hypothesis is tested by considering the following three questions. First, what are the impacts on ensemble forecast skill of generating IC perturbations with a multiscale ensemble data assimilation system (MULTI), compared to downscaling larger-scale IC perturbations from a coarser domain (LARGE)? Second, what role does the mesoscale component (i.e., resolved by both MULTI and LARGE) of the IC perturbation differences have in determining the ensemble forecast skill? Third, what role do the small-scale (i.e., only resolved with MULTI) IC perturbations have in the differences between MULTI and LARGE ensemble forecast skill? The impacts of the IC perturbations are evaluated in terms of 2-h reflectivity forecasts over a range of neighborhood radii less than 48 km and in terms of 9-h mesoscale (i.e., 48-km neighborhood radius) hourly accumulated precipitation forecasts. A perfect-model OSSE framework is used to isolate the impacts of IC error and the corresponding IC perturbations. In the OSSE framework, the governing dynamics and physics of the experiment model are identical to that of the truth simulation, leaving IC/LBC uncertainty as the only source of error to be sampled in the ensemble design. It is possible that some results could be sensitive to the choice of configuration for the experiment and truth simulations. Future work with real data experiments without the perfect-model OSSE framework are therefore still needed to determine how much forecast skill improvement is obtained by optimal IC perturbations in more realistic scenarios that also include model and physics errors. However, such experiments could also have results that are dependent on particular physics configurations.

The impact of the different IC perturbation methods on the spread of the directly perturbed nonprecipitation variables is first evaluated. The LARGE IC perturbations are much more underdispersive than MULTI on scales less than ~50 km. However, as expected from the results of Durran and Gingrich (2014), the downscale cascade of perturbation energy results in similar perturbation spectra between MULTI and LARGE within ~1 h. The total spread of nonprecipitation variables is dominated by the larger scales that initially show less spread for MULTI than LARGE for all variables except for level 12 (~750 hPa) moisture.

In addition to the spread of the directly perturbed nonprecipitation variables, the skill of the ensemble

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

forecasts of convective precipitation on different time and space scales is also evaluated. Comparison of the MULTI and LARGE ensembles addresses the question of how the forecast skill is affected by the differences between multiscale IC perturbations generated with multiscale ensemble data assimilation versus coarser-resolution downscaled IC perturbations. The comparison of MULTI and LARGE reveals statistically significant skill advantages for MULTI at the 1-h lead time for all thresholds and at several lead times after 4 h for the 2.54 and 6.35 mm h^{-1} thresholds. The storm-scale reflectivity forecasts are more skillful for MULTI than LARGE for about 45 (at higher thresholds) to 65 (at lower thresholds) min. On average, the MULTI IC perturbations therefore represent a more optimal method of sampling the IC uncertainty for SSEFs of midlatitude convection than LARGE. The comparisons of MULTI48 with LARGE and MULTI with MULTI48 provide further understanding of the reasons for the MULTI advantages over LARGE.

Comparison of MULTI48 and LARGE addresses the question of how the differences in the mesoscale component of the IC perturbations affect the forecast skill. The differences in ensemble forecast skill are explained mainly by the mesoscale component of the differences between the IC perturbation methods, with the exception of the first hour of reflectivity forecasts using neighborhood radii of 0–4 km. The comparison of MULTI48 with LARGE is generally similar to the comparison of MULTI and LARGE for mesoscale precipitation forecasts at 2.54 and 6.35 mm h^{-1} thresholds and storm-scale reflectivity forecasts with neighborhood radii >4 km. It was demonstrated with the 20 May case study, and confirmed with other cases as well (not shown), that the MULTI48 perturbations have less amplitude and are more consistent with the analysis error in the vicinity of the analyzed MCSs and corresponding cold pools, compared to LARGE. This leads to subjectively and objectively better probabilistic forecasts of both reflectivity and hourly accumulated precipitation. However, the advantages of MULTI48 over LARGE for mesoscale precipitation at ~5–9-h lead times are less statistically significant than the advantages of MULTI over LARGE. Furthermore, at the 12.7 mm h^{-1} threshold, the mesoscale and small-scale IC perturbations have impacts on skill of similar magnitude, showing that the small-scale IC perturbations also play an important role. The small scales (i.e., 0–4-km neighborhoods) of the reflectivity forecasts during the first hour are also an exception to the dominance of the mesoscale IC perturbations.

Comparison of MULTI and MULTI48 addresses the question of how the presence of small-scale IC perturbations in MULTI affects the forecast skill. Advantages of the small-scale IC perturbations were found for both reflectivity and hourly precipitation, although for some lead times and spatial scales the reflectivity forecasts were made slightly less skillful by the small-scale IC perturbations. Since Fig. 2 and Durran and Gingrich (2014) both suggest that small-scale perturbations rapidly develop as a result of downscale energy propagation, the appearance of mesoscale forecast advantages for MULTI, compared to MULTI48, at much later lead times is particularly noteworthy. For the 20 May case study, the impact of the small-scale IC perturbations on the mesoscale precipitation forecasts resulted from their impact on new convection that developed during the early forecast period. The new cells originated from small-scale features, explaining their sensitivity to the small-scale IC perturbations. At later times, such cells also influenced the mesoscale convective systems, explaining the upscale growth of this impact onto the mesoscale hourly accumulated precipitation forecasts throughout the 9-h forecast period for this case. It should also be noted, however, that the systematic impact on skill at earlier lead times is small or statistically insignificant (Fig. 4).

In summary, this study shows some of the ways that IC perturbations for SSEFs of midlatitude convection more optimally sample the analysis uncertainty when generated at full resolution with ensemble-based multiscale data assimilation than when downscaled from a coarser ensemble. The greater consistency between the mesoscale IC perturbations and analysis uncertainty and the presence of flow-dependent smaller-scale IC perturbations both contribute to forecast advantages for the multiscale IC perturbation method. This work provides two new scientific findings that have not been shown in past work. First, an advantage of the small-scale (i.e., order of 10 km) component of multiscale IC perturbations found in this work had not been previously shown. Second, the advantage demonstrated in this study of generating the larger-scale component of the IC perturbations on the same grid as the forecast model had not been previously shown to the authors' knowledge.

Acknowledgments. The work is primarily supported by NSF Awards AGS-1046081 and AGS-1359703. Some of the computing for this project was performed at the OU Supercomputing Center for Education and Research (OSCER) at the University of Oklahoma (OU). This work also used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation Grant ACI-1053575, and Yellowstone machine (ark:/85065/d7wd3xhc) at NCAR's Computational and Information

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

Systems Laboratory, sponsored by NSF. Much of this paper closely follows, or is excerpted from, the lead author's Ph.D. dissertation (Johnson 2014).

REFERENCES

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Cintineo, R. M., and D. J. Stensrud, 2013: On the predictability of supercell thunderstorm evolution. *J. Atmos. Sci.*, **70**, 1993–2011, doi:10.1175/JAS-D-12-0166.1.

Clark, A. J., W. A. Gallus Jr., and T.-C. Chen, 2008: Contributions of mixed physics versus perturbed initial/lateral boundary conditions to ensemble-based precipitation forecast skill. *Mon. Wea. Rev.*, **136**, 2140–2156, doi:10.1175/2007MWR2029.1.

——, ——, 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, doi:10.1175/2009WAF2222222.1.

——, and Coauthors, 2012: An overview of the 2010 Hazardous Weather Testbed experimental forecast program spring experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74, doi:10.1175/BAMS-D-11-00040.1.

Denis, B., J. Côté, and R. Laprise, 2002: Spectral decomposition of two-dimensional atmospheric fields on limited-area domains using the discrete cosine transform (DCT). *Mon. Wea. Rev.*, **130**, 1812–1829, doi:10.1175/1520-0493(2002)130<1812:SDOTDA>2.0.CO;2.

Dey, S. R. A., 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, doi:10.1175/MWR-D-14-00172.1.

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

Durran, D. R., and M. Gingrich, 2014: Atmospheric predictability: Why butterflies are not of practical importance. *J. Atmos. Sci.*, **71**, 2476–2488, doi:10.1175/JAS-D-14-0007.1.

Ehrendorfer, M., 1997: Predicting the uncertainty of numerical weather forecasts: A review. *Meteor. Z.*, **6**, 147–183.

Hamill, T. M, 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, doi:10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

Harnisch, F., and C. Keil, 2015: Initial conditions for convective-scale ensemble forecasting provided by ensemble data assimilation. *Mon. Wea. Rev.*, **143**, 1583–1600, doi:10.1175/MWR-D-14-00209.1.

Hohenegger, C., and C. Schär, 2007a: Predictability and error growth dynamics in cloud-resolving models. *J. Atmos. Sci.*, **64**, 4467–4478, doi:10.1175/2007JAS2143.1.

——, and ——, 2007b: Atmospheric predictability at synoptic versus cloud-resolving scales. *Bull. Amer. Meteor. Soc.*, **88**, 1783–1793, doi:10.1175/BAMS-88-11-1783.

——, D. Lüthi, and C. Schär, 2006: Predictability mysteries in cloud-resolving models. *Mon. Wea. Rev.*, **134**, 2095–2107, doi:10.1175/MWR3176.1.

——, A. Walser, W. Langhans, and C. Schär, 2008: Cloud-resolving ensemble simulations of the August 2005 Alpine flood. *Quart. J. Roy. Meteor. Soc.*, **134**, 889–904, doi:10.1002/qj.252.

Johnson, A., 2014: Optimal design of a multi-scale ensemble system for convective scale probabilistic forecasts: Data assimilation and initial condition perturbation methods. Ph.D. dissertation, School of Meteorology, University of Oklahoma, 182 pp.

——, and X. Wang, 2012: Verification and calibration of neighborhood and object-based probabilistic precipitation forecasts from a multimodel convection-allowing ensemble. *Mon. Wea. Rev.*, **140**, 3054–3077, doi:10.1175/MWR-D-11-00356.1.

——, ——, F. Kong, and M. Xue, 2011a: Hierarchical cluster analysis of a convection-allowing ensemble during the Hazardous Weather Testbed 2009 Spring Experiment. Part I: Development of the object-oriented cluster analysis method for precipitation fields. *Mon. Wea. Rev.*, **139**, 3673–3693, doi:10.1175/MWR-D-11-00015.1.

——, ——, M. Xue, and F. Kong, 2011b: Hierarchical cluster analysis of a convection-allowing ensemble during the Hazardous Weather Testbed 2009 Spring Experiment. Part II: Ensemble clustering over the whole experiment period. *Mon. Wea. Rev.*, **139**, 3694–3710, doi:10.1175/MWR-D-11-00016.1.

——, and Coauthors, 2014: Multiscale characteristics and evolution of perturbations for warm season convection-allowing precipitation forecasts: Dependence on background flow and method of perturbation. *Mon. Wea. Rev.*, **142**, 1053–1073, doi:10.1175/MWR-D-13-00204.1.

——, X. Wang, J. R. Carley, L. J. Wicker, and C. Karstens, 2015: A comparison of multiscale GSI-based EnKF and 3DVar data assimilation for midlatitude convective-scale precipitation forecasts. *Mon. Wea. Rev.*, **143**, 3087–3108, doi:10.1175/MWR-D-14-00345.1.

Kong, F., K. K. Droegemeirer, and N. L. Hickmon, 2007: Multiresolution ensemble forecasts of an observed tornadic thunderstorm system. Part II: Storm-scale experiments. *Mon. Wea. Rev.*, **135**, 759–782, doi:10.1175/MWR3323.1.

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, doi:10.1002/qj.2238.

Leoncini, G., R. S. Plant, S. L. Gray, and P. A. Clark, 2010: Perturbation growth at the convective scale for CSIP IOP18. *Quart. J. Roy. Meteor. Soc.*, **136**, 653–670, doi:10.1002/qj.587.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289–307, doi:10.1111/j.2153-3490.1969.tb00444.x.

Mahajan, R., D. T. Kleist, C. Thomas, J. C. Derber, and R. Treadon, 2016: Implementation plans of hybrid 4D EnVar for the NCEP GFS and future directions. *20th Conf. on Integrated Observing and Assimilation Systems for the Atmosphere, Oceans and Land Surface*, New Orleans, LA, Amer. Meteor. Soc., J3.3. [Available online at https://ams.confex.com/ams/96Annual/webprogram/Paper288830.html.]

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600, doi:10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Peña, M., and E. Kalnay, 2004: Separating fast and slow modes in coupled chaotic systems. *Nonlinear Processes Geophys.*, **11**, 319–327, doi:10.5194/npg-11-319-2004.

Peralta, C., Z. B. Bouallegue, S. E. Theis, C. Gebhardt, and M. Buchhold, 2012: Accounting for initial condition uncertainties in COSMO-DE-EPS. *J. Geophys. Res.*, **117**, D07108, doi:10.1029/2011JD016581.

Perkey, D. J., and R. A. Maddox, 1985: A numerical investigation of a mesoscale convective system. *Mon. Wea. Rev.*, **113**, 553–566, doi:10.1175/1520-0493(1985)113<0553:ANIOAM>2.0.CO;2.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC

convective events. *Mon. Wea. Rev.*, **136**, 78–97, doi:10.1175/2007MWR2123.1.

Rotunno, R., and C. Snyder, 2008: A generalization of Lorenz's model for the predictability of flows with many scales of motion. *J. Atmos. Sci.*, **65**, 1063–1076, doi:10.1175/2007JAS2449.1.

Schwartz, C. S., 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, doi:10.1175/MWR-D-13-00100.1.

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

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

——, and J. B. Klemp, 2007: A time-split nonhydrostatic atmospheric model for research and NWP applications. *J. Comput. Phys.*, **135**, 3465–3485.

Snook, N., M. Xue, and Y. Jung, 2011: Analysis of a tornadic mesoscale convective vortex based on ensemble Kalman filter assimilation of CASA X-Band and WSR-88D radar data. *Mon. Wea. Rev.*, **139**, 3446–3468, doi:10.1175/MWR-D-10-05053.1.

Stensrud, D. J., and Coauthors, 2009: Convective-scale warn-on-forecast system: A vision for 2020. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1499, doi:10.1175/2009BAMS2795.1.

Stratman, D. R., M. C. Coniglio, S. E. Koch, and M. Xue, 2013: Use of multiple verification methods to evaluate forecasts of convection from hot- and cold-start convection-allowing models. *Wea. Forecasting*, **28**, 119–138, doi:10.1175/WAF-D-12-00022.1.

Theis, S. E., A. Hense, and U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: A pragmatic approach. *Meteor. Appl.*, **12**, 257–268, doi:10.1017/S1350482705001763.

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297–3319, doi:10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2.

Vié, B., O. Nuissier, and V. Ducrocq, 2011: Cloud-resolving ensemble simulations of Mediterranean heavy precipitation events: Uncertainty on initial conditions and lateral boundary conditions. *Mon. Wea. Rev.*, **139**, 403–423, doi:10.1175/2010MWR3487.1.

Wang, X., and T. Lei, 2014: GSI-based four-dimensional ensemble–variational (4DEnsVar) data assimilation: Formulation and single-resolution experiments with real data for NCEP Global Forecast System. *Mon. Wea. Rev.*, **142**, 3303–3325, doi:10.1175/MWR-D-13-00303.1.

——, D. Parrish, D. 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, doi:10.1175/MWR-D-12-00141.1.

Wang, Y., M. Bellus, J.-F. Geleyn, X. Ma, W. Tian, and F. Weidle, 2014: A new method for generating initial condition perturbations in a regional ensemble prediction system: Blending. *Mon. Wea. Rev.*, **142**, 2043–2059, doi:10.1175/MWR-D-12-00354.1.

Whitaker, J. S., T. M. Hamill, X. Wei, Y. Song, and Z. Toth, 2008: Ensemble data assimilation with the NCEP Global Forecast System. *Mon. Wea. Rev.*, **136**, 463–482, doi:10.1175/2007MWR2018.1.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences: An Introduction*. 2nd ed. Academic Press, 467 pp.

Xue, M., and Coauthors, 2010: CAPS realtime storm scale ensemble and high resolution forecasts for the NOAA Hazardous Weather Testbed 2010 Spring Experiment. *25th Conf. on Severe Local Storms*, Denver, CO, Amer. Meteor. Soc., 7B.3. [Available online at https://ams.confex.com/ams/25SLS/techprogram/paper_176056.htm.]

Yussouf, N., E. R. Mansell, L. J. Wicker, D. M. Wheatley, and D. J. Stensrud, 2013: The ensemble Kalman filter analyses and forecasts of the 8 May 2003 Oklahoma City tornadic supercell storm using single- and double-moment microphysics schemes. *Mon. Wea. Rev.*, **141**, 3388–3412, doi:10.1175/MWR-D-12-00237.1.

Zhang, F., C. Snyder, and R. Rotunno, 2003: Effects of moist convection on mesoscale predictability. *J. Atmos. Sci.*, **60**, 1173–1185, doi:10.1175/1520-0469(2003)060<1173:EOMCOM>2.0.CO;2.

——, A. M. Odins, and J. W. Nielsen-Gammon, 2006: Mesoscale predictability of an extreme warm-season precipitation event. *Wea. Forecasting*, **21**, 149–166, doi:10.1175/WAF909.1.

——, N. Bei, R. Rotunno, C. Snyder, and C. C. Epifanio, 2007: Mesoscale predictability of moist baroclinic waves: Convection-permitting experiments and multistage error growth dynamics. *J. Atmos. Sci.*, **64**, 3579–3594, doi:10.1175/JAS4028.1.

——, Y. Weng, Y.-H. Kuo, J. S. Whitaker, and B. Xie, 2010: Predicting Typhoon Morakot's catastrophic rainfall with a convection-permitting mesoscale ensemble system. *Wea. Forecasting*, **25**, 1816–1825, doi:10.1175/2010WAF2222414.1.

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 06:33 PM UTC