Multiresolution Ensemble Forecasts of an Observed Tornadic Thunderstorm System. Part I: Comparsion of Coarse- and Fine-Grid Experiments

FANYOU KONG

Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma

KELVIN K. DROEGEMEIER

Center for Analysis and Prediction of Storms, and School of Meteorology, University of Oklahoma, Norman, Oklahoma

NICKI L. HICKMON*

Warning Decision Training Branch, NOAA, Norman, Oklahoma

(Manuscript received 4 October 2004, in final form 19 July 2005)

ABSTRACT

Using a nonhydrostatic numerical model with horizontal grid spacing of 24 km and nested grids of 6- and 3-km spacing, the authors employ the scaled lagged average forecasting (SLAF) technique, developed originally for global and synoptic-scale prediction, to generate ensemble forecasts of a tornadic thunderstorm complex that occurred in north-central Texas on 28–29 March 2000. This is the first attempt, to their knowledge, in applying ensemble techniques to a cloud-resolving model using radar and other observations assimilated within nonhorizontally uniform initial conditions and full model physics. The principal goal of this study is to investigate the viability of ensemble forecasting in the context of explicitly resolved deep convective storms, with particular emphasis on the potential value added by fine grid spacing and probabilistic versus deterministic forecasts. Further, the authors focus on the structure and growth of errors as well as the application of suitable quantitative metrics to assess forecast skill for highly intermittent phenomena at fine scales.

Because numerous strategies exist for linking multiple nested grids in an ensemble framework with none obviously superior, several are examined, particularly in light of how they impact the structure and growth of perturbations. Not surprisingly, forecast results are sensitive to the strategy chosen, and owing to the rapid growth of errors on the convective scale, the traditional SLAF methodology of age-based scaling is replaced by scaling predicated solely upon error magnitude. This modification improves forecast spread and skill, though the authors believe errors grow more slowly than is desirable.

For all three horizontal grid spacings utilized, ensembles show both qualitative and quantitative improvement relative to their respective deterministic control forecasts. Nonetheless, the evolution of convection at 24- and 6-km spacings is vastly different from, and arguably inferior to, that at 3 km because at 24-km spacing, the model cannot explicitly resolve deep convection while at 6 km, the deep convection closure problem is ill posed and clouds are neither implicitly nor explicitly represented (even at 3-km spacing, updrafts and downdrafts only are marginally resolved). Despite their greater spatial fidelity, the 3-km grid spacing experiments are limited in that the ensemble mean reflectivity tends to be much weaker in intensity, and much broader in aerial extent, than that of any single 3-km spacing forecast owing to amplitude reduction and spatial smearing that occur when averaging is applied to spatially intermittent phenomena. The ensemble means of accumulated precipitation, on the other hand, preserve peak intensity quite well.

Although a single case study obviously does not provide sufficient information with which to draw general conclusions, the results presented here, as well as those in Part II (which focuses solely on 3-km grid spacing experiments), suggest that even a small ensemble of cloud-resolving forecasts may provide greater skill, and greater practical value, than a single deterministic forecast using either the same or coarser grid spacing.

* Current affiliation: Oklahoma Climatological Survey, Norman, Oklahoma.

Corresponding author address: Dr. Fanyou Kong, Center for Analysis and Prediction of Storms, University of Oklahoma, Rm. 1110 SEC, 100 E. Boyd St., Norman, OK 73019.
E-mail: fkong@ou.edu

1. Introduction

Ensemble forecasting—or the creation of multiple, concurrently valid forecasts from slightly different initial conditions, from different models, from the same model initialized at different times, and/or via the use of different physics options within the same or multiple models—has become the cornerstone of medium-range (6–10 days) operational global numerical weather prediction (NWP) (e.g., Kalnay 2003). Extension of this methodology to the regional scale (1–3 days), frequently referred to as short-range ensemble forecasting (SREF), has been underway for some time (e.g., Brooks et al. 1995; Du and Tracton 2001; Hamill et al. 2000; Hou et al. 2001), though for global models within a framework of hydrostatic dynamics and horizontal grid spacing that cannot explicitly resolve convective clouds.

In the context of nonhydrostatic cloud-resolving models, considerable attention has been directed during the past decade toward the explicit numerical prediction of convective storms using fine-scale observations from Doppler radars and other sensing systems (e.g., Droegemeier et al. 1996; Xue et al. 1996; Carpenter et al. 1997, 1999; Droegemeier 1997; Sun and Crook 1998; Weygandt et al. 1998; Crook and Sun 2002, 2004; Xue et al. 2003; Alberoni et al. 2003). However, the numerous sensitivities evident at small scales (e.g., Brooks et al. 1992; Crook 1996; Hu and Xue 2002; Adlerman and Droegemeier 2002; Martin and Xue 2004; Dawson and Xue 2006) strongly suggest that probabilistic, rather than deterministic approaches, will be required for practicable storm-scale NWP.

Initial efforts directed toward comparing fine-grid forecasts from nonhydrostatic models against coarse-grid ensemble forecasts from hydrostatic models began several years ago with the Storm and Mesoscale Ensemble Experiment (SAMEX) (Hou et al. 2001). Four different models, operating at grid spacings of approximately 30 km, were used to generate a total of 25 forecasts per day, with no finer-grid forecasts created owing to technical difficulties. In some cases, perturbations were applied to initial conditions while in others, multiple physics parameterizations were used. Not surprisingly, the grand ensemble of 25 forecasts exhibited quantitative skill far superior to any of the ensembles generated by a given model (Hou et al. 2001). SAMEX also demonstrated the importance of specifying lateral boundary condition perturbations in a manner consistent with those applied to the interior of the domain.

Although studies in recent years have explored storm-scale ensemble forecasting (hereafter SSEF) in a simple context via the use of cloud models initialized with horizontally homogeneous environments and thermal impulses to trigger convection (e.g., Sindic-Rancic et al. 1997; Elmore et al. 2002a,b, 2003), full-physics storm-scale ensembles that include terrain, horizontally varying initial conditions, and the assimilation of real observations—particularly from Doppler radar—have yet to be attempted. The present paper takes the first step in that direction, building upon the work of Levit et al. (2004) by applying the scaled lagged average forecasting (SLAF) technique (Ebisuzaki and Kalnay 1991), suitably modified, to a tornadic thunderstorm complex that occurred in north-central Texas on 28 March 2000 (Xue et al. 2003). Specifically, we use multiple nesting (grid spacings of 24, 6, and 3 km) within the Advanced Regional Prediction System (ARPS; Xue et al. 2000, 2001, 2003) to produce a five-member ensemble on each of the grids and perform a variety of quantitative comparisons against available observations.

Owing to the lack of a unique method for linking the three grids, we explore several approaches and also examine solution sensitivity to physics options and other parameter variations. Our principal goal is to investigate the viability of ensemble forecasting in the context of explicitly resolved deep convective storms with particular emphasis on the potential value added by fine grid spacing and probabilistic versus deterministic forecasts. Although a single case study obviously does not provide sufficient information with which to draw general conclusions, the present work represents a first step. An additional goal is to study the structure and growth of errors, as well as the application of suitable quantitative metrics, to assess forecast skill for highly intermittent phenomena. In Part II of this paper (Kong et al. 2006, manuscript submitted to *Mon. Wea. Rev.*, hereafter Part II), we focus exclusively on 3-km grid spacing forecasts and examine a number of strategies for creating their ensembles. Further, we perform detailed quantitative verification against rain gauge–calibrated precipitation estimates from the Fort Worth, Texas, Weather Surveillance Radar-1988 Doppler (WSR-88D) and examine the impact of radar data on ensemble skill.

Section 2 provides an overview of the tornadic storm case while section 3 describes the model configuration and experiment design. Ensemble perturbation structure and error growth are examined in section 4, and qualitative performance as well as quantitative skill are assessed in section 5. We conclude in section 6 with a summary and outlook for future work.

2. The Fort Worth tornadic thunderstorm system

We chose the 28–29 March 2000 Fort Worth, Texas, tornadic thunderstorm system because it is well documented and has been used successfully in fine-grid forecast experiments with the ARPS (Xue et al. 2003). The storm complex, shown by the KFWS WSR-88D [Next-Generation Weather Radar (NEXRAD)] Doppler radar images in Fig. 1, produced two tornadoes between 0015 and 0115 UTC on 29 March 2000, one of which traversed the metropolitan Fort Worth area from 0018 to 0028 UTC, causing two deaths, many injuries, and extensive damage to buildings. The second tornado passed through Arlington and Grand Prairie, Texas, between 0105 and 0115 UTC. Torrential rain produced flooding, and softball-sized hail caused three additional casualties. Total storm damage was estimated at $450 million (NCDC 2000).

The synoptic setting for this event consisted of a Pacific trough that moved quickly inland during the preceding 24 h, and at 0000 UTC on 29 March, the trough axis was located over the Texas panhandle, with winds approaching 50 kt at 500 hPa over Fort Worth (Fig. 2a). As this trough brought cooler air southward and eastward through the plains, a surface low in the Texas panhandle on the morning of 28 March helped draw warm, moist low-level air from the Gulf of Mexico northward into central Texas. The National Centers for Environmental Prediction (NCEP) operational Rapid Update Cycle (RUC) analysis (not shown) depicted a dryline bulging eastward over north-central Texas by 2100 UTC, with a maximum in CAPE over the same region (Fig. 2b). The NCEP Eta Model (not shown) predicted no precipitation south of the Red River (in north-central Texas) in the 12 h prior to 0000 UTC on 29 March, in large part because of its relatively coarse horizontal grid spacing (32 km) and hydrostatic dynamics. However, forecasters were well aware of the likelihood of severe weather, and the National Oceanic and Atmospheric Administration (NOAA) Storm Prediction Center issued tornado watches for north Texas beginning at 2053 UTC, more than 3 h prior to the Fort Worth tornado.

3. Model configuration and experiment design

a. General approach and model configuration

The design of our experiments is guided by the fact that SSEF differs from traditional global or short-range ensemble forecasting in several ways. First, in light of available computing resources, the extremely fine grid spacing required by SSEF necessitates the use of multiple nested grids. This complicates the construction of initial perturbations because no unique strategy exists for linking multiple grids, as described below. Second, large domains generally are desirable for reducing the impact on spread, within small domains, of lateral boundary conditions (see Nutter et al. 2004), but this comes at the expense of fine grid spacing, which is particularly important in SSEF. For example, in global ensemble forecasting, the grid spacing or spectral truncation of ensemble members is coarser than that of the control run, usually by a factor of 3–5. Such coarsening, however, is not feasible in SSEF, where individual convective elements in a control run using 2-km grid spacing, for example, likely would be unresolved using a spacing of 6–10 km for the ensembles. Finally, techniques used in hydrostatic global and regional models for generating the most rapidly growing modes (e.g., Hamill et al. 2000) may not be applicable at the convective scale owing to intrinsic differences in dynamics and energetics. As discussed below, a modified form of the SLAF technique is applied here to partly account for such differences.

The model used in this study, ARPS, is a three-dimensional, nonhydrostatic compressible numerical weather prediction system (Xue et al. 2000, 2001, 2003) with comprehensive physics and a self-contained data ingest, quality control, retrieval, and assimilation system (Xue et al. 2003). In this study we employ grid nesting (Fig. 3) using 24-, 6-, and 3-km spacing for the coarse-, medium-, and fine-grid domains, respectively. The 3-km grid spacing domain (hereafter 3-km domain) is centered over Fort Worth with sufficient coverage to contain the principal features of interest while maintaining some distance from the lateral boundaries. All grids use 53 terrain-following vertical layers, with nonlinear stretching, via a hyperbolic tangent function, that yields a spacing of 20 m at the ground that expands to approximately 800 m at the top of the domain (located at approximately 20-km altitude). A more complete list of model parameters and options used is provided in Table 1.

b. Construction of perturbations in the 24-km grid forecasts

Several approaches are available for creating ensemble initial conditions, including Monte Carlo (random perturbations) (e.g., Mullen and Baumhefner 1989), breeding of growing modes (e.g., Toth and Kalnay 1993, 1997), lagged average forecasting (e.g., Hoffman and Kalnay 1983), singular vectors (e.g., Hamill et al. 2000), physics perturbations (e.g., Stensrud et al. 2000), and ensemble Kalman-filter-based techniques (e.g., Houtekamer and Mitchell 1998; Hamill and Snyder 2000; Wang and Bishop 2003). These methods have

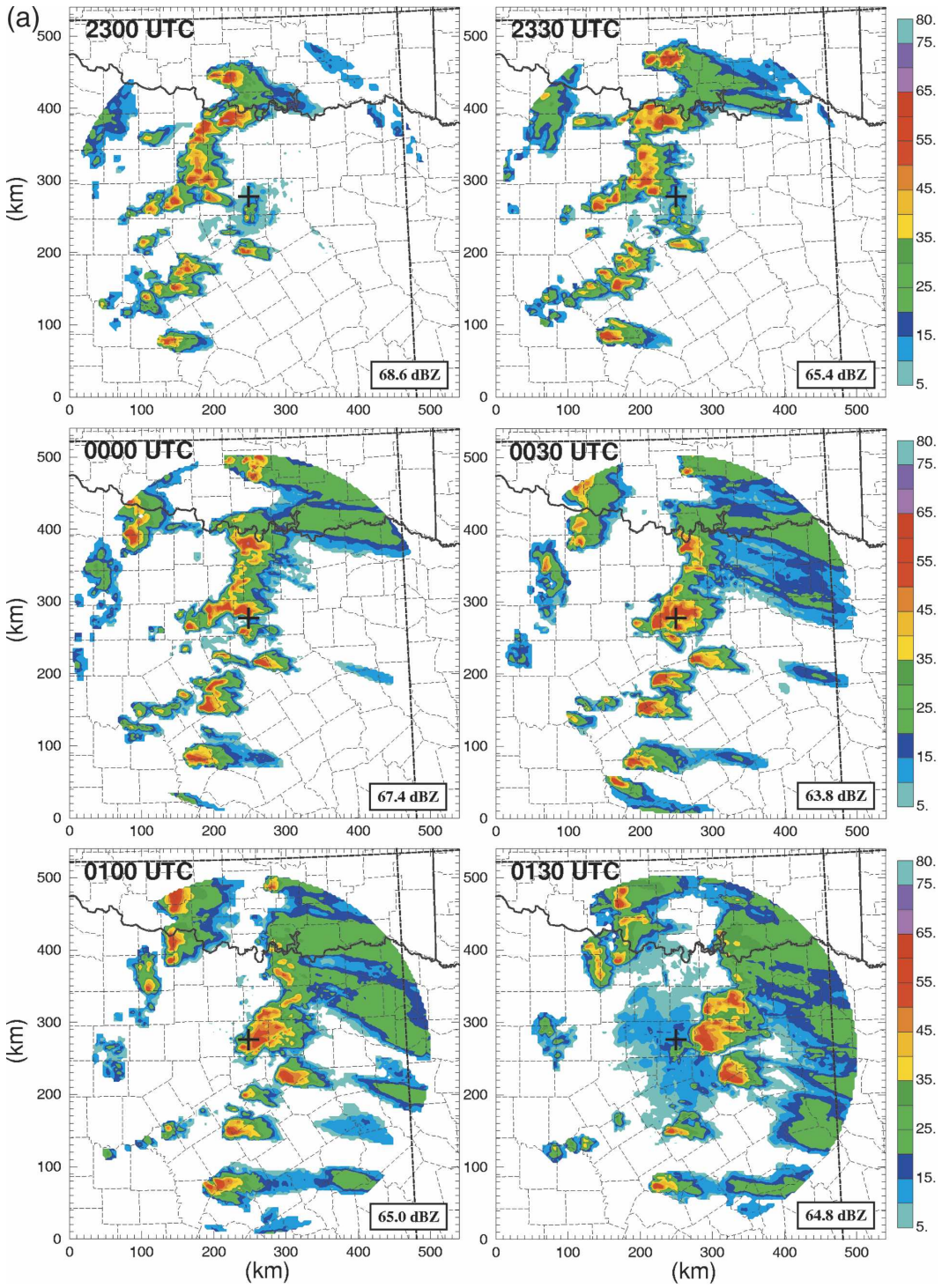

FIG. 1. Fort Worth (KFWS) WSR-88D lowest-elevation-angle reflectivity every 30 min from 2300 UTC on 28 Mar through 0230 UTC on 29 Mar 2000. The time (UTC) is shown in upper-left corner of each panel, and the "+" mark shows the location of Fort Worth, TX.

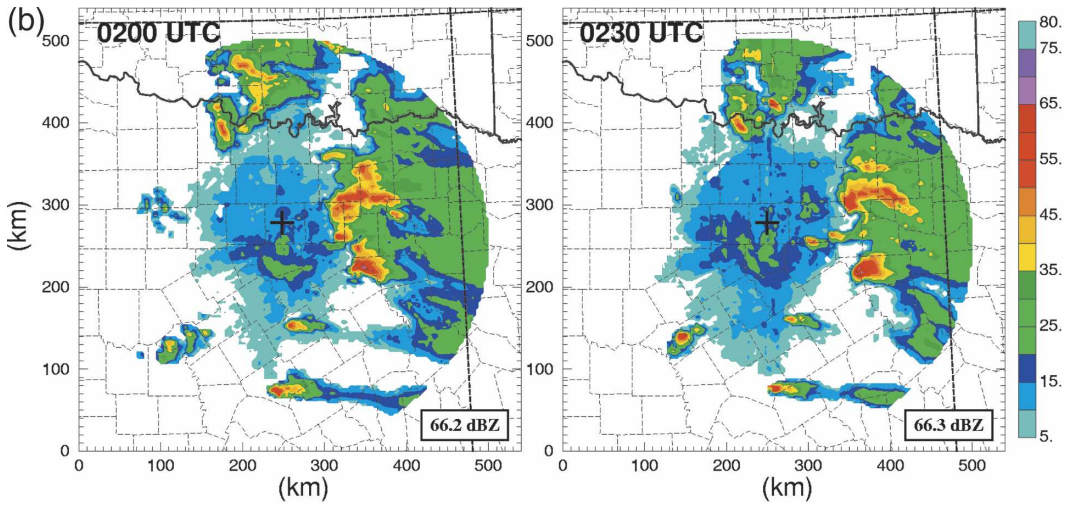

FIG. 1. (*Continued*)

been applied with success to large-scale hydrostatic NWP models but may not directly be applicable to nonhydrostatic SSEF. Nevertheless, because this study represents a first attempt to study SSEF in the context of an observed storm system using full-physics NWP, we employ a modified version of an existing large-scale technique, the SLAF method (Ebisuzaki and Kalnay 1991; Hou et al. 2001), using a single model and a fixed set of physics options for each nested domain. Toth and Kalnay (1993) showed that SLAF perturbations represent realistic short-term forecast errors stemming from errors in the analysis. Consequently, the errors grow faster than their Monte Carlo counterparts in global models yet slower than bred or singular vectors. A comparison of performance among methods is beyond the scope of the present study but is being examined using idealized cloud-scale experiments to be reported on subsequently.

For each of the three grids, we generate a five-member SLAF ensemble consisting of one control forecast and four perturbed members. The latter are constructed by forming the difference fields between a previous ARPS forecast and a verifying analysis, followed by appropriate scaling. We then add and subtract these scaled difference fields from the same verifying analysis (which serves as the initial condition for the control run) to form two (paired) ensemble members. A five-member SLAF thus requires two previous ARPS forecasts. Although a larger ensemble no doubt would produce more reliable statistics (e.g., Toth and Kalnay 1997; Hou et al. 2001), our experience with idealized cloud-scale simulations suggests that five members is adequate, perhaps minimally so, for achieving the goals of this initial study while keeping at a manageable level the complexity of the experiment design and required computing resources.

Figure 4 shows the general procedure outlined above applied to the 24-km ensemble. Symbols P1 and P2 represent ARPS forecasts initiated at 0600 and 0000 UTC, respectively, on 28 March 2000, while forecast P0 is initiated at 1200 UTC in the same manner and serves as the *control run*. Initial conditions for P0-P2 are generated using the ARPS Data Assimilation System (ADAS) (Brewster 1996) with the corresponding NCEP Eta analysis as the background state.[1] (Note that in a continuous operational cycle, the background field and ensembles likely would be created by the same model.) Observations used by ADAS include standard surface data, wind profiler, and rawinsonde data, Aircraft Communications Addressing and Reporting System (ACARS) commercial aircraft wind and temperature data, Geostationary Operational Environmental Satellite (GOES) visible and infrared satellite data, and Oklahoma Mesonet data. No radar data are used in the 24-km domain.

Ensemble members S1 and S2 are generated by subtracting and adding, respectively, the difference fields between the 6-h forecast from P1, valid at 1200 UTC, and the 1200 UTC ADAS assimilated analysis that is used to initiate forecast P0. This set of difference fields, labeled *perturbation 1* in Fig. 4, also serves as a reference for perturbation amplitudes. Ensemble members S3 and S4, with scaled perturbations (*perturbation 2*) subtracted and added, respectively, likewise are created

[1] The operational Eta Model utilized a horizontal grid spacing of 32 km but the gridded binary (GRIB) files used here had a spacing of 40 km.

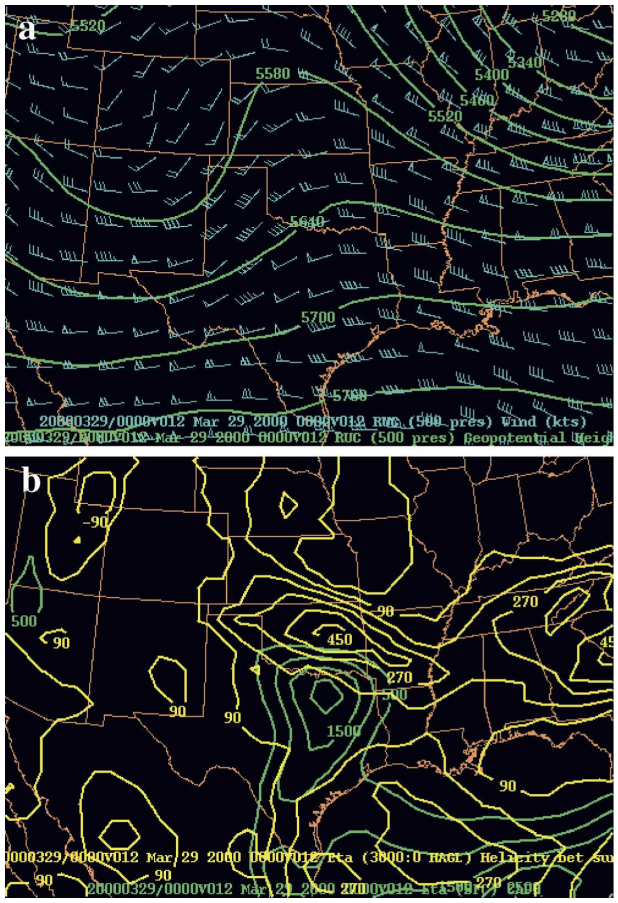

FIG. 2. Fields valid at 0000 UTC on 29 Mar 2000: (a) 500-hPa winds and geopotential heights from the RUC analysis and (b) storm-relative environmental helicity and CAPE from the eta analysis. Images provided courtesy of COMET Case Study.

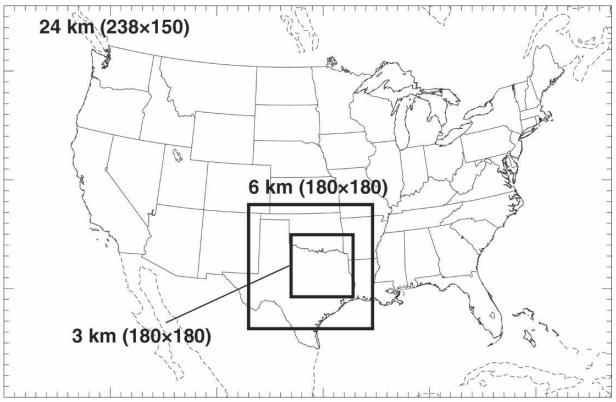

FIG. 3. Nested grid configuration used in the ARPS for all experiments. Horizontal grid spacing and array dimensions are shown for each grid.

from P2 and P0. If we were to follow the traditional SLAF technique used in medium-range ensemble forecasting (Ebisuzaki and Kalnay 1991), in which perturbations are scaled by their age, then a scaling factor of unity should be used for perturbations S1 and S2 and a factor of 0.5 for perturbations S3 and S4. However, the linear error growth assumption associated with this scaling method does not hold true in the present study, especially for deep convection at fine grid spacings. Consequently, we amend the SLAF methodology and scale according to the reference amplitudes described above, using as metrics the domain-wide rms difference of the three wind components along with potential temperature, pressure, and water vapor mixing ratio. With that accomplished (details are provided in section 4), each of the five ensemble members (P0, S1, S2, S3, and S4) is integrated for 18 h using the same model configuration, with lateral boundaries perturbed in a manner consistent with the initial conditions.

c. *Construction of perturbations on nested grids*

Grid nesting complicates the construction of ensembles in SSEF because no unique strategy exists for linking fine grids to their coarser parents. Further, in an operational context, fine grids might be spawned irregularly in time and space, thus precluding the assurance that fine-grid background fields from previous forecasts will be available on a regular basis. Assuming that a coarse-grid forecast always is available, one could generate perturbations from it and then interpolate them to finer grids. This approach could be applied to both SLAF and breeding methods, though an obvious drawback is the lack of dynamical consistency among grids, including potentially the use of different physics options at different grid spacing.

With that in mind, and to begin addressing the many questions associated with the construction of ensembles using multiple nested grids, we employ a single strategy in the experiments described herein and examine two additional strategies in Part II. None is intended to represent an optimal methodology, yet each produces different error growth structures and forecast skill.

Figure 5 shows a nesting procedure under the assumption that previous forecasts from only the coarsest grid are available. As indicated by the uppermost curved arrow, a 6-km control forecast (labeled cn6) is initiated at 1800 UTC on 28 March 2000 using the 6-h, 24-km grid ARPS forecast, interpolated to 6 km, as the background state. WSR-88D level III reflectivity and other observations are assimilated by ADAS into this 6-km analysis valid at 1800 UTC. Owing to the absence of 6-km forecasts prior to 1800 UTC, ensemble perturbations at 6-km grid spacing are constructed by first interpolating onto the 6-km grid the two previous 24-km ARPS forecasts valid at 1800 UTC (i.e., P1 and P2),

TABLE 1. Physical and computational parameters used in the simulations.

Parameters	Value
Horizontal grid spacing (array size) for grid 1	24 km (238 × 150 points)
Horizontal grid spacing (array size) for grid 2	6 km (180 × 180 points)
Horizontal grid spacing (array size) for grid 3	3 km (180 × 180 points)
Vertical grid stretching function	Hyperbolic tangent
Vertical grid spacing (number of levels)	20–800 m (53 levels)
Large time step: grid 1/grid 2/grid 3	20/10/5 s
Small time step: grid 1/grid 2/grid 3	20/10/5 s
Fourth-order mixing coefficient for grid 1	5×10^{-4} s^{-1}
Fourth-order mixing coefficient for grids 2 and 3	3×10^{-4} s^{-1}
Nondimensional divergence damping coefficient	0.05
Rayleigh damping coefficient (applied above 12 km only)	3.33×10^{-3} s^{-1}
Lateral boundary conditions	Externally forced, linear time interpolation
Top and bottom boundary conditions	Rigid wall with upper sponge
Horizontal and vertical advection scheme	Fourth-order with leapfrog time step
Cumulus parameterization—grids 1 and 2	Kain–Fritsch
Microphysics	Lin–Tao five-category ice scheme
Turbulence parameterization	1.5-order turbulent kinetic energy (TKE) closure
Radiation parameterization	Shortwave: Chou (1990); longwave: Chou and Suarez (1994)
Land surface and vegetation scheme	Noilhan and Planton (1989); Pleim and Xiu (1995)

scaling their difference from the analysis (cn6), and then adding this difference to, and subtracting it from, the new analysis at 1800 UTC to produce four ensemble members: s1 and s3 having negative perturbations, along with s2 and s4 having positive perturbations. For both the 24- and 6-km ensembles, the Kain–Fritsch cumulus parameterization scheme (Kain and Fritsch 1993) and the Lin–Tao explicit five-category ice-phase microphysics scheme (Lin et al. 1983; Tao and Simpson 1993) are used.

The 3-km ensemble, initiated at 2300 UTC, follows a similar procedure. The control member (cn3) uses as a background state the 5-h ARPS forecast from experiment cn6, interpolated to 3 km, into which are assimilated WSR-88D level III reflectivity and radial wind data, along with other observations noted previously. Two pairs of ensemble perturbations at 3 km are constructed by first interpolating P1 and P2, valid at 2300 UTC, onto the 3-km grid, scaling their difference from the control run (cn3), and then adding this difference to and subtracting it from the new 3-km analysis at 2300 UTC to produce four perturbed members (s1, s2, s3, and s4). As discussed below and in contrast to the 24- and 6-km spacing experiments, observations are assimilated into both the control *and* perturbation experiments at 3-km spacing.

4. Scaling of perturbations

Examining perturbation size and error growth rate for each model grid reveals that traditional SLAF scaling, which normalizes perturbation amplitude based upon the age of the forecast used to generate the perturbation, does not apply for our case. To illustrate, Figs. 6 and 7 show the 500-hPa and near-surface poten-

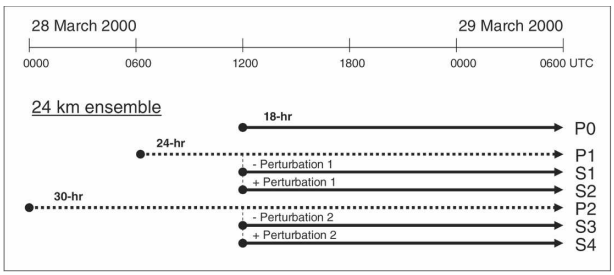

FIG. 4. Construction of SLAF ensembles for the 24-km grid spacing domain.

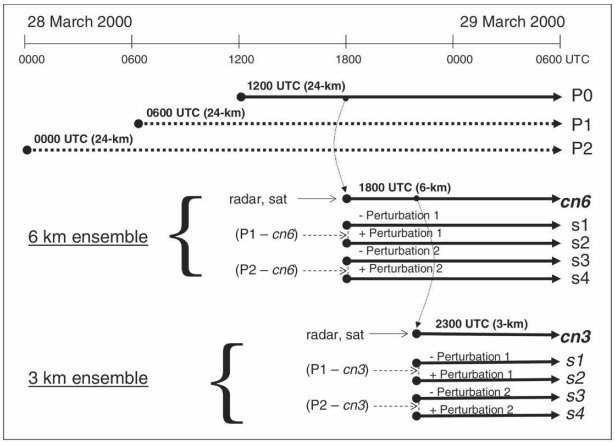

FIG. 5. Construction of SLAF ensembles for the 6- and 3-km grid spacing domains.

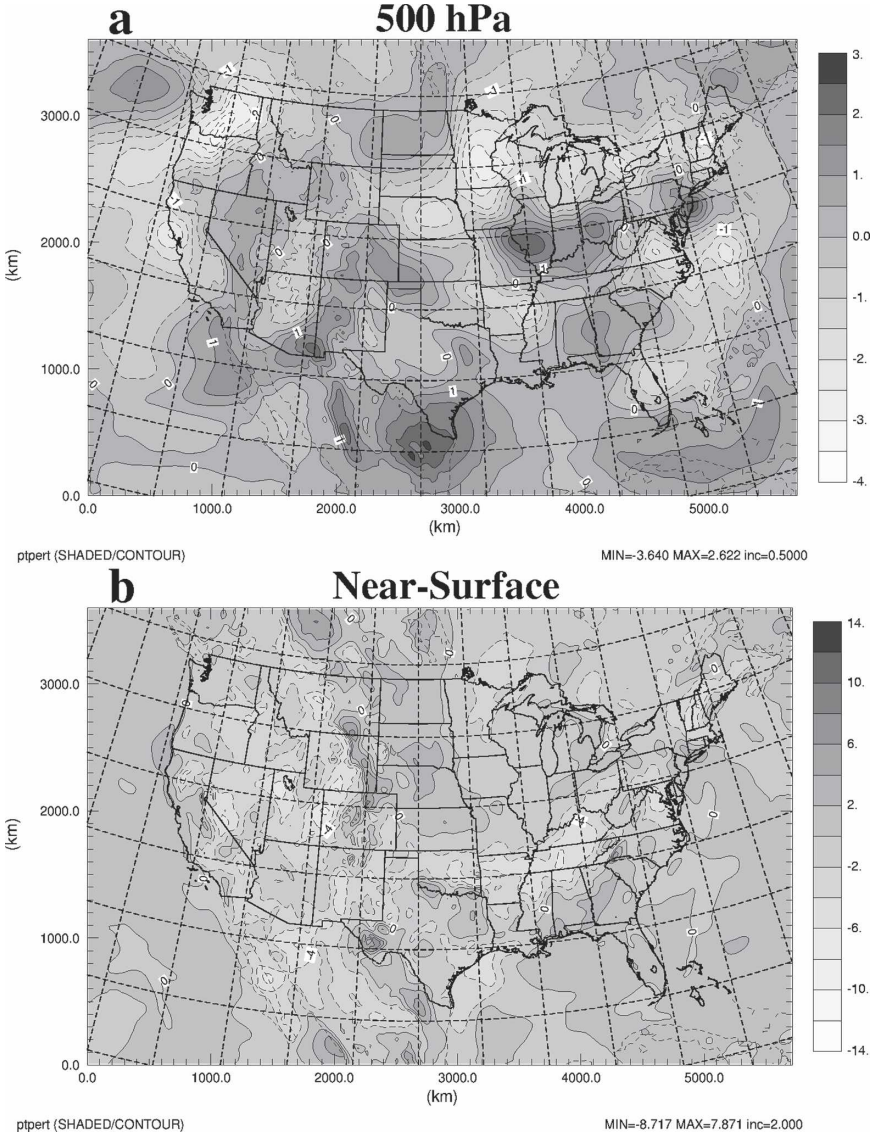

FIG. 6. Potential temperature for linearly scaled perturbation 1 (P1 – P0) at (a) 500 hPa and (b) near surface ($k = 5$) for the 24-km grid spacing ensemble. Values are °C.

tial temperature for ensemble perturbation 1 (forecast P1 minus P0 analysis) and perturbation 2 (forecast P2 minus P0 analysis), respectively, at 1200 UTC on 28 March 2000 from the 24-km ensemble. Both perturbations are scaled linearly with forecast age. The amplitudes in perturbation 2 (Fig. 7) are roughly half those of perturbation 1 (Fig. 6), instead of the same, as anticipated with traditional SLAF. Table 2 lists selected rms errors for the two perturbations *before* scaling, which confirms that perturbations from the previous 6-h forecast (P1 minus P0) and previous 12-h forecast (P2 minus P0) have the same amplitude, except for potential temperature.

Figure 8 shows rms errors for the 24-km unperturbed forecast (P2) versus the corresponding ADAS analysis. Most errors grow fast initially but then become relatively constant after 6 h. This suggests that error growth in limited-area mesoscale models reaches saturation much faster than in their global counterparts,[2] as has been shown in previous studies, presumably because of lateral boundary condition effects (Warner et al. 1997; Nutter et al. 2004). Other nested domain forecasts exhibit similar behavior, as shown in Part II.

Based upon these results, linear scaling between two

[2] The early peak in rms error, like that exhibited in the spread (Fig. 9), likely results from an imbalance in the initial conditions.

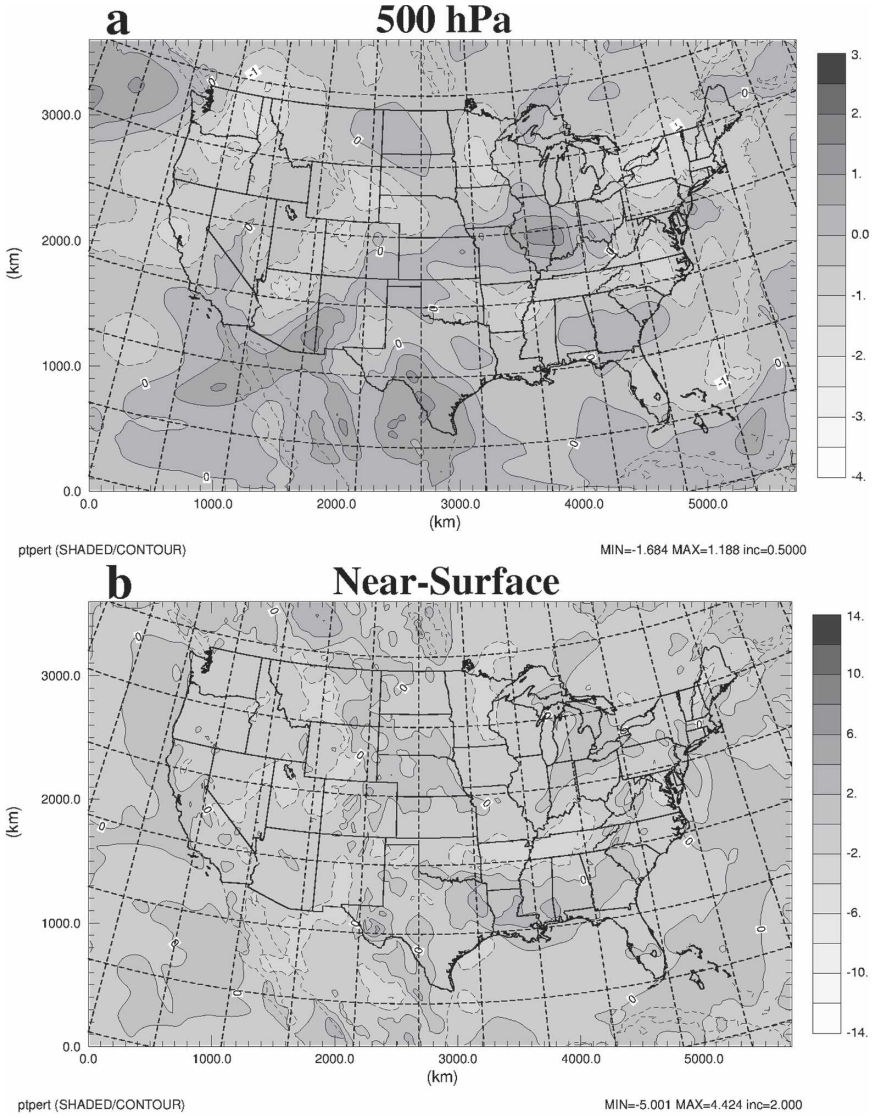

FIG. 7. As in Fig. 6 but for perturbation 2.

previous forecasts 6 h apart, based upon age, is insufficient to maintain equity between the two perturbation amplitudes. We thus replace the traditional age-based SLAF scaling with an amplitude-based scaling, provided that the previous forecasts used in constructing the initial perturbation have reached saturation. Specifically, using perturbation 1 as a reference, we scale perturbation 2 based upon the actual amplitude ratios between them. Although this procedure is, strictly speaking, no longer SLAF as originally defined (Ebisuzaki and Kalnay 1991), we continue using this acronym for simplicity because scaling remains a fundamental component of the methodology. It is important to note that our scaling is applied independent of wavelength whereas in reality, the power spectrum of errors is nonuniform and a function of both time and space.

Figures 9–11 present comparisons of 24-km ensembles using the traditional SLAF and revised amplitude-based scaling methods. For the latter scaling, the

TABLE 2. Domain-wide rms errors of selected initial perturbations from the 24-km ensemble forecast prior to scaling.

	Perturbation 1 (P1 – P0)	Perturbation 2 (P2 – P0)
U (m s^{-1})	3.60	3.50
V (m s^{-1})	3.42	3.51
θ (K)	1.28	2.06
q_v (g kg^{-1})	0.60	0.60
P (hPa)	1.19	1.29

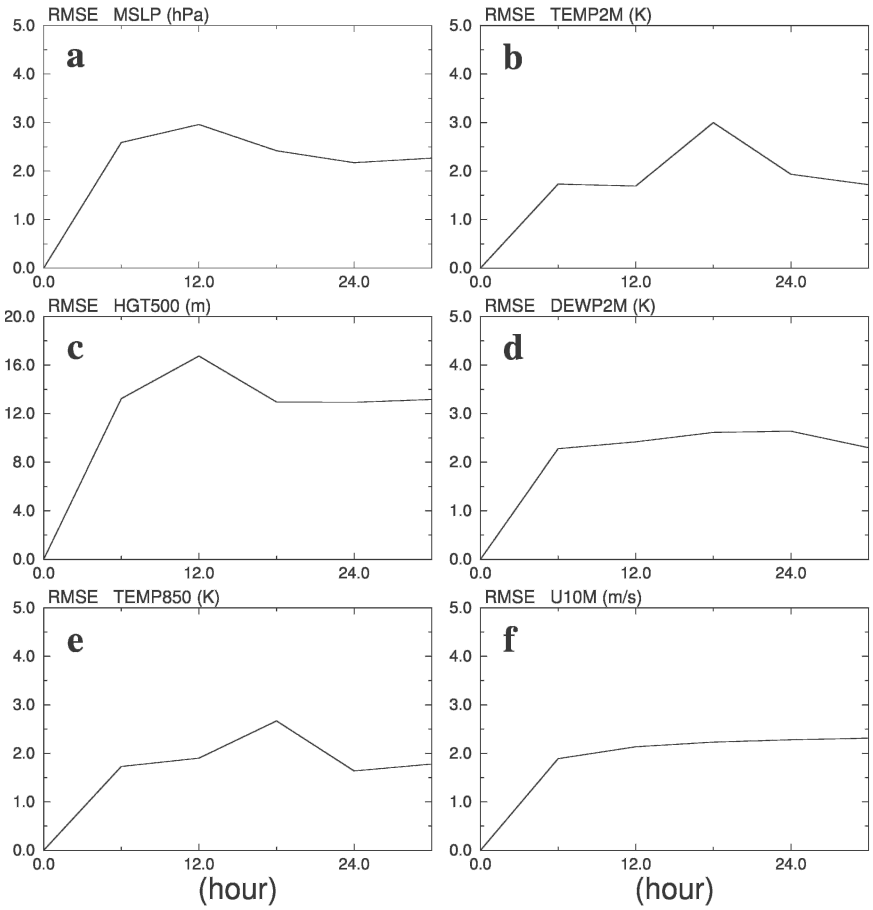

FIG. 8. Rms error between the 24-km unperturbed forecast and corresponding ADAS analysis, computed at 6-h intervals beginning at 0000 UTC on 28 Mar 2000. Fields shown are (a) mean sea level pressure (hPa), (b) 2-m air temperature (K), (c) 500-hPa geopotential height (m), (d) 2-m dewpoint (°C), (e) 850-hPa temperature (K), and (f) 10-m zonal wind (m s^{-1}).

domain-mean ensemble spreads (defined according to Hou et al. 2001) are larger (Fig. 9), and the alignment of the western edge of the 6-h accumulated precipitation region over Fort Worth agrees better with observations, as does the limited eastward extent of precipitation near the Oklahoma–Texas border (Figs. 10 and 11; cf. Fig. 12).

Toth and Kalnay (1997) showed that ensemble skill is sensitive to the amplitude of the initial perturbation. Consequently, in addition to the scaling factor of 1.0 applied to perturbation 1, we ran other experiments in which scaling factors of 0.5, 1.5, and 2.0 were applied to form new reference values. The results, shown in Fig. 13, demonstrate that an increase (decrease) in initial perturbation amplitude leads to an increase (decrease) in ensemble spread (note that the 1X curve is identical to the amplitude-scaled curve in Fig. 9). The impacts of initial perturbation amplitude are discussed in the next section, and in Part II.

5. Results

In this section we present ensemble precipitation forecasts from each nested grid and compare them with available observations. The predicted radar reflectivity for the 3-km grid also is compared against observations, taking into account known differences between the two (Smedsmo et al. 2005).[3] A more detailed analysis of the 3-km grid experiments, including additional forecast configurations, is presented in Part II.

[3] By comparing vertical profiles of reflectivity produced by a 1-km grid spacing ARPS forecast against WSR-88D level II volume scan data for the same Fort Worth tornado case as in this paper, Smedsmo et al. (2005) concluded that a direct comparison between modeled and observed radar reflectivity is problematic. Their study found that significant differences exist between mean reflectivity profiles below the freezing level, due partly to the limitation of correctly predicting frozen hydrometeor species. The model reflectivity field was much less variable than the observed field, differing by 24% in std dev.

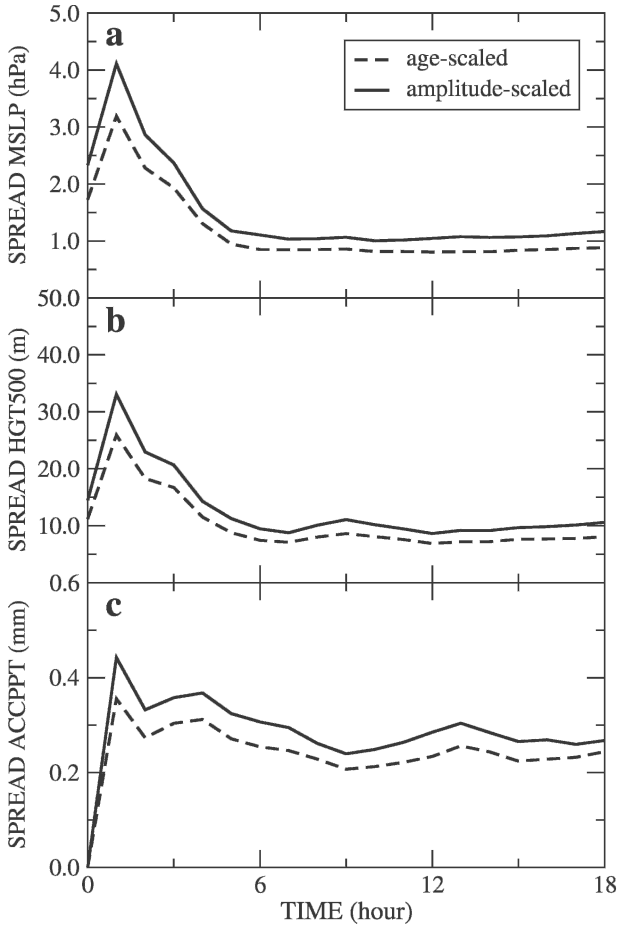

FIG. 9. Domain-wide mean forecast spread of (a) mean sea level pressure, (b) 500-hPa geopotential height (m), and (c) accumulated precipitation (mm) from the S3 and S4 24-km ensemble for perturbations that are age-scaled (dash) and amplitude-scaled (solid).

a. 24-km ensemble forecast

Six-hour accumulated precipitation from the five 24-km grid ensemble members, valid at 0000 on UTC 29 March (near the time of the Fort Worth tornado), is shown in Fig. 14, along with the ensemble mean. Each member exhibits diversity and captures elements of the principal precipitation features shown in the stage IV analysis (Fig. 12). As expected, members S1 and S3 are similar while S2 and S4 are similar. Precipitation over the eastern and northeastern United States is captured to some extent by all members, though generally with a high bias in both amplitude and area. Forecasts over central and northern Texas show a mostly eastward bias in the overall pattern, suggesting that the predicted features move too fast. The large accumulation over northeast Texas predicted by the control forecast, as well as by members S2 and S3, also does not agree with observations. This discrepancy is reflected in the Eta Model CAPE field in Fig. 2b, in which the CAPE axis is located east of the observed precipitation. The 12-h forecast of surface CAPE in the ARPS control run P0 (not shown) is very similar to that in the Eta.

The ensemble mean forecast at 24-km grid spacing (Fig. 14b) compares more favorably with the stage IV analysis than the control run in terms of the western boundary of the precipitation region in central Texas, though the forecast still contains the region of spurious precipitation in northeastern Texas. The location of precipitation over the Rocky Mountains, and especially the local maximum over central Arizona and southern Montana, are predicted reasonably well by the mean and control forecasts, though again with a high bias in area and amplitude. Precipitation along the Oklahoma–Texas border is more realistic in the mean forecast in that it stretches westward into Oklahoma, due in large part to the contribution from ensemble member S1.

A notable benefit of ensemble forecasting is the capability of generating probability products that quantify the relative frequency of occurrence of a given condition or parameter. This is especially valuable for explicitly resolved deep convection because probabilities highlight the likely occurrence of extreme or intense events. Figure 11 shows the uncalibrated probability of 6-h accumulated precipitation greater than or equal to 2.54 mm based upon the five ensemble members described above using both age and amplitude scaling. The probability at a point is simply the ratio of the number of forecasts meeting a stated criterion divided by five ensemble members. Thus, the probability can only equal 0%, 20%, 40%, 60%, 80%, and 100%. Although five members is arguably the lower limit for a meaningful ensemble size, we do find reasonable qualitative agreement with the stage IV precipitation data shown in Fig. 12. Discrepancies do exist, however, as the ensemble fails to predict precipitation over southwest Oregon and the Texas panhandle. Further, the forecasts place heavy precipitation along the Red River and eastward, into north-central Texas, when in reality (Fig. 12), the accumulation had a north–south orientation with the heaviest values to the south and west of those in the forecast.

To quantify forecast skill, we examine a number of traditional measures including bias and equitable threat scores (ETS),[4] and for probability forecasts the Brier score (BS) and ranked probability score (RPS). Details regarding these scores may be found in Wilks (1995), Stensrud et al. (2000), and Hou et al. (2001). Because

[4] We have not adjusted the ETS for bias.

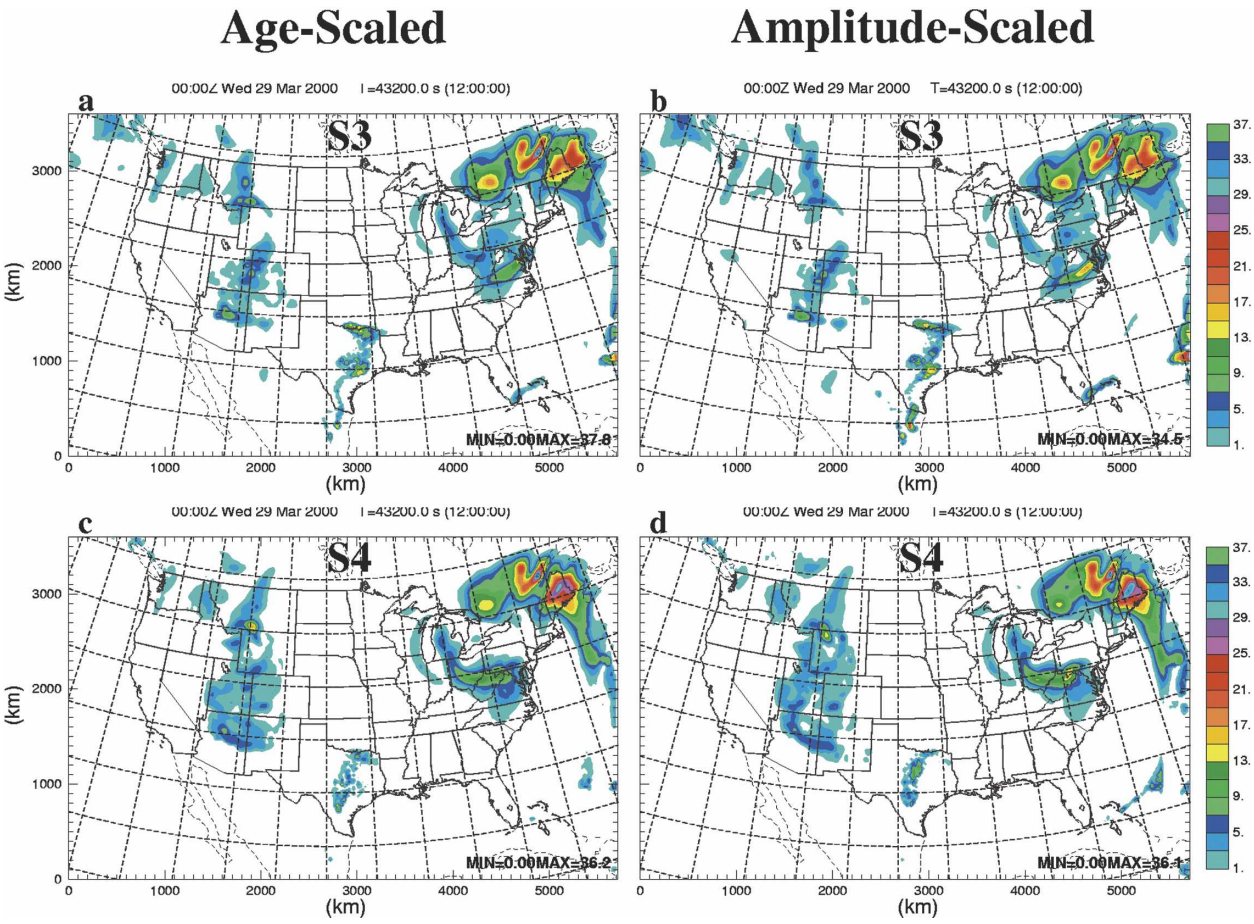

FIG. 10. The 6-h accumulated precipitation (mm) valid at 0000 UTC on 29 Mar 2000 from individual 24-km ensemble members: (a), (c) age-scaled S3 and S4 members, respectively, and (b), (d) amplitude-scaled S3 and S4 members, respectively.

the focus of the present paper is the convective system associated with tornadic storms in north central Texas, we compute all scores for the 24-km ensemble using a verification domain defined by the area of the 6-km nested grid (see Fig. 3). It is important to note that with an ensemble of five members, these scores must be viewed with caution because the confidence intervals, which are not computed because of the small sample size, likely are as large or larger than the differences exhibited. Further, when computing statistics for our three domains (24, 6, and 3 km) using 4-km grid spacing stage IV precipitation data, we linearly average the precipitation data to 24 and 6 km and interpolate them to 3 km.

Tables 3 and 4 show bias and ETS for 6-h accumulated precipitation from the 24-km ensemble using the following thresholds: 0.254, 2.54, 12.7, and 25.4 mm (i.e., 0.01, 0.1, 0.5, and 1.0 in.). The ETS measures skill in predicting the area of precipitation above a given threshold relative to a random forecast, with a skillful forecast having a positive ETS. Because little precipitation falls during the first 6 h of any forecast, we verify the 6-h accumulation for two separate periods (see Fig. 4): the first (12-h forecast) ending at 0000 UTC on 29 March 2000 and the second (18-h forecast) ending at 0600 UTC. The bias score simply measures the ratio of the number of "yes" forecasts to the number of "yes" observations. Because a forecast with no bias has a bias score of unity, Table 3 shows that all members, except S2, underforecast precipitation for the 0.254-mm threshold, while the ensemble mean surpasses all individual members. Bias in the ensemble mean is among the smallest values for the 2.54-mm threshold. Scores for the 25.4-mm threshold and the 12-h values of the 12.7-mm threshold in parentheses correspond to ETS values less than or equal to zero (Table 4), highlighting the fact that for such ETS, the model has no skill.

Table 4 shows that for the lowest two thresholds, ETS in the second 6-h period exhibits more skill than the first. For the 25.4-mm threshold, all values except one are less than or equal to zero. For the 12.7-mm threshold, only the second 6-h forecasts show skill. The ensemble mean is more skillful than the control fore-

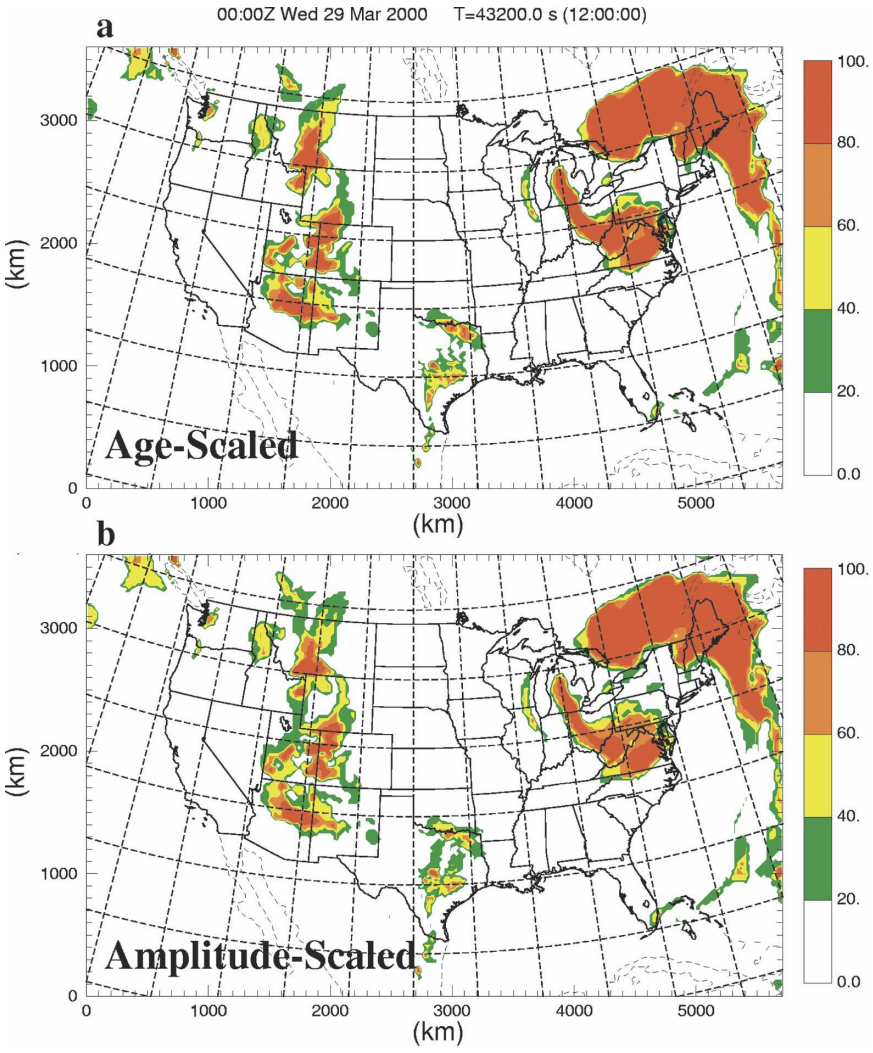

FIG. 11. Probability of 6-h accumulated precipitation greater than or equal to 2.54 mm computed from the 24-km ensemble forecast valid at 0000 UTC on 29 Mar 2000 using (a) age scaling and (b) amplitude scaling.

cast at the lowest precipitation threshold at 12 h, less skillful at 18 h, yet more skillful for the 2.54-mm threshold at both times.

Table 5 shows the BS and RPS from the ensembles using both age and amplitude scaling for selected precipitation thresholds. The BS ranges between zero and unity, while the RPS ranges between zero and the number of categories minus one (i.e., 4 in this study). A perfect probability forecast has both BS = 0 and RPS = 0. Table 5 shows that the amplitude-scaled ensemble is slightly more skillful than the age-scaled ensemble, though unlike ETS, both probability scores show less skill during the second 6-h period than the first. We again urge caution in generalizing these results because of the small ensemble size (and thus lack of statistical confidence testing) and application to a single weather event.

Table 6 shows BS values from the experiments using different scaling factors for the initial perturbation (note that a scaling factor of unity is identical to the 24-km ensemble labeled "amplitude-scaled" in Table 5). Some sensitivity is evident, with scaling factors of 1.0 and 1.5 scoring the best. As expected, RPS scores show a similar behavior. This suggests that scaling factor of unity is probably appropriate for the 24-km ensembles.[5]

In summary, the 24-km ensembles show value in cap-

[5] With finer grid spacing, a scaling factor of 0.5 results in more localized high-intensity precipitation with higher probability but also significant underdispersion. A scaling factor of 1.5 produces much broader precipitation with lower peak probability values, with spreads comparable to the corresponding rmse. Additional results on scaling are presented in Part II.

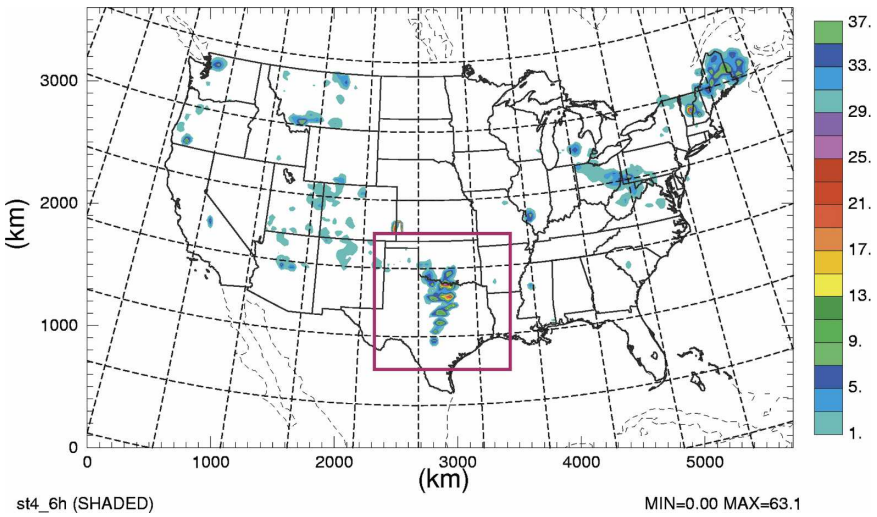

FIG. 12. Stage IV 6-h accumulated precipitation ending at 0000 UTC on 29 Mar 2000. The box indicates the 6-km grid spacing domain.

turing the location and overall pattern of precipitation, with probabilities highlighting regions of greatest intensity even though the model is unable to resolve individual convective storms.[6] Although the forecast precipitation rates over Fort Worth are extremely small relative to those observed by radar (~1 mm h^{-1} versus greater than 200 mm h^{-1}), ensemble probabilities do suggest the presence of deep convection, thus offering potentially useful guidance to forecasters.

b. 6-km ensemble forecasts

Figure 15 shows predicted 6-h accumulated precipitation, valid at 0000 UTC on 29 March 2000, for the 6-km grid spacing ensemble members as well as the ensemble mean. Comparing with the stage IV analysis in Fig. 12 (the small box outlines the 6-km domain), the precipitation from cn6 (control), and from members s1 and s3, covers far too great an area, extending eastward and southeastward beyond the observed locations as in the 24-km ensemble. Further, both show notably spurious precipitation in southeast Texas. This is not surprising because the 24-km forecasts provide the background state for the 6-km ensemble members. Unlike the 24-km ensemble, precipitation from members s1 and s3 is greatly overpredicted in amplitude. In central Texas, the north–south precipitation region in member s4 is aligned more closely in space with that in the stage IV analysis, while that in s2 differs considerably, show-

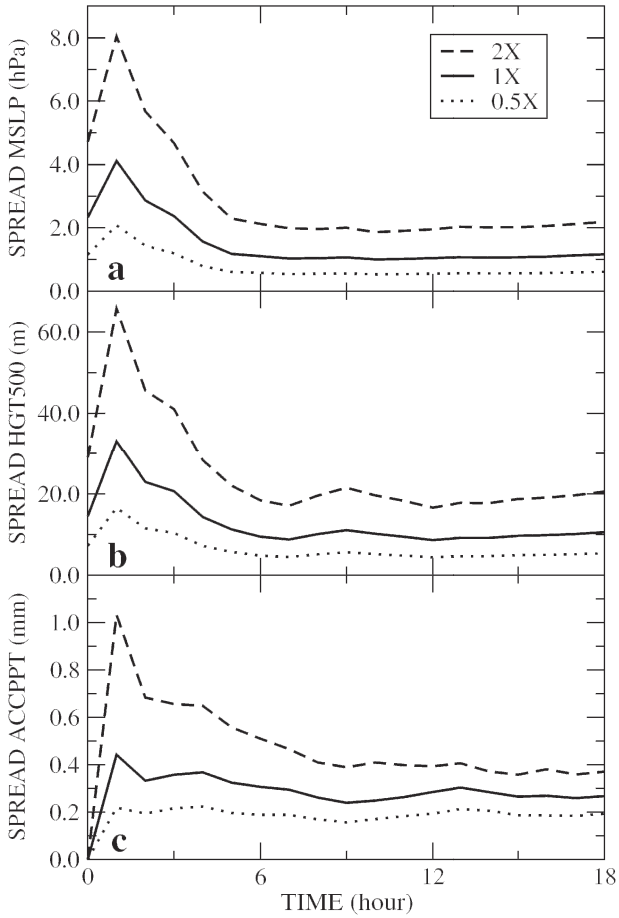

FIG. 13. Domain-wide mean forecast spread of (a) mean sea level pressure, (b) 500-hPa geopotential height (m), and (c) accumulated precipitation (mm) from 24-km ensembles with different factors on the reference magnitude. The 1X curve is identical to the amplitude-scaled one in Fig. 9.

[6] Interestingly, forecasts from ARPS are notably superior to those from the operational Eta Model owing, in the ARPS, to somewhat finer grid spacing and nonhydrostatic dynamics.

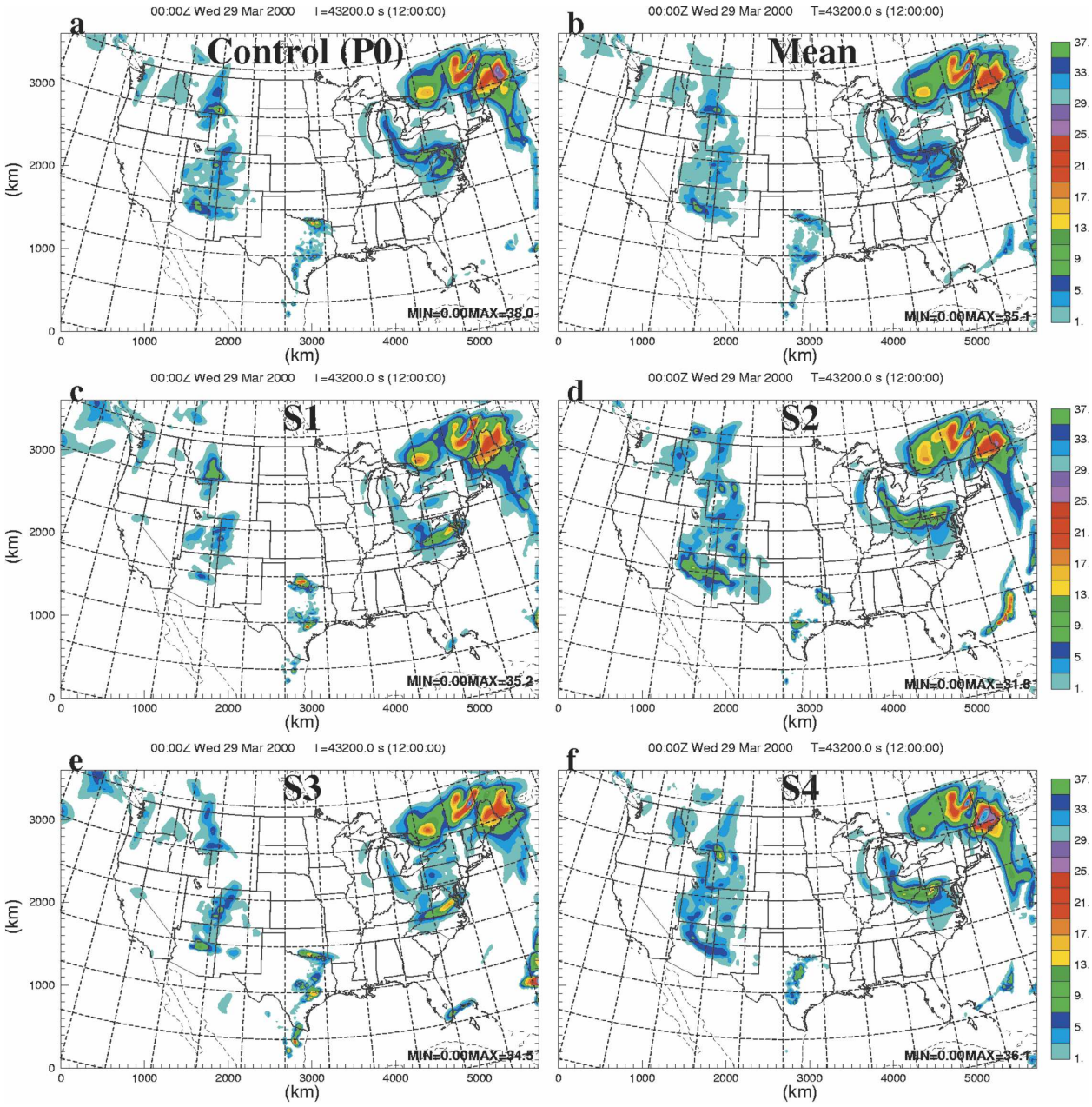

FIG. 14. The 6-h accumulated precipitation (mm), valid at 0000 UTC on 29 Mar 2000, from the 24-km ensemble forecasts: (a) control (P0), (b) ensemble mean, and individual ensemble members (c) S1, (d) S2, (e) S3, and (f) S4.

ing a glaring absence of precipitation in the region where storms were actually located. The probability of precipitation greater than or equal to 2.54 mm, shown in Fig. 16a, also fails to capture the structure present in the observations, and precipitation in south-central Texas is greatly overpredicted.

All 6-km grid spacing forecasts employ both cumulus parameterization (implicit treatment of convection) and grid-scale (explicit) microphysics parameterizations. This grid spacing is problematic in that convective clouds are not entirely subgrid scale nor are they fully resolved (e.g., Molinari 1993). Consequently, no well-defined closure assumption exists, and the implicit and explicit schemes can in fact work against one another. In a midlatitude squall-line simulation, Bélair and Mailhot (2001) found that at 6-km grid spacing, both implicit and explicit microphysics schemes were equally active at the leading edge of the line, thus mak-

TABLE 3. Bias scores of 6-h accumulated precipitation for the 24-km ensemble using multiple thresholds and two forecast periods: a 12-h forecast ending at 0000 UTC on 29 Mar 2000 and an 18-h forecast ending at 0600 UTC. Values in parentheses refer to ETS ≤ 0.

	P0	S1	S2	S3	S4	Mean
			≥0.254 mm			
12 h	0.8801	0.8738	1.3155	0.9621	0.6845	1.8233
18 h	0.8671	0.6601	1.4819	0.7085	0.8384	1.5574
			≥2.54 mm			
12 h	0.9417	1.1083	1.0250	1.4667	0.8000	1.0833
18 h	0.8614	0.8315	1.3221	0.7491	0.7341	0.9813
			≥12.7 mm			
12 h	(0.3125)	(0.8125)	(0.1875)	(1.0625)	(0.1250)	(0.0000)
18 h	1.4103	0.8718	1.4872	0.7949	2.1538	0.3333
			≥25.4 mm			
12 h	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
18 h	(1.2500)	(0.0000)	(0.0000)	(0.0000)	6.2500	(0.0000)

ing the implicit–explicit partitioning of precipitation ambiguous. Nevertheless, they recommended the use of an implicit cumulus scheme. For the Fort Worth case, which uses the Kain–Fritsch implicit scheme, the model produces unrealistic wave-type rainbands aligned northwest–southeast over the area of interest (not shown). These features are smoothed in the accumulated precipitation fields, and although their origin is not entirely clear, they are manifest as internal gravity waves and have been seen in numerous other cases.

To examine the impact of using implicit cumulus parameterization at 6-km grid spacing, the probability of accumulated precipitation from an experimental ensemble that *excludes* the Kain–Fritsch scheme is presented in Fig. 16b. It clearly shows that the explicit scheme alone is unable to capture the intense local precipitation present in the observations (cf. Fig. 12), though individual members of this ensemble produce isolated storms of anomalously high intensity (often termed "grid point storms"), with 6-h accumulated precipitation maxima as large as 284 mm (not shown). This result supports Bélair and Mailhot's (2001) recommendation of using both implicit and explicit convective parameterizations at 6-km grid spacing.

Statistical verification scores for the 6-km forecasts, computed over the same two 6-h time periods as for the 24-km forecasts, are shown in Tables 7–9. The bias scores for the two lowest thresholds show that all members except s4 overforecast 6-h accumulated precipitation at least during one time period, and that in general, the biases are larger than for the 24-km ensemble (cf. Table 3). ETS scores (Table 8) show that all members are more skillful than a random forecast for the two lowest precipitation thresholds, with the exception of cn6 and s2 at 06 h, for which the ETS is too small to be judged skillful. ETS for the 12.7- and 25.4-mm thresholds, however, are mixed: some members (cn6, s1, and s2) are skillful during at least one of the 6-h periods

TABLE 4. Equitable threat scores of 6-h accumulated precipitation for the 24-km ensemble using multiple thresholds and two forecast periods: a 12-h forecast ending at 0000 UTC on 29 Mar 2000 and an 18-h forecast ending at 0600 UTC.

	P0	S1	S2	S3	S4	Mean
			≥0.254 mm			
12 h	0.0517	0.1680	0.0108	0.0606	0.1233	0.1342
18 h	0.3277	0.3203	0.1472	0.2348	0.1232	0.2380
			≥2.54 mm			
12 h	0.0309	0.1580	0.0224	0.0575	0.2389	0.0910
18 h	0.2651	0.1671	0.1285	0.1875	0.1511	0.3227
			≥12.7 mm			
12 h	−0.0018	−0.0034	−0.0012	−0.0039	0.0000	0.0000
18 h	0.1218	0.0810	0.3762	0.1587	0.1590	0.1510
			≥25.4 mm			
12 h	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
18 h	−0.0011	0.0000	0.0000	0.0000	0.0341	0.0000

TABLE 5. Brier scores (BS) and ranked probability scores (RPS) for the 24-km ensemble for two forecast periods: a 12-h forecast ending at 0000 UTC on 29 Mar 2000 and an 18-h forecast ending at 0600 UTC.

	BS		RPS	
	Age-scaled	Amplitude-scaled	Age-scaled	Amplitude-scaled
12 h	0.0624	0.0594	0.2242	0.2079
18 h	0.0881	0.0867	0.2653	0.2614

while s3 shows no skill at all. Most members, except for cn6 and s4, are more skillful during the second period than the first, in agreement with the 24-km ensemble.

Table 9 shows ensemble probability scores including those for the experimental ensemble in which the Kain–Fritsch cumulus scheme was not used (No-KF). The BS and RPS generally are in agreement with those of the 24-km ensemble. It is interesting that despite obvious differences introduced by removing the Kain–Fritsch scheme, the probability scores from the experiments with and without it are similar.

Summarizing, the 6-km grid spacing forecasts tend to greatly overpredict the amplitude and aerial coverage of precipitation—much more so than in the 24-km experiments. In some 6-km forecasts, the results bear no resemblance to observations, and "grid point storms," having extremely high amplitude, are present at multiple locations. This reinforces the notion that subgrid closure for deep convection does not exist at this grid spacing using parameterization schemes currently available. It also raises questions about the physical meaningfulness of such forecasts and whether they should be used to bridge the resolution mismatch between coarse (24 km) and fine (3 km) grids. We explore this issue in Part II.

c. 3-km ensemble forecasts

Using idealized simulations, Weisman et al. (1997) suggested that horizontal grid spacing of 4 km in a nonhydrostatic cloud-resolving model may be sufficient to capture the meso-convective-scale features of organized deep convective systems. In the context of more realistic NWP in regions of significant terrain, Mass et al. (2002) showed considerable improvement in conventional skill scores when horizontal grid spacing was refined from 24 to 12 km, with more modest improvement from 12 to 4 km (note that because his study involved nonconvective, orographically forced precipitation, its results must be applied with caution to the present study). In the latter case, finer spacing led to greater fidelity in key features, though larger speed and position errors partially offset the benefit. Several years of experience at the Center for Analysis and Prediction of Storms (CAPS) in running daily forecasts at grid spacings ranging from 9 to 3 km, along with recent experience with the Weather Research and Forecast (WRF) model (Weiss et al. 2004), suggests that 3–4 km represents a lower bound on horizontal grid spacing for explicitly resolved deep convection, even though energetics are improperly represented (Bryan et al. 2003).

The 3-km grid spacing ensemble presented here is initiated at 2300 UTC and extends for 7 h (see Fig. 5). The most important new information assimilated is radar data. If data assimilation is applied to only the control forecast, as is the case for the 24- and 6-km grid spacing experiments, certain of the ensemble members do not contain perturbations of sufficient amplitude to trigger deep convection (see Part II).[7] Applying such assimilation to both the control (cn3) *and* perturbed members (s1–s4) inevitably reduces ensemble spread and violates the foundational principles of SLAF. We nevertheless follow that procedure here using level III reflectivity and radial velocity data from the nine radars shown in Fig. 17 and examine alternative strategies in Part II. This special treatment at 3 km will be considered in subsequent comparisons that seek to identify the potential benefits of fine grid spacing.

Figure 18 shows 1-h near-surface reflectivity forecasts from individual members as well as the ensemble mean at 3-km grid spacing. The ensemble compares reasonably well with radar observations in Fig. 1 in the orientation, nature, and location of deep convection, particularly in light of how the operational Eta Model failed to explicitly predict any precipitation over northern and central Texas 12 h prior to the event. However, notable discrepancies emerge upon detailed examination. The area of convection in north Texas is predicted relatively well by members s1 and s3 in that they place storms in roughly the right region with intensities (mea-

TABLE 6. Brier scores for 6-h accumulated precipitation from 24-km ensembles with different reference perturbation amplitudes. The scaling factors in first row are applied to the original magnitude of perturbation 1 (P1 − P0; see Fig. 4) to form new references.

	0.5×	1.0×	1.5×	2.0×
12 h	0.0671	0.0594	0.0590	0.0621
18 h	0.0894	0.0867	0.0864	0.0882

[7] Additional experiments were conducted in which data assimilation was applied to all members at 24- and 6-km grid spacing, consistent with the approach used at 3-km spacing. The results showed reduced spread and no quantitative improvement in skill.

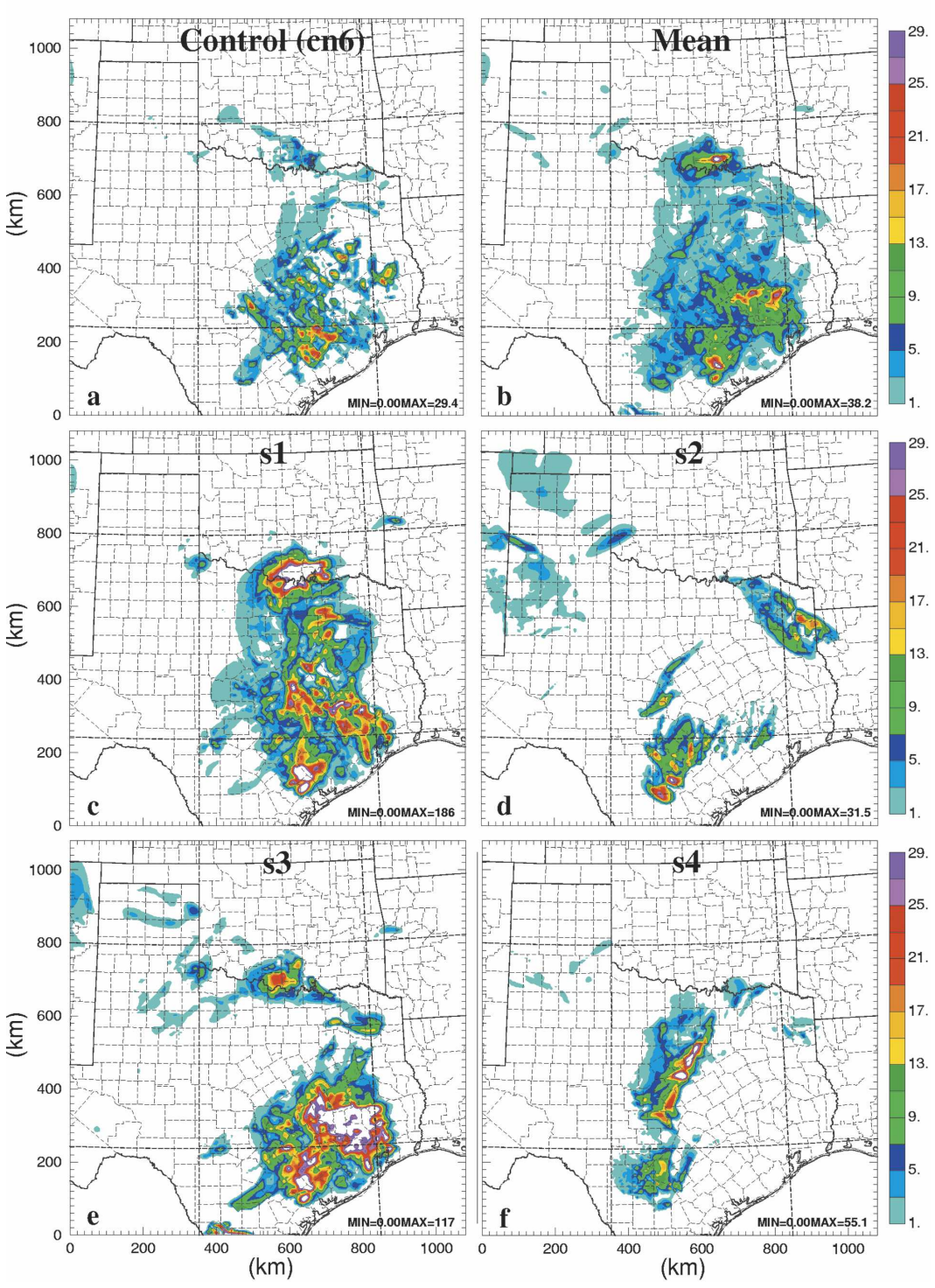

FIG. 15. As in Fig. 14, except for the 6-km grid spacing ensemble.

sured by reflectivity) within 10% of those observed. Consistent with observations, all members predict the existence of multiple cells south of Fort Worth. However, members s1 and s3 produce spurious scattered storms in the southeastern portion of the domain, largely as a result of forcing by the background field. The assimilation of radar data plays an important role in improving the forecast of this storm system, as was

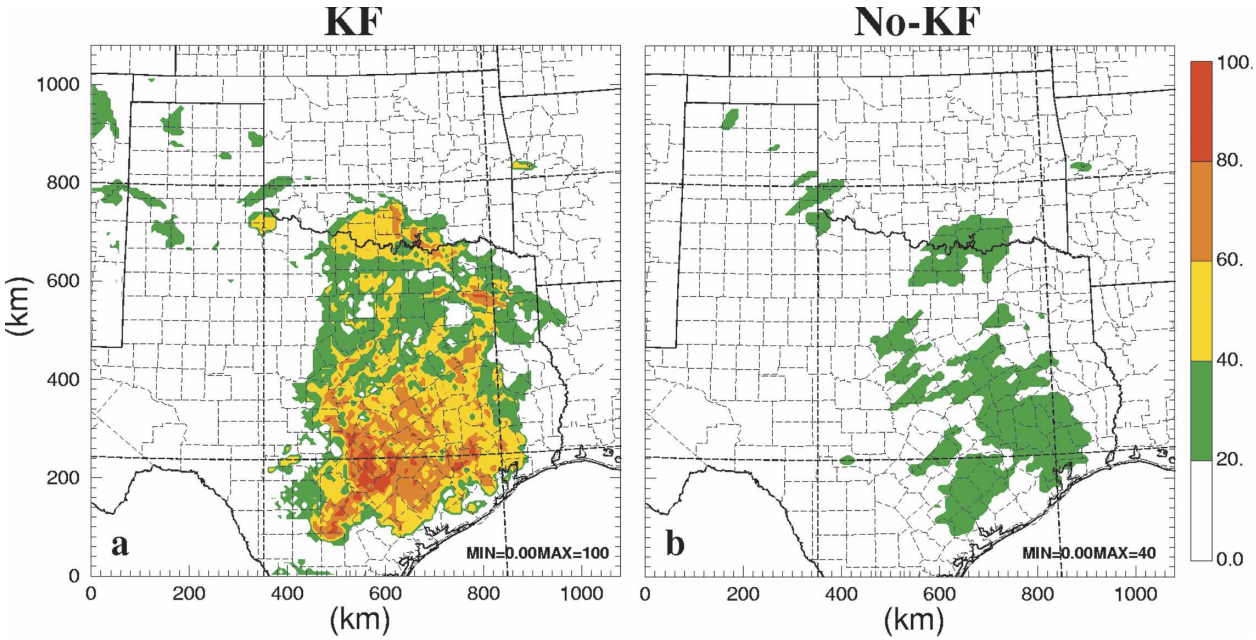

FIG. 16. Probability of 6-h accumulated precipitation greater than or equal to 2.54 mm from the 6-km ensemble forecast, valid at 0000 UTC on 29 Mar 2000 (a) with and (b) without the Kain–Fristch cumulus parameterization scheme.

demonstrated by Xue et al. (2003). We examine in detail the impact of radar data on the 3-km ensembles in Part II.

Owing to the spatially intermittent nature of deep convection, the ensemble mean reflectivity forecast (Fig. 18b) covers a somewhat broader area than any of the individual forecasts due to smearing, and each storm tends to be weaker, thus hindering the usefulness of the mean as an ensemble product.[8] For this reason, probability maps become a natural alternative.

[8] Note that the forecast spread contains features that resemble the storms themselves (see Part II), thus making the spread more challenging to interpret.

Figure 19 shows hourly probability of surface reflectivity greater than or equal to 35 and 45 dBZ for the first three forecast hours. The highest probabilities are in reasonably good agreement with observed reflectivity from the KFWS (Fort Worth) WSR-88D radar (Fig. 1). The spurious echoes generated by some of the ensemble members over the southeastern portion of the domain have relatively low probability in the total ensemble, thus demonstrating the value of probability forecasts. The 2- and 3-h forecasts show better agreement with radar observations near Fort Worth as storms move eastward through the area. The second line of convection emerging in the far northwestern re-

TABLE 7. As in Table 3 but for the 6-km grid spacing experiment.

	cn6	s1	s2	s3	s4	Mean
			≥0.254 mm			
06 h	1.1037	1.7143	1.3283	1.6788	0.7784	2.3661
12 h	0.9604	0.7681	1.2927	1.1242	0.8960	1.7160
			≥2.54 mm			
06 h	1.2475	2.7316	1.1668	2.3825	0.9492	2.4567
12 h	1.0610	0.8275	1.3291	1.3002	0.9059	1.4017
			≥12.7 mm			
06 h	(0.7633)	4.3004	(0.7420)	(5.1095)	0.7880	(0.7562)
12 h	1.1149	1.1792	1.2048	(1.5850)	1.7887	0.6325
			≥25.4 mm			
06 h	(0.0816)	5.6735	(0.2653)	(11.7755)	(0.7959)	(0.3673)
12 h	0.6320	2.2880	1.2640	(1.0800)	(4.0960)	(0.0320)

TABLE 8. As in Table 4 but for the 6-km grid spacing experiment.

	cn6	s1	s2	s3	s4	Mean
			≥0.254 mm			
06 h	0.0988	0.1476	0.0483	0.0592	0.1831	0.1463
12 h	0.2915	0.2377	0.1354	0.1591	0.1517	0.2138
			≥2.54 mm			
06 h	0.0106	0.1062	0.0116	0.0258	0.1736	0.0767
12 h	0.2118	0.1733	0.1395	0.0407	0.1217	0.2029
			≥12.7 mm			
06 h	−0.0038	0.0345	−0.0017	−0.0021	0.0569	−0.0038
12 h	0.1289	0.1984	0.2688	−0.0095	0.0269	0.1638
			≥25.4 mm			
06 h	−0.0001	0.0238	−0.0003	−0.0014	−0.0007	−0.0004
12 h	0.0134	0.0846	0.0158	−0.0020	−0.0031	−0.0001

gion of the KFWS scan area at 0000 UTC (Fig. 1) is not captured by the model owing to the absence of organized convergence in the background field near the model inflow boundary (and perhaps partly because of the proximity of the lateral boundary itself).

Figure 20 shows stage IV 1-h accumulated precipitation valid at 0000 UTC on 29 March 2000, remapped to a grid of 3-km spacing from its original 4-km grid as described previously. For comparison, the 1-h accumulated precipitation from each member of the 3-km ensemble forecasts, as well as from the mean, is shown in Fig. 21. Not surprisingly, the 3-km ensemble contains significant detail compared to its 24- and 6-km counterparts, and generally agrees more closely with Fig. 20. Although all members have a somewhat similar precipitation pattern, simulations s2 and s4 exhibit lower amplitude while s1 and s3 generate spurious accumulations over the southeastern portion of the domain. The elongated structure of accumulated precipitation in the forecasts, in comparison to the more circular features present in the observations, is due to a larger number of storms present within the model. Unlike reflectivity (or precipitation rate), which is an instantaneous quantity, hourly accumulated precipitation is a time-integrated variable, and thus the ensemble mean preserves peak intensity quite well.

Figure 22 shows probabilities of 1-h accumulated precipitation greater than or equal to 2.54 and 12.7 mm for the 1-h forecast at 3-km grid spacing. Visually, the model captures the general structure of the observed precipitation (cf. Fig. 20) in much more detail than its counterparts at coarser grid spacings (cf. Figs. 13b and 16a), while de-emphasizing the spurious convection in the southeastern part of the domain.

Table 10, which shows bias scores over the first 3 h of the forecast, indicates that the ensemble mean overpredicts 1-h accumulated precipitation for every threshold, with the largest contribution from members s1 and s3.

TABLE 9. As in Table 5 but for the 6-km grid spacing experiment. Also shown is an ensemble in which the Kain–Fritsch cumulus parameterization scheme was not used (No-KF).

	BS		RPS	
	KF	No-KF	KF	No-KF
06 h	0.0585	0.0559	0.2190	0.2063
12 h	0.0890	0.0882	0.2970	0.3021

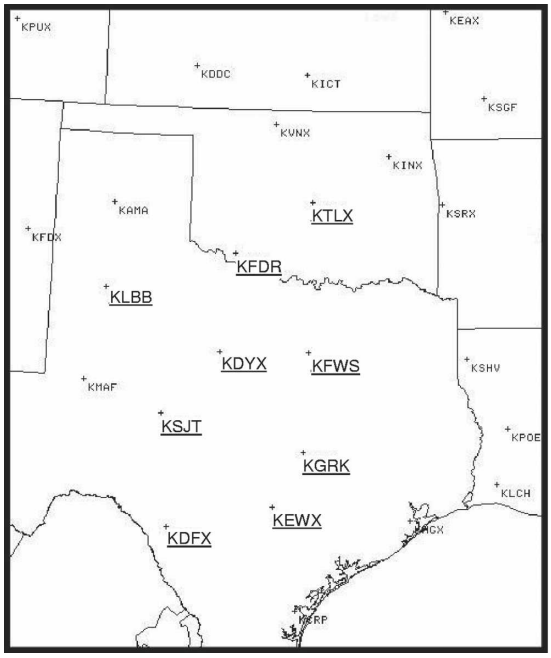

FIG. 17. Location and station identifiers of selected WSR-88D (NEXRAD) radars. Underlined identifiers indicate sites from which level III radial velocity and reflectivity data were used in the 3-km grid spacing ensembles.

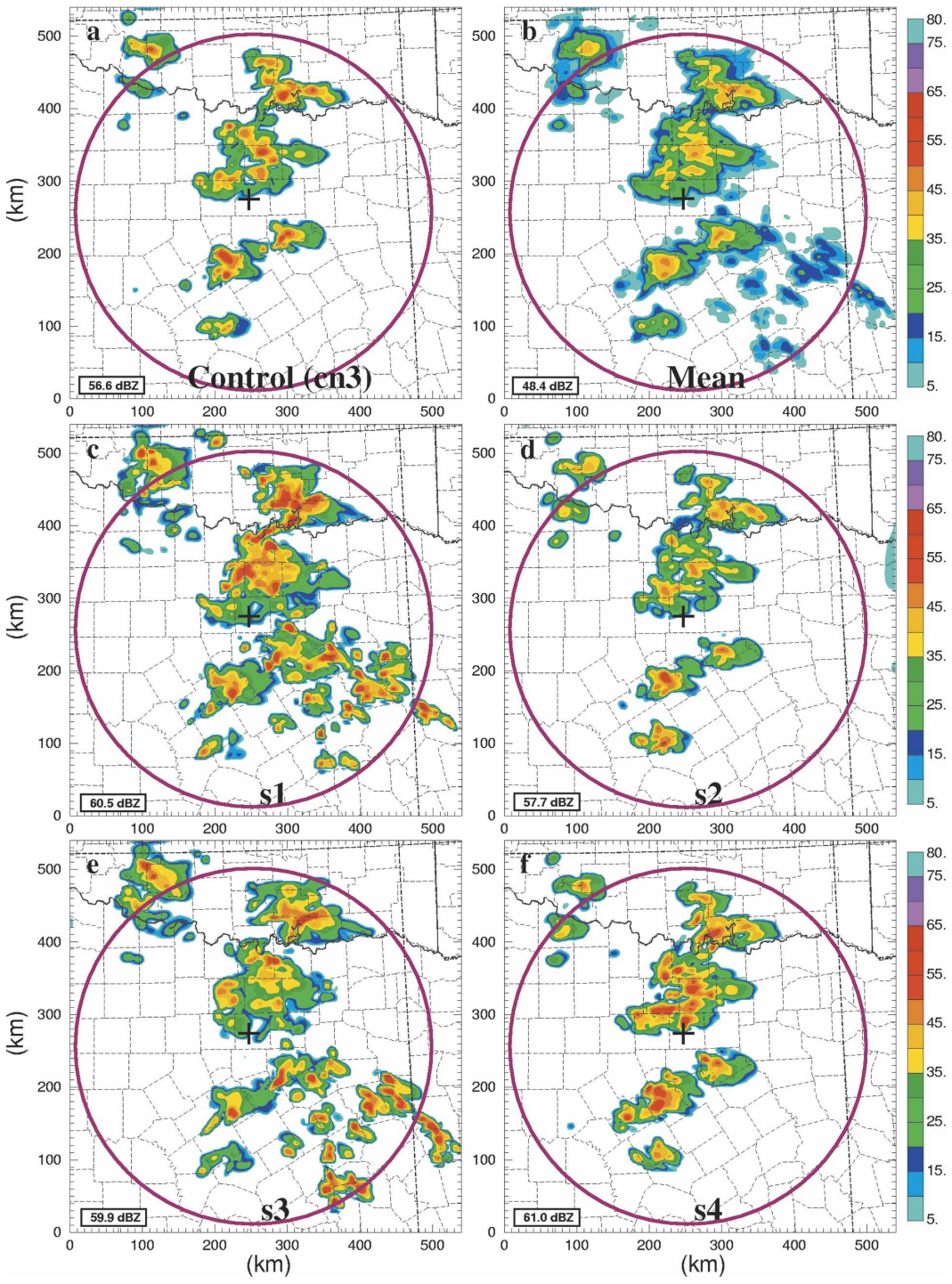

FIG. 18. The 1-h forecast surface reflectivity valid at 0000 UTC on 29 Mar 2000 from individual 3-km ensemble members: (a) control (cn3), (b) ensemble mean, (c) s1, (d) s2, (e) s3, and (f) s4. The large circles mark WSR-88D (KFWS) radar scan range in Fig. 1.

The mean bias scores are comparable to those at 24-km spacing for the lowest threshold but are much larger than those at both 24- and 6-km spacing for the higher thresholds. The ETS in Table 11 shows that for nearly every verification time and threshold, at least one ensemble member is more skillful than the control run (cn3), while the ensemble mean is not necessarily so. Exceptions exist for thresholds of 2.54 and 12.7 mm at

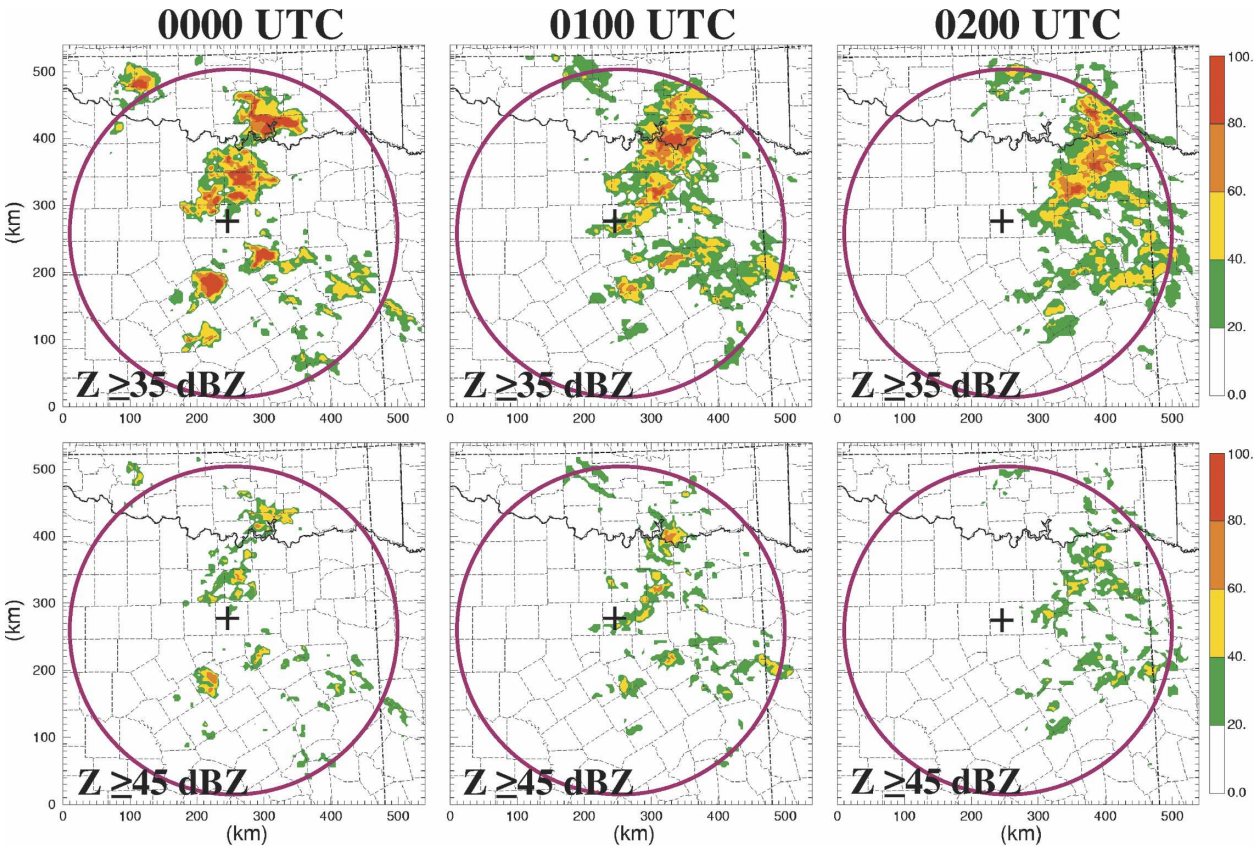

FIG. 19. Probability of surface reflectivity greater than or equal to (top) 35 and (bottom) 45 dB*Z* from the 3-km ensemble valid every hour from 0000 to 0200 UTC on 29 Mar 2000. The large circles mark WSR-88D (KFWS) radar scan range in Fig. 1.

3 h. In contrast to the coarser grid spacing forecasts, the BS and RPS (Table 12) both show improving skill with time over the verification period.

In summary, the 3-km ensemble resolves the overall storm system in considerable detail compared to its 24- and 6-km counterparts (though not with a cell-to-cell match between model and observations), enabling comparison between modeled and radar observed reflectivity. Probability forecasts appear to have practical value in identifying regions of potentially intense local weather, though unlike the coarser-grid ensembles, the ensemble mean and spread (especially for reflectivity field) at 3 km are not terribly useful owing to the highly intermittent nature of deep convection.

6. Summary

We extended the concept of ensemble forecasting down to the scale of individual convective storms by applying a full-physics numerical prediction system, initialized with observations that included WSR-88D Doppler radar data, to an observed tornadic thunder-

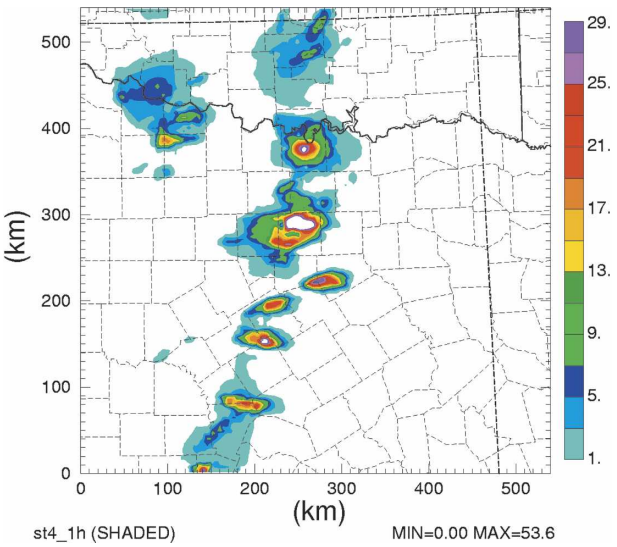

FIG. 20. Stage IV hourly accumulated precipitation ending at 0000 UTC on 29 Mar 2000.

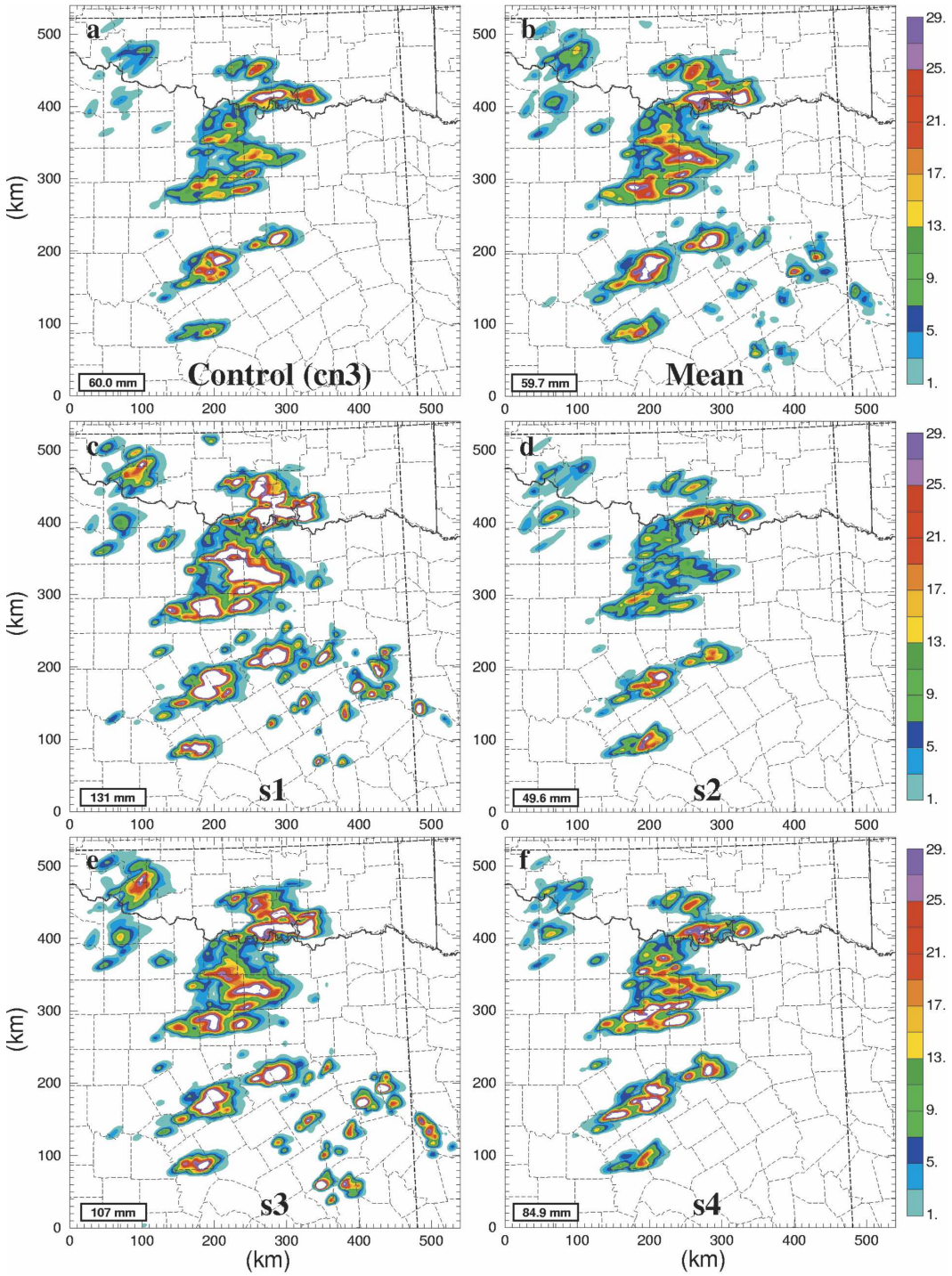

FIG. 21. As in Fig. 14, but for 1-h accumulated precipitation from the 3-km grid spacing ensemble forecast.

storm complex that passed through the Fort Worth, Texas, area on 28–29 March 2000. Using grids of 24, 6, and 3 km within the Advanced Regional Prediction System (ARPS), we constructed five-member ensembles for each grid using a modified version of the scaled lagged average forecasting (SLAF) technique. Initial perturbation sizes and structures, as well as error growth features, suggested the use of an amplitude-based rather than an age-based scaling strategy, the veracity of which was verified in model forecasts prin-

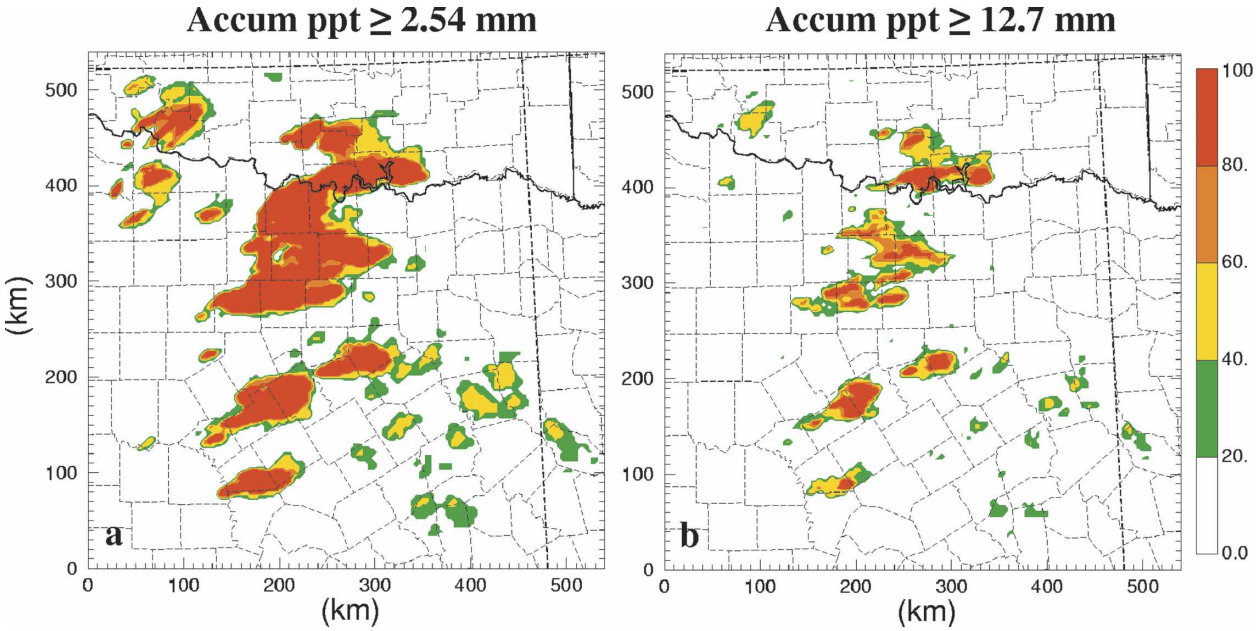

FIG. 22. Probability of 1-h accumulated precipitation greater than or equal to (a) 2.54 and (b) 12.7 mm for the 3-km ensemble forecast valid at 0000 UTC on 29 Mar 2000.

cipally by improved spread and somewhat better forecast skill. Although SLAF is relatively simple and for larger-scale models, as well as the experiments shown here, produces errors that grow too slowly, it nevertheless provides a foundation upon which to build at the storm scale.

Accumulated precipitation was more skillfully forecast by the ensemble mean than the control case at 24-km grid spacing, though in no case could the model at this spacing explicitly resolve the most energetic structures of the storm system. The 6-km ensemble captured some aspects of convective system morphology, but in general vastly overpredicted precipitation owing to "grid point storms" and unrealistic wave-type features associated with the use of an implicit cumulus parameterization scheme. With explicit cloud microphysics and the assimilation of WSR-88D level III radar data for all ensemble members, the 3-km grid spacing

TABLE 10. Bias scores of 1-h accumulated precipitation from the 3-km ensemble for given thresholds. The three forecast periods end at 0000, 0100, and 0200 UTC on 29 Mar. The values in parentheses refer to ETS ≤ 0.

	cn3	s1	s2	s3	s4	Mean
			≥0.254 mm			
1 h	0.6823	1.1748	0.7125	1.1617	0.7363	1.1465
2 h	0.7291	1.4528	0.7874	1.5076	0.8203	1.6740
3 h	0.7299	1.2206	0.7198	1.2902	0.7966	1.6364
			≥2.54 mm			
1 h	0.9163	1.6989	0.9236	1.6340	1.0702	1.4216
2 h	1.0415	2.5351	0.8990	2.2998	1.3261	2.4917
3 h	1.7816	3.5925	1.6071	3.4787	1.9833	3.9714
			≥12.7 mm			
1 h	1.4595	4.4464	0.9869	3.8249	2.3348	2.6608
2 h	1.8481	5.9341	1.5215	(4.3496)	3.3725	2.8395
3 h	2.6426	5.4955	2.6517	4.3844	(3.6126)	3.1471
			≥25.4 mm			
1 h	(1.2326)	11.6395	(0.4651)	(6.8372)	3.7558	3.6047
2 h	(12.4286)	(38.2857)	11.9048	(20.8095)	29.5714	(5.5714)
3 h	3.6106	5.7080	(3.6460)	4.4513	(5.5221)	0.7788

TABLE 11. ETS of 1-h accumulated precipitation for the 3-km ensemble for given thresholds. The three forecast periods end at 0000, 0100, and 0200 UTC on 29 Mar.

	cn3	s1	s2	s3	s4	Mean
			≥0.254 mm			
1 h	0.2618	0.2365	0.2692	0.2323	0.2633	0.2216
2 h	0.1927	0.1689	0.1984	0.0925	0.2080	0.1538
3 h	0.1494	0.2372	0.1169	0.1598	0.1299	0.2128
			≥2.54 mm			
1 h	0.1766	0.1791	0.1699	0.1668	0.1880	0.1783
2 h	0.1035	0.1334	0.0773	0.0390	0.1363	0.1257
3 h	0.1566	0.1070	0.0591	0.0509	0.0734	0.1125
			≥12.7 mm			
1 h	0.0500	0.0530	0.0378	0.0407	0.0469	0.0492
2 h	0.0338	0.0011	0.0499	−0.0018	0.0793	0.0479
3 h	0.1095	0.0804	0.0067	0.0589	−0.0081	0.0400
			≥25.4 mm			
1 h	−0.0015	0.0012	−0.0008	−0.0023	0.0230	0.0056
2 h	−0.0006	−0.0006	0.0417	−0.0006	0.0168	−0.0005
3 h	0.0693	0.0104	−0.0027	0.0120	−0.0030	0.0085

ensemble compared favorably to observed reflectivity and hourly accumulated stage IV precipitation estimates. All ensemble members at this spacing predicted overall storm system structure and movement reasonably well, with notable spread and in some cases large areas of spurious convection. Probabilities were found to be especially useful when applied at fine model grid spacings because they tend to focus attention on high intensity and rare events. Because forecast spread of reflectivity and other features exhibits a structure similar to the features themselves, its value may be of limited practical value unless suitable modifications are developed for application to intermittent flows.

Equitable threat scores from the 24-km ensemble showed that the mean forecast was, for the most part, more skillful than the control. Such was not the case for the 3-km ensemble, in which at least one member performed more skillfully than the control. Brier and ranked probability scores for the 24- and 6-km ensembles showed comparable performance, with skill decreasing throughout the forecast period. In contrast, the 3-km ensemble exhibited improving skill with time during the 3-h verification period.

TABLE 12. Brier scores and RPS for the 3-km ensembles. The three forecast periods end at 0000, 0100, and 0200 UTC on 29 Mar.

	BS	RPS
1 h	0.0828	0.3113
2 h	0.0790	0.2802
3 h	0.0755	0.2576

Recalling our principal goal of examining the potential value added to quantitative precipitation forecasting through the use of fine grid spacing and an ensemble methodology, we have shown that 3-km ensembles have greater value than a single deterministic forecast at the same grid spacing and to both ensembles and single deterministic forecasts at coarser spacing. These results are consistent with the conjecture by Mass et al. (2002) that ensemble forecasting may provide improved results as benefits from continued decreases in horizontal grid spacing for single forecasts reach a point of diminishing return. Yet, the single case study described here obviously does not provide sufficient information with which to draw general conclusions, and thus considerable work remains to more fully understand ensemble forecasting of deep convection and other intense local weather. Specifically, efforts must be directed toward examining other strategies for linking fine and coarser grids, toward evaluating other techniques for generating initial perturbations (e.g., bred vectors, singular vectors), and toward assessing the impact of lateral boundary conditions. Quantitative measures of both skill and value are needed for highly intermittent phenomena, and experiments involving ensembles based upon multiple models, variations in physics, and combinations thereof must be undertaken.

Acknowledgments. This research was funded by the National Science Foundation under Grants ATM99-81130 and ATM01-30396 to the second author. Comments provided by two anonymous reviewers, and an especially thorough review by Dr. Steve Mullen of the

University of Arizona, substantially improved the manuscript. The numerical simulations were performed at the Oklahoma Supercomputing Center for Education and Research, which is funded by the University of Oklahoma and was created in part by a grant from the Williams Energy Marketing and Trading Corporation.

REFERENCES

Adlerman, E. L., and K. K. Droegemeier, 2002: The sensitivity of numerically simulated cyclic mesocyclogenesis to variations in model physical and computational parameters. *Mon. Wea. Rev.,* **130,** 2671–2691.

Alberoni, P. P., and Coauthors, 2003: Quality and assimilation of radar data for NWP—A review. European Commission Rep. EUR 20600, COST 717 Review Rep., 38 pp.

Bélair, S., and J. Mailhot, 2001: Impact of horizontal resolution on the numerical simulation of a midlatitude squall line: Implicit versus explicit condensation. *Mon. Wea. Rev.,* **129,** 2362–2376.

Brewster, K., 1996: Implementation of a Bratseth analysis scheme including Doppler radar. Preprints, *15th Conf. on Weather Analysis and Forecasting,* Norfolk, VA, Amer. Meteor. Soc., 92–95.

Brooks, H. E., C. A. Doswell III, and R. A. Maddox, 1992: On the use of mesoscale and cloud-scale models in operational forecasting. *Wea. Forecasting,* **7,** 120–132.

——, M. S. Tracton, D. J. Stensrud, G. Dimego, and Z. Toth, 1995: Short-range ensemble forecasting: Report from a workshop (25–27 July 1994). *Bull. Amer. Meteor. Soc.,* **76,** 1617–1624.

Bryan, G. H., J. C. Wyngaard, and J. M. Fritsch, 2003: Resolution requirements for the simulation of deep moist convection. *Mon. Wea. Rev.,* **131,** 2394–2416.

Carpenter, R. L., Jr., K. K. Droegemeier, G. M. Bassett, W. L. Qualley, and R. Strasser, 1997: Project Hub-CAPS: Storm-scale NWP for commercial aviation. Preprints, *Seventh Conf. on Aviation, Range, and Aerospace Meteorology,* Long Beach, CA, Amer. Meteor. Soc., 474–479.

——, ——, ——, S. S. Weygandt, D. E. Jahn, S. Stevenson, W. L. Qualley, and R. Strasser, 1999: Storm-scale numerical weather prediction for commercial and military aviation. Part I: Results from operational tests in 1998. Preprints, *Eighth Conf. on Aviation, Range, and Aerospace Meteorology,* Dallas, TX, Amer. Meteor. Soc., 209–211.

Chou, M.-D., 1990: Parameterization for the absorption of solar radiation by O_2 and CO_2 with application to climate study. *J. Climate,* **3,** 209–217.

——, and M. J. Suarez, 1994: An efficient thermal infrared radiation parameterization for use in general circulation models. NASA Tech. Memo. 104606, 85 pp.

Crook, N. A., 1996: Sensitivity of moist convection forced by boundary layer processes to low-level thermodynamic fields. *Mon. Wea. Rev.,* **124,** 1767–1785.

——, and J. Sun, 2002: Assimilating radar, surface, and profiler data for the Sydney 2000 Forecast Demonstration Project. *J. Atmos. Oceanic Technol.,* **19,** 888–898.

——, and ——, 2004: Analysis and forecasting of the low-level wind during the Sydney 2000 Forecast Demonstration Project. *Wea. Forecasting,* **19,** 151–167.

Dawson, D. T., II, and M. Xue, 2006: Numerical forecasts of 15–16 June 2002 Southern Plains MCS: Impact of mesoscale data and cloud analysis. *Mon. Wea. Rev.,* in press.

Droegemeier, K. K., 1997: The numerical prediction of thunderstorms: Challenges, potential benefits, and results from real time operational tests. *WMO Bull.,* **46,** 324–336.

——, and Coauthors, 1996: The 1996 CAPS spring operational forecasting period: Realtime storm-scale NWP. Part I: Goals and methodology. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., 297–300.

Du, J., and M. S. Tracton, 2001: Implementation of a real-time short-range ensemble forecasting system at NCEP: An update. Preprints, *Ninth Conf. on Mesoscale Processes,* Fort Lauderdale, FL, Amer. Meteor. Soc., 355–356.

Ebisuzaki, W., and E. Kalnay, 1991: Ensemble experiments with a new lagged analysis forecasting scheme. Research Activities in Atmospheric and Oceanic Modeling Rep. 15, WMO, 423 pp.

Elmore, K. L., D. J. Stensrud, and K. C. Crawford, 2002a: Ensemble cloud model applications to forecasting thunderstorms. *J. Appl. Meteor.,* **41,** 363–383.

——, ——, and ——, 2002b: Explicit cloud-scale models for operational forecasts: A note of caution. *Wea. Forecasting,* **17,** 873–884.

——, S. J. Weiss, and P. C. Banacos, 2003: Operational ensemble cloud model forecasts: Some preliminary results. *Wea. Forecasting,* **18,** 953–964.

Hamill, T. M., and C. Snyder, 2000: A hybrid ensemble Kalman filter–3D variational analysis scheme. *Mon. Wea. Rev.,* **128,** 2905–2919.

——, S. L. Mullen, C. Snyder, and Z. Toth, 2000: Ensemble forecasting in the short to medium range: Report from a workshop. *Bull. Amer. Meteor. Soc.,* **81,** 2653–2664.

Hoffman, R. N., and E. Kalnay, 1983: Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus,* **35A,** 100–118.

Hou, D., E. Kalnay, and K. K. Droegemeier, 2001: Objective verification of the SAMEX'98 ensemble forecasts. *Mon. Wea. Rev.,* **129,** 73–91.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.,* **126,** 796–811.

Hu, M., and M. Xue, 2002: Sensitivity of model thunderstorms to modifications to the environmental conditions by a nearby thunderstorm in the prediction of 2000 Fort Worth tornado case. Preprints, *15th Conf. on Numerical Weather Prediction/19th Conf. on Weather Analysis and Forecasting,* San Antonio, TX, Amer. Meteor. Soc., 2.1.

Kain, J. S., and J. M. Fritsch, 1993: Convective parameterization for mesoscale models: The Kain–Fritsch scheme. *The Representation of Cumulus Convection in Numerical Models, Meteor. Monogr.,* No. 46, Amer. Meteor. Soc., 165–170.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 341 pp.

Levit, N. L., K. K. Droegemeier, and F. Kong, 2004: High-resolution storm-scale ensemble forecasts of the 28 March 2000 Fort Worth tornadic storms. Preprints, *20th Conf. on Weather Analysis and Forecasting/16th Conf. on Numerical Weather Prediction,* Seattle, WA, Amer. Meteor. Soc., CD-ROM, 23.6.

Lin, Y. L., R. D. Farley, and H. D. Orville, 1983: Bulk parameterization of the snow field in a cloud model. *J. Climate Appl. Meteor.,* **22,** 1065–1092.

Martin, W. J., and M. Xue, 2004: Initial condition sensitivity analysis of a mesoscale forecast using very-large ensembles. Preprints, *20th Conf. on Weather Analysis and Forecasting/ 16th Conf. on Numerical Weather Prediction,* Seattle, WA, Amer. Meteor. Soc., CD-ROM, J8.2.

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.,* **83,** 407–430.

Molinari, J., 1993: An overview of cumulus parameterization in mesoscale models. *The Representation of Cumulus Convection in Numerical Models, Meteor. Monogr.,* No. 46, Amer. Meteor. Soc., 155–158.

Mullen, S. L., and D. P. Baumhefner, 1989: The impact of initial condition uncertainty on numerical simulations of large-scale explosive cyclogenesis. *Mon. Wea. Rev.,* **117,** 2800–2821.

NCDC, 2000: *Storm Data*. Vol. 42, No. 3, 172 pp.

Noilhan, J., and S. Planton, 1989: A simple parameterization of land surface processes for meteorological models. *Mon. Wea. Rev.,* **117,** 536–549.

Nutter, P., D. Stensrud, and M. Xue, 2004: Effects of coarsely resolved and temporally interpolated lateral boundary conditions on the dispersion of limited-area ensemble forecasts. *Mon. Wea. Rev.,* **132,** 2358–2377.

Pleim, J. E., and A. Xiu, 1995: Development and testing of a surface flux and planetary boundary layer model for application in mesoscale models. *J. Appl. Meteor.,* **34,** 16–32.

Sindic-Rancic, G., Z. Toth, and E. Kalnay, 1997: Storm-scale ensemble experiments with the ARPS model: Preliminary results. Preprints, *12th Conf. on Numerical Weather Prediction,* Amer. Meteor. Soc., 279–280.

Smedsmo, J. L., E. Foufoula-Georgiou, V. Vuruputur, F. Kong, and K. K. Droegemeier, 2005: On the vertical structure of modeled and observed deep convective storms: Insights for precipitation retrieval and microphysical parameterization. *J. Appl. Meteor.,* **44,** 1866–1884.

Stensrud, D. J., J.-W. Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensembles. *Mon. Wea. Rev.,* **128,** 2077–2107.

Sun, J., and N. A. Crook, 1998: Dynamical and microphysical retrieval from Doppler radar observations using a cloud model and its adjoint. Part II: Retrieval experiments of an observed Florida convective storm. *J. Atmos. Sci.,* **55,** 835–852.

Tao, W. K., and J. Simpson, 1993: Goddard Cumulus Ensemble Model. Part I: Model description. *Terr. Atmos. Oeceanic Sci.,* **4,** 35–72.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NWC: The generation of perturbations. *Bull. Amer. Meteor. Soc.,* **74,** 2317–2330.

——, and ——, 1997: Ensemble forecasting at NCEP: The breeding method. *Mon. Wea. Rev.,* **125,** 3297–3318.

Wang, X., and C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.,* **60,** 1140–1158.

Warner, T. T., R. A. Peterson, and R. E. Treadon, 1997: A tutorial on lateral boundary conditions as a basic and potentially serious limitation to regional numerical weather prediction. *Bull. Amer. Meteor. Soc.,* **78,** 2599–2617.

Weisman, M. L., W. C. Shamarock, and J. B. Klemp, 1997: The resolution dependence of explicitly modeled convective systems. *Mon. Wea. Rev.,* **125,** 527–548.

Weiss, S. J., J. S. Kain, J. J. Levit, M. E. Baldwin, and D. R. Bright, 2004: Examination of several different versions of the WRF model for the prediction of severe convective weather: The SPC/NSSL Spring Program 2004. Preprints, *22d Conf. on Severe Local Storms,* Hyannis, MA, Amer. Meteor. Soc., CD-ROM, 17.1.

Weygandt, S. S., A. Shapiro, and K. K. Droegemeier, 1998: The use of wind and thermodynamic retrievals to create initial forecast fields from single-Doppler observations of a supercell thunderstorm. Preprints, *11th Conf. on Weather Analysis and Forecasting,* Phoenix, AZ, Amer. Meteor. Soc., 286–288.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 467 pp.

Xue, M., and Coauthors, 1996: The 1996 CAPS spring operational forecasting period: Realtime storm-scale NWP. Part II: Operational summary and examples. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., 297–300.

——, K. K. Droegemeier, and V. Wong, 2000: The Advanced Regional Prediction System (ARPS)—A multiscale nonhydrostatic atmospheric simulation and prediction model. Part I: Model dynamics and verification. *Meteor. Atmos. Phys.,* **75,** 161–193.

——, and Coauthors, 2001: The Advanced Regional Prediction System (ARPS)—A multiscale nonhydrostatic atmospheric simulation and prediction tool. Part II: Model physics and applications. *Meteor. Atmos. Phys.,* **76,** 134–165.

——, D. Wang, J. Gao, K. Brewster, and K. K. Droegemeier, 2003: The Advanced Regional Prediction System (ARPS), storm-scale numerical weather prediction and data assimilation. *Meteor. Atmos. Phys.,* **76,** 143–165.