Multiresolution Ensemble Forecasts of an Observed Tornadic Thunderstorm System.
Part II: Storm-Scale 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*
NOAA/Warning Decision Training Branch, Norman, Oklahoma

(Manuscript received 4 January 2006, in final form 9 June 2006)

ABSTRACT

In Part I, the authors used a full physics, nonhydrostatic numerical model with horizontal grid spacing of 24 km and nested grids of 6- and 3-km spacing to generate the ensemble forecasts of an observed tornadic thunderstorm complex. The principal goal was to quantify the value added by fine grid spacing, as well as the assimilation of Doppler radar data, in both probabilistic and deterministic frameworks. The present paper focuses exclusively on 3-km horizontal grid spacing ensembles and the associated impacts on the forecast quality of temporal forecast sequencing, the construction of initial perturbations, and data assimilation. As in Part I, the authors employ a modified form of the scaled lagged average forecasting technique and use Stage IV accumulated precipitation estimates for verification. The ensemble mean and spread of accumulated precipitation are found to be similar in structure, mimicking their behavior in global models. Both the assimilation of Doppler radar data and the use of shorter (1–2 versus 3–5 h) forecast lead times improve ensemble precipitation forecasts. However, even at longer lead times and in certain situations without assimilated radar data, the ensembles are able to capture storm-scale features when the associated control forecast in a deterministic framework fails to do so. This indicates the potential value added by ensembles although this single case is not sufficient for drawing general conclusions. The creation of initial perturbations using forecasts of the same grid spacing shows no significant improvement over simply extracting perturbations from forecasts made at coarser spacing and interpolating them to finer grids. However, forecast quality is somewhat dependent upon perturbation amplitude, with smaller scaling values leading to significant underdispersion. Traditional forecast skill scores show somewhat contradictory results for accumulated precipitation, with the equitable threat score most consistent with qualitative performance.

1. Introduction

In Kong et al. (2006, hereafter Part I), the authors 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 including Weather Surveillance Radar-1988 Doppler (WSR-88D) radar data, to an observed tornadic thunderstorm complex that passed through the Fort Worth, Texas, area on 28–29 March 2000. Using domains of 24-, 6-, and 3-km horizontal grid spacing within the Advanced Regional Prediction System (ARPS; Xue et al. 2000, 2001, 2003), five-member ensembles were constructed for each grid using a modified version of the scaled lagged average forecasting (SLAF) technique (Ebisuzaki and Kalnay 1991). In general, the ensembles were superior to single forecasts.
for all grid spacings, though at 6-km spacing, the results were problematic owing in part to the absence of a well-defined closure for deep convection. Using explicit cloud microphysics and assimilating WSR-88D Level III radar data for all members, the 3-km grid spacing ensembles captured explicitly the details of storm evolution and compared favorably with observations. Associated probabilities were found to be especially useful at this grid spacing because they identify intense local events.

The present study extends the results of Part I by focusing exclusively on 3-km grid spacing ensembles. Particular emphasis is given to temporal forecast sequencing and the manner in which the 3-km grid is linked to its coarser spacing parents. The absence of a unique strategy for linking the grids—regardless of the method used to define the initial perturbations—as well as the associated impact on error growth are notably important for ensemble forecasting using nesting and are examined here. Further, in operational prediction, fine-grid forecasts might be spawned only when conditions warrant, thus precluding the regular availability of fine-grid background fields from previous forecasts from which to construct perturbations. An alternative is to generate perturbations directly from a coarse-grid forecast (assuming it is available routinely) and interpolate them to finer grids. However, a number of problems exist with this approach, most notably the lack of consistency among grids owing to the possible use of different physical parameterizations. Several perturbation generation and scaling strategies are examined here, as well as the impact on forecast quality of radar data assimilation. As in Part I, both qualitative and quantitative analyses are employed using radar-based estimates of accumulated precipitation and other observations.

Section 2 describes the construction of initial perturbations while section 3 examines analysis error, perturbation structure, and growth. Experiments to assess various sensitivities are presented in section 4, and in section 5, a variety of statistics are used to evaluate forecast skill. A summary and outlook are offered in section 6.

2. Construction of perturbations

a. Model and general approach

All experiments utilize the ARPS, a three-dimensional, nonhydrostatic compressible numerical weather prediction system with comprehensive physics and a self-contained data ingest, quality control, retrieval and assimilation system (Xue et al. 2000, 2001, 2003). The nested grid configuration and all model parameters, unless otherwise noted, are identical to those in Part I (cf. Fig. 3 and Table 1 in Part I), including the assimilation of Level III data from multiple WSR-88D radars (cf. Fig. 17 in Part I).

To address the many questions associated with constructing ensemble perturbations using multiple nested grids, as noted above and in Part I, three strategies are used within the SLAF framework, ranging from simply interpolating coarse-grid perturbations onto finer grids to producing dynamically consistent perturbations on the fine grids themselves. The choice of these particular approaches was based upon a logical progression from simple to complex, the desire to obtain physically meaningful perturbations, computational considerations, and the potential for operational utilization. None of the approaches is unique or optimal, yet each produces different perturbation structures and forecast results. All methods involve one or more ensemble experiments with different forecast lead times (LTs) and initialization procedures, and a total of twelve experiments are conducted, as summarized in Table 1. Among them, experiment 5 was described in detail in Part I and is used here as the baseline for comparison.

b. Perturbation method 1

Figure 1 (cf. Fig. 5 in Part I) shows a particularly simple but operationally feasible method for generating nested 3-km grid spacing ensemble forecasts assuming that no previous 3-km forecasts are available. As discussed in Part I, a 6-km control forecast (cn6) is initiated at 1800 UTC using as a background state the 6-h, 24-km forecast (P0). Similarly, the 3-km control forecast (cn3) is initiated at 2300 UTC using the 5-h, 6-km control forecast (cn6) as a background. No perturbations are applied to either of the nested control runs. WSR-88D Level III reflectivity and radial wind data, along with other observations including 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 are assimilated into the 6-km analysis at 1800 UTC and into the 3-km analysis at 2300 UTC, using the ARPS Data Assimilation System (ADAS; Brewster 1996). All initializations with ADAS are performed via a cold-start procedure (i.e., without

---

1 Hereafter, we omit the phrase “grid spacing” associated with 24, 6, and 3 km unless otherwise noted.
temporal nudging or incremental analysis updating as in Part I).

Two sets of paired perturbations for the 3-km ensembles (s1 through s4) are constructed by first interpolating onto the 3-km grid the two previous 24-km forecasts (i.e., P1 and P2), valid at 2300 UTC, and calculating their amplitude-scaled differences from the control run (cn3). The perturbations then are added to and subtracted from the new analysis at 2300 UTC to produce four initial conditions for the perturbed members. The lateral boundary conditions for the perturbed members are produced in a consistent manner.

Nine experiments (Table 1) are conducted using this methodology to better understand the influence on forecast quality of the ADAS analysis, WSR-88D Level III radar data (via ADAS), perturbation amplitude, and changes in forecast lead time. Experiments 1, 2, 3, and 5 are identical apart from the starting and ending times. In them, ADAS is applied to the initial conditions of all ensemble members, including the control, and radar data also are utilized. Experiment 4 is identical to experiment 3 except that ADAS is applied only to the control member of the ensemble.

For a majority of the experiments, the amplitudes of unscaled perturbation 1 (i.e., the difference between P1 and cn3 for all variables) are used as a reference for

---

**Table 1.** Details of the 3-km ensemble experiments. The scaling reference factor is a multiplier applied to the magnitudes of unscaled perturbation 1 (P1 – cn3). ADAS refers to whether the ARPS Data Assimilation System is applied to all ensemble members including the control (all), only the control (control), or none of the members including the control (none). Radar data refers to whether WSR-88D Level III data are included in the ADAS analysis.

<table>
<thead>
<tr>
<th>Expt.</th>
<th>Method No.</th>
<th>Begin time (UTC)</th>
<th>End time (UTC)</th>
<th>Reference factor</th>
<th>ADAS</th>
<th>Radar data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1800</td>
<td>0200</td>
<td>1.0</td>
<td>all</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2000</td>
<td>0400</td>
<td>1.0</td>
<td>all</td>
<td>yes</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2200</td>
<td>0600</td>
<td>1.0</td>
<td>all</td>
<td>yes</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2200</td>
<td>0600</td>
<td>1.0</td>
<td>control</td>
<td>yes</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2300</td>
<td>0600</td>
<td>1.0</td>
<td>all</td>
<td>yes</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>2300</td>
<td>0600</td>
<td>1.0</td>
<td>all</td>
<td>no</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>2300</td>
<td>0600</td>
<td>1.0</td>
<td>none</td>
<td>no</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>2300</td>
<td>0600</td>
<td>1.0</td>
<td>all</td>
<td>yes</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>2300</td>
<td>0600</td>
<td>1.0</td>
<td>all</td>
<td>yes</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>2300</td>
<td>0600</td>
<td>1.0</td>
<td>control</td>
<td>yes</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>2300</td>
<td>0600</td>
<td>0.5</td>
<td>all</td>
<td>yes</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>2300</td>
<td>0600</td>
<td>1.5</td>
<td>all</td>
<td>yes</td>
</tr>
</tbody>
</table>

---

**Fig. 1.** Construction of SLAF ensembles for the 3-km domain using perturbation method 1.
applying amplitude scaling (see Part I). To evaluate the impact of initial perturbation amplitude, the reference values for all fields are reduced by 50% and increased by 50% in experiments 11 and 12, respectively, relative to experiment 5. A scaling factor of 2.0 was tested at 3-km grid spacing in Part I but resulted in solution instability.

Two experiments are conducted to evaluate the impact of data assimilation. Experiment 6 is identical to experiment 5 except that radar data are excluded from ADAS, while experiment 7 excludes ADAS completely (and thus assimilated observations), using as the initial state only the background fields interpolated from the coarser-grid forecasts. For all 3-km ensemble experiments, only the Lin–Tao explicit five-category ice-phase microphysics scheme (Lin et al. 1983; Tao and Simpson 1993) is applied.

c. Perturbation method 2

Figure 2 illustrates the second method used here for constructing ensemble perturbations. Although the 6-km control run (cn6) remains the same as in method 1, two additional 6-km unperturbed forecasts (p1 and p2) are created: p1 is initiated at 1200 UTC on 28 March 2000 using the 6-h, 24-km forecast from P1 to provide the background fields, while p2 is initiated at 0600 UTC with background fields provided by the 6-h, 24-km forecast from P2. The same ADAS analysis procedure used for the control run (cn6) is applied to both p1 and p2. Two 3-km ensemble perturbations then are constructed from amplitude-scaled differences between the two previous 6-km forecasts valid at 2300 UTC (i.e., p1 and p2, interpolated onto the 3-km grid) and the 3-km analysis (cn3) at 2300 UTC. Each perturbation then is added to and subtracted from the analysis to form a pair of initial conditions (s1 through s4), resulting in a total of four perturbed members plus the control. Experiment 8 (Table 1) is produced using this method but otherwise is identical to experiment 5.

d. Perturbation method 3

The final method for generating perturbations, illustrated in Fig. 3, involves creating additional 3-km forecasts to provide dynamically consistent background fields for the perturbed ensemble members at the same grid spacing (i.e., rather than interpolating from coarser-grid output, as in methods 1 and 2). The 24- and 6-km forecasts remain the same as in method 2. One of the additional 3-km forecasts, labeled p1, is initiated at 2000 UTC using the 8-h forecast from the 6-km grid (p1) and the other, labeled p2, is initiated at 1700 UTC using the 11-h forecast from the 6-km grid (p2). Amplitude-based SLAF members are constructed at 3-km grid spacing using the differences (plus and minus) between a previous 3-km forecast and the 3-km analysis, resulting in 4 perturbed ensemble members (s1 through s4).

Two ensembles, experiments 9 and 10, are generated using perturbation method 3, both initiated at 2300 UTC on 28 March 2000. In experiment 9, ADAS is
applied to each perturbed member (s1 through s4) and to the control (cn3), while for experiment 10, ADAS is applied only to the control.

3. Behavior of error and perturbations

a. Perturbation structure

Figure 4 shows the initial potential temperature perturbations, prior to scaling, at 500 hPa from experiments 5, 8, and 9, corresponding to perturbation methods 1, 2, and 3, respectively. Regions of localized negative potential temperature anomaly correspond to observed storm location for all experiments because the analysis contains finescale thermodynamic information from radar data assimilation. For perturbation method 3, the two additional unperturbed 3-km forecasts (p1 and p2) in experiment 9 do in fact produce several storms at analysis time (not shown), appearing as positive potential temperature anomalies in the bottom panels of Fig. 4.

The same information for the near-surface (vertical level 5, approximately 125 m AGL) potential temperature and water vapor mixing ratio are shown, respectively, in Figs. 5 and 6. Less finescale structure is evident in comparison to 500 hPa, presumably as a result of surface-friction-related smoothing and the presence of explicit convection at 500 hPa. The large-scale National Centers for Environmental Prediction (NCEP) operational ensemble system, which uses bred vectors, behaves similarly in some cases (Kalnay 2003), though without resolving convection. Among the three ensemble perturbation methods tested, the structure of the perturbations is quite similar except for localized features associated with explicitly resolved storms in the background fields of experiment 9 (bottom panels in Figs. 5 and 6).

Figures 4–6 also show that perturbations 1 and 2 have comparable magnitudes even though they are formed without any scaling, using forecasts having different ages. This is reinforced by Table 2, which shows root-mean-square errors (rmses; relative to the ADAS analysis) of selected perturbations prior to scaling for the three different perturbation construction methods. If linear scaling were used, perturbation 2 (which is older) would have only half the amplitude of perturbation 1. These results support the conclusion in Part I that the linear error growth assumption associated with traditional SLAF is not appropriate for explicitly resolved deep convection in limited-area models.

b. Error growth

To document the growth of analysis errors, Fig. 7 shows several rms curves for one of the unperturbed

2 Patil et al. (2001) attribute the local low dimensionality in the 500-hPa wind vector field from the National Weather Service global ensemble system (using bred vectors) to locally growing dominant Lyapunov vectors, or “errors of the day,” which as noted by Kalnay (2003) are also contained in SLAF perturbations.
Fig. 4. Potential temperature perturbations (°C) at 2300 UTC 28 Mar 2000 for nonscaled perturbations (left) 1 and (right) 2 at 500 hPa for the 3-km ensembles using (top) perturbation method 1 (experiment 5), (middle) method 2 (experiment 8), and (bottom) method 3 (experiment 9).
Fig. 5. As in Fig. 4, but near the surface (vertical model level \( k = 5 \), approximately 125 m AGL).
Fig. 6. As in Fig. 5, but for water vapor mixing ratio (g kg⁻¹).
forecasts (p1). Errors grow very fast initially and reach saturation after approximately 1 h. As a result, the linear “age factor” scaling of traditional SLAF is inappropriate and thus amplitude-based scaling, as described in Part I, is used for all 3-km ensemble experiments in this paper. In other ensemble forecasting studies at the global or regional scales, the 500-hPa height is frequently analyzed. In our experiments, the 500-hPa height error (not shown) initially grows in a manner similar to other fields but decreases several hours into the forecast, implying the strong influence of the lateral boundaries (see Nutter et al. 2004).

The error growth pattern for the perturbed forecasts (s1 through s4) depends upon the variable considered and the amplitude of the initial perturbation. For mean sea level pressure (MSLP) and the 10-m wind (u component), the rms curves behave similar to Fig. 7 while the 850-hPa temperature shows relatively little growth, owing to a relatively large initial perturbation amplitude (figure not shown).

c. Ensemble spread

Ensemble spread, defined as the standard deviation of ensemble members relative to the ensemble mean (Hou et al. 2001), is a measure of forecast variability. In an ideal ensemble, small (large) spread indicates relative certainty (uncertainty) that the ensemble mean is close to the state of the real atmosphere (Wilks 1995), and in this regard, spread and its temporal evolution should ideally be very similar to forecast error and its growth (Hou et al. 2001).

Figure 8 shows the ensemble mean and spread of the 1-h forecast accumulated precipitation, surface radar reflectivity (computed as in Part I), and 500-hPa height for experiment 5 valid at 0000 UTC on 29 March 2000. Unlike at larger scales, where the spread has a wavelike structure similar to the ensemble mean but with a different pattern and amplitude (Tracton and Kalnay 1993; Wilks 1995), the spread in our experiments (Fig. 8b) is very similar to the corresponding ensemble mean (Fig. 8a), both of which are shaped like convective cells with overlapping centroids and comparable amplitudes. Likewise, the reflectivity spread (Fig. 8d) is similar to the corresponding ensemble mean (Fig. 8c) except that the largest values (roughly half the mean reflectivity in magnitude) are scattered largely at the edges of the storm cores and in the southeast portion of the domain, where spurious convection is present. The mean and spread of the 500-hPa height field (Figs. 8e,f), on the other hand, are notably dissimilar, with higher spreads aligning along the mesohigh (ridge) and over the region of convection.

Time series of selected domain-averaged ensemble spreads are shown in Fig. 9 for experiments 5, 11, and 12. As for the 24-km grid ensembles (Part I), an increase (decrease) in initial perturbation amplitude leads to an increase (decrease) in ensemble spread. The spreads in MSLP and 10-m zonal wind (Figs. 9a,c) are comparable to the error growth curves in Figs. 7a,c, especially with the initial perturbation reference factor of 1.5 (experiment 12). The spread in 500-hPa geopotential height [Fig. 9b; as for other large-scale fields such as 850-hPa height (not shown)] does not exhibit growth throughout the simulation, which is a notable departure from larger-scale ensembles and likely is a result of the perturbation method, small ensemble size, and relatively small domain size. Interestingly, the spread in accumulated precipitation (Fig. 9d) exhibits more significant growth during the first 3 h of the forecast because it reflects explicit storm-scale features.

### 4. Sensitivity experiments

In this section, quantitative precipitation forecasts (QPFs) from the 3-km ensemble experiments, valid at 0000, 0100, and 0200 UTC on 29 March 2000, are evaluated by grouping them according to variations in forecast lead time, methodology for creating the initial

---

3 In global ensemble forecasting, spread typically is normalized by the climatological standard deviation.
analysis, perturbation construction method, and scaling reference factor. To facilitate comparison, Fig. 10 shows the Stage IV hourly accumulated precipitation remapped onto the 3-km domain at the same times noted above. WSR-88D reflectivity data were also used in the evaluation (as in Part I), but owing to largely redundant findings with the Stage IV data, the discussion is omitted.

a. Forecast lead time

It was speculated in Part I that the relatively short (1 h) forecast lead time, along with the assimilation of radar data, contributed to the high quality of the forecast in the current experiment 5. This conjecture is explored with experiments 1, 2, and 3, which are identical to experiment 5 except that they are initiated, respectively, 5 h earlier (1800 UTC), 3 h earlier (2000 UTC), and 1 h earlier (2200 UTC). Figure 11 shows hourly uncalibrated (i.e., purely the relative frequency of occurrence without any correction) probabilities of hourly accumulated precipitation ≥2.54 mm from 0000 to 0200 UTC on 29 March 2000 for experiments 1, 2, 3, and 5. Because a 5-member ensemble is used in all experiments, probabilities are limited to 6 values: 0%, 20%, 40%, 60%, 80%, and 100%. The results indicate that forecast quality generally decreases with increasing lead time. In reality, at both 1800 and 2000 UTC, no significant storm activity was present within or adjacent to the 3-km model domain, and as a result, the assimilation of radar data had little impact in experiments 1 and 2. Nevertheless, these two ensembles still show an ability to capture some storm-scale structure, though with very low probability, while both control members (deterministic framework) predict no convection at all over the same region. Higher probabilities associated with spurious storms in the southeastern portion of the domain are produced from the control runs and members s1 and s3 (not shown) in experiments 1 and 2.

At 2200 UTC, when experiment 3 is initialized, significant radar echoes are present east of the Texas Panhandle. Figure 11 shows that experiment 3 performs somewhat better over northern Texas owing to radar data assimilation. However, it still underpredicts the intensity of storms south of Fort Worth and does not capture the second area of storms along the Texas–Oklahoma border. The probabilities also are weaker than in experiment 5.

Starting 1 h later, at 2300 UTC, experiment 5 best captures the tornadic storm event among the 4 experiments in this group, with precipitation cores better aligned with those in Fig. 10 and having higher probabilities. The second area of convection northwest of the main line is also captured. Nevertheless, the southern portion of the principal north–south line weakens shortly after forecast initiation, presumably as a result of its proximity to the inflow boundary [similar behavior has been noted in other experiments and in coarser-
Fig. 8. The (left) ensemble mean and (right) spread in experiment 5 of the (top) hourly accumulated precipitation, (middle) surface reflectivity, and (bottom) 500-hPa height, valid at 0000 UTC 29 Mar 2000. The figures in the boxes depict maxima.
grid spacing daily forecasts produced at the Center for Analysis and Prediction of Storms (CAPS)]. Both experiments 3 and 5 generate spurious storms over the southeastern part of the domain with probability values up to 60%.

b. Data assimilation options

In experiment 6, all five ensemble members are initialized without radar data, though other observational data are assimilated as in experiment 5 (see Part I). Hourly accumulated precipitation probabilities (Fig. 12) indicate that in the absence of radar data, the predicted storms are widely scattered and have low probabilities. Interestingly, some storms do form over the central and northern portions of the domain, in general agreement with observations and contributed solely by the perturbed members. The control member only produces scattered storms over the eastern Texas–Oklahoma border some 4–6 h into the forecast (beyond the times shown in Fig. 12). The broad, high-probability region of spurious convection in experiment 6 over eastern Texas is the result of perturbed members s1 and s3. This suggests that these storms are not associated with radar data assimilation but rather are controlled by mesoscale features embedded within the background fields generated from coarser-resolution fore-
Fig. 11. Hourly uncalibrated probability of hourly accumulated precipitation ≥2.54 mm from 0000 to 0200 UTC 29 Mar 2000 for the 3-km ensembles initiated at (top) 1800 (experiment 1), (second row) 2000 (experiment 2), (third row) 2200 (experiment 3), and (bottom) 2300 UTC (experiment 5). Forecast LT (h) is shown in parentheses above each panel and the plus sign indicates the location of Fort Worth.
casts, augmented by ensemble perturbations as discussed in Part I.\textsuperscript{5}

To assess the impact of data assimilation via ADAS, Fig. 12 shows hourly accumulated precipitation probabilities for experiment 7, in which all ensemble members are initiated directly from interpolated background fields without the assimilation of new observations. As anticipated, the control member does not generate storms until some 5 h following initialization (i.e., at 0400 UTC, beyond the times shown in Fig. 12). Perturbed members $s_2$ and $s_4$ (perturbations added) behave similarly to the control run while $s_1$ and $s_3$ (perturbations subtracted), however, produce storms immediately upon forecast initiation, thus contributing to the hourly accumulated precipitation probability field. Nevertheless, the broad and spurious region of precipitation over eastern Texas remains.

Experiments 6 and 7 suggest that the absence of radar data, and the failure to assimilate new observations into the background analysis, can in this particular case lead to a poor deterministic forecast (here measured by the performance of the control run). However, an ensemble with as few as five members, using appropri-
ately perturbed initial and boundary conditions, was able to capture some signal associated with convection, though with a low probability of occurrence. Considering that deep convection at a given point in space and time has a relatively low probability of occurrence, this result suggests that probabilistic rather than deterministic forecasting of convection is likely to be of the greatest operational utility, in agreement with the notions put forward by Brooks et al. (1992).

In operational global ensemble forecasting (e.g., Toth and Kalnay 1997), data assimilation is applied only to the control member so that perturbations representative of analysis errors can be identified. Should this same approach be used in Storm Scale Ensemble Forecasting (SSEF), where cloud microphysics processes and radar data assimilation play potentially crucial roles, the latter by adding storms that should be present and removing those that should not? Experiments 4 and 10, which are identical to experiments 3 and 9 (see Table 1) except that ADAS data assimilation (including radar data) is applied only to the control member (cn3), address this question. Overall, these two sets of experiments produce quite similar hourly accumulated precipitation forecasts (figures not shown), but with much lower probabilities than when data assimilation is applied to all members. One of the principal reasons involves a delay in convective development owing to model start-up in the absence of radar data assimilation. However, assimilating data in forecasts other than the control arguably runs counter to the notion that perturbations are meant to represent analysis errors. More work is needed to fully address this issue.

c. Perturbation construction strategy

All ensemble experiments described previously were constructed using a simple approach, perturbation method 1, in which no unperturbed forecast at 3-km grid spacing other than the control member is required and with perturbations for the 3-km grid calculated from the 24-km unperturbed forecasts by interpolating differences between them and the analysis to finer grid domains. The rationale for this method is rooted in the fact that at most operational prediction centers, only relatively coarse-grid forecasts are routinely available. However, perturbations generated using coarser grids would not be expected to represent analysis errors on finer grids. To address this issue, additional experiments are made using increasingly sophisticated methods for creating the initial perturbations (cf. section 2 and Fig. 2).

Figure 13 shows hourly accumulated precipitation probabilities from experiments 8 and 9 (along with experiment 5 for comparison), which utilize perturbation methods 2 (Fig. 2) and 3 (Fig. 3), respectively. Visually, the probabilities in experiments 5 and 8 are quite similar except that experiment 8 produces somewhat smaller values over a more limited area. Although no general conclusions can be drawn, these results suggest that additional forecasts created at intermediate grid spacing—here at a spacing (6 km) that is problematic with regard to convective closure—may add little value relative to interpolation from coarser spacing (24 km), where closure is better defined.

Using perturbation method 3, two more unperturbed forecasts are easily produced at 3-km grid spacing (Fig. 3) so that initial perturbations for the 3-km ensembles are computed directly from the difference fields between the 3-km unperturbed forecasts (p1 and p2, valid at the analysis time) and the current analysis. In this manner, no interpolation is required because unperturbed forecasts are available for each grid. Additionally, such perturbations should reflect analysis errors and yield better dispersion of the ensemble forecasts. The obvious drawback to operational implementation is that unperturbed forecasts would need to be generated routinely for every nested grid, a significant computational burden and possibly a waste of resources during periods when convection is absent.

Comparing experiment 9 with experiment 5 in Fig. 13, the hourly accumulated precipitation probability fields are quite similar, though the initial perturbations in the former contain more small-scale structure, as revealed in Figs. 4–6. Additional quantitative evaluations in the next section provide further assessment of the impact of different perturbation methods.

d. Amplitude-scaling reference factor

As mentioned previously, experiments 1–10 utilize a reference amplitude of unity for perturbation scaling. For larger-scale flow and coarser-model grid spacings, ensemble forecast skill has been shown to exhibit sensitivity to perturbation amplitude (e.g., Toth and Kalnay 1997). To assess such impact at the storm scale, we repeat experiment 5 (scaling factor of unity) using a perturbation amplitude reduced by 50% (experiment 11, scaling factor of 0.5) and increased by 50% (experiment 12, scaling factor of 1.5). The former leads to a more spatially compact precipitation pattern having higher probabilities while the latter yields a much broader precipitation pattern with lower peak probabilities (figures not shown). Overall, the precipitation in experiment 11 appears to agree best with observations, with minimal spurious convection in the southeastern part of the domain.
The ensemble spread in experiment 11 is significantly underdispersed (Fig. 9), with all members being nearly identical (not shown). Experiment 12 (with a scaling factor of 1.5), on the other hand, exhibits spread comparable to the corresponding rms values. However, it also exhibits the greatest amount of spurious convection in the southeastern part of the domain (not shown). Also noteworthy is that for the relatively few ensemble members in this study, a poor relationship should exist between ensemble spread and error, owing to low bias in the spread (Murphy 1988). These results, in combination with those from the 24-km grid spacing experiments described in Part I, suggest that a perturbation amplitude–scaling factor of unity is justified in this particular storm case.

5. Analysis of forecast skill

The qualitative discussion of the preceding section is now expanded to include verification statistics for assessing the skill of all 3-km ensemble experiments. Two traditional scores, the rmse⁶ and equitable threat score (ETS), are applied to hourly accumulated precipitation forecasts for both individual members and the ensemble mean. Two measures of probabilistic forecast skill, the Brier score (BS) and ranked probability score (RPS), are calculated for hourly accumulated precipitation.

---

⁶ Hereafter, the terms error and rmse are used loosely, for in fact, they properly refer to the innovation (difference between forecast and observation).
tation and surface radar reflectivity. ETS is also assessed for the latter. The application of these scores follows Part I and their mathematical definition and interpretation can be found in Wilks (1995), Stensrud et al. (2000), and Hou et al. (2001). Because complete hourly Stage IV accumulated precipitation data are available for our case, the predicted accumulated precipitation for all experiments is verified every hour for the entire forecast period (i.e., 8 h for experiments 1 through 4 and 7 h for all others).

Figure 14 shows rmse of hourly accumulated precipitation for each ensemble member as well as the ensemble mean for all twelve experiments. Because the abscissa refers to simulated time, the UTC verification times differ among some experiments. As noted below, rmse is not sufficiently robust for application to explicitly resolved convection owing to the associated highly intermittent nature of the associated precipitation. For example, the very low rmse values in the early portion of the forecast for experiments 1 and 2 (Figs. 14a,b) are misleading because no significant precipitation occurred within the 3-km domain prior to 2300 UTC (not shown).

Nevertheless, several important observations can be made from Fig. 14. First, the ensemble means, in general, are more skillful than the control forecasts with the exception of the first few hours in experiments 1, 2, 6, and 7 and are more skillful than the individual members as well. Second, the two negative (s1 and s3) perturbation forecasts tend to be similar, the same being true for the two positive (s2 and s4) perturbation forecasts. Finally, the relative accuracy of individual members (or member clusters) tends to change with time.

The rmse curves of all ensemble means, except for experiments 11 and 12, are replotted in Fig. 15 for actual verification times (2300–0600 UTC), and thus only portions of experiments 1 and 2 are shown. Experiments 5, 8, and 9 (representing the three different perturbation methodologies) exhibit similar trends, with experiment 5 superior from 0100 to 0400 UTC. Experiments 4 and 10, in which the ADAS analysis is applied only to the control forecasts, have smaller rmse values than their counterparts, experiments 3 and 9, for which ADAS is applied to all members. This is contrary to the result shown in the previous section. Ironically, even experiments 6 and 7, in which no radar data are assimilated, exhibit smaller rmse values than experiments 5, 8, and 9.

These results raise, among other things, questions regarding the appropriateness of using rmse for highly intermittent phenomena. First of all, fine-grid forecasts containing realistic looking, finescale features can exhibit larger rmse than their coarser-grid counterparts that produce smoother, less realistic looking solutions owing to the so-called double penalty phenomenon. Additionally, a forecast can yield very low rmse values simply because it produces little precipitation (as in our experiments 1, 2, 6, and 7) compared to a forecast like experiment 5, which exhibits convective structures closer to observations but, owing to amplitude and phase errors, has a larger overall error. These same problems also occur for categorical error measures such as ETS (e.g., Mass et al. 2002; Sousounis et al. 2004; Marshall et al. 2004).

ETS measures the skill, relative to a random forecast, of predicting the area of hourly accumulated precipitation greater than or equal to a given threshold. The larger the ETS (up to unity), the more skillful the forecast. ETS > 0 implies skill relative to a random forecast while ETS ≤ 0 means no skill. A perfect forecast yields ETS = 1. As in Du et al. (1997) and Stensrud et al. (2000), we calculate ETS for precipitation thresholds of 0.254, 2.54, 12.7, and 25.4 mm.

Figure 16 shows, for experiment 5, the ETS of hourly accumulated precipitation for individual forecast members and the ensemble mean (see also Part I). For the lowest two thresholds (0.254 and 2.54 mm), all members are skillful (ETS > 0) except for the final hour of the forecast. Skill deteriorates with increasing threshold and forecast lead time. Based upon this statistic, the ensemble mean does not exhibit greater skill than the control member throughout most of the forecast period—a behavior that differs from large-scale ensemble forecasting (Toth and Kalnay 1997). Instead, one or more perturbed members exhibit (from time to time) higher scores than the control (e.g., s1 in Fig. 16a and s3, s4 in Fig. 16c). For the 25.4-mm threshold, the ensemble mean essentially has no skill while some members (e.g., cn3 and s3) exhibit marginal skill over portions of the forecast period.

Figure 17 shows, for all 12 experiments, ETS of hourly accumulated precipitation ≥ 2.54 mm. Experiments...
Fig. 14. Rmse of accumulated precipitation for individual 3-km ensemble members and the mean computed using the Stage IV analysis. Abscissa is the forecast L.T.
ments 1 and 2 (Figs. 17a,b) exhibit no forecast skill for nearly all members throughout the entire period while experiments 5, 8, 9, and 11 (Figs. 17e,h,i,k) yield the highest skill up to 6 h. In general, for experiments 3–5, 8, and 9–12, the skill scores deteriorate with forecast lead time except between 3–4 h and during the final hour. ETS in experiments 6 and 7 (Figs. 17f,g) indicates that even without the ADAS analysis and/or the assimilation of radar data, some ensemble members as well as the ensemble mean yield skill following a 1–2-h spinup process, in agreement with our qualitative evaluation in the previous section. Figure 17 also shows that for all experiments, the ensemble means do not exhibit skill superior to their corresponding control members throughout most of the forecast period, in contrast to large-scale ensemble forecasting, as noted previously. Instead, for each experiment, one or more perturbed members exhibit (from time to time) higher scores than their corresponding control members (e.g., Figs. 17c,d,f,g).

Unlike the rmse in Fig. 15, the relative forecast skill among experiments, measured by their ensemble mean ETS as shown in Fig. 18, is consistent with the qualitative evaluation of the previous section, especially during the most active period of convection from 0000 to 0200 UTC 29 March 2000. Experiments 5, 8, and 9, which include radar data assimilation for all members along with consistent perturbations and short lead times, exhibit the greatest skill. Experiments 1 and 2, with very long lead times and thus without the benefit of radar data owing to the lack of storms at initialization time, are the least skillful (exhibiting no skill for the ensemble means because ETS < 0). Applying the ADAS analysis to all members increases the ETS as experiment 3 is more skillful than experiment 4 and experiment 9 is more skillful than experiment 10. Another finding from Fig. 18 that contrasts with the behavior of rmse in Fig. 15 is that experiments 8 and 9,
which utilize more complex methods for creating the initial perturbations, yield better skill than the baseline case, experiment 5, which utilizes the simplest approach (perturbation method 1).

For probabilistic QPF, the Brier score, also called the probability score (PS) in early publications, is one of the most common verification metrics. It measures the difference between the forecast probability of an event
and its occurrence in observations (see Brier 1950; Murphy 1973) and is essentially the mean-square error (mse) of the probability forecast. In this study, we calculate the BS for hourly accumulated precipitation using five categories (separated by thresholds of 0.254, 2.54, 12.7, and 25.4 mm).

Figure 19 shows the BS for all experiments except 11 and 12. The BS generally decreases during the first 6–7 h, except for experiments 3 and 4, indicating an improvement in skill for the latter. Unfortunately, the BS suffers from the same limitations as rmse. Consequently, an unambiguous determination of skill is not possible and more effective measures are needed.

Figure 20 evaluates the impact on forecast skill of perturbation scaling for experiments 5, 11, and 12. Once again, rmse and ETS yield contradictory results, with experiment 12 (scaling factor of 1.5) having a lower (better) rmse during most of the forecast period and experiment 11 (scaling factor of 0.5) having a larger (better) ETS. The two probabilistic scores, BS and RPS, are similar, with experiment 12 slightly more skillful during the first 3.5 h but the least skillful thereafter.

6. Summary and outlook

Our principal goal in this two-part study was 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 determin-
coarser-grid forecasts. An analysis of error growth supported the finding in Part I that the scaling of initial SLAF perturbations, in the context of explicitly resolved convection, should be based upon error amplitude. The ensemble spread of forecast accumulated precipitation and radar reflectivity exhibited structures different from nonprecipitation variables and from fields in large-scale ensemble forecasting. Unlike the latter, in which ensemble spread exhibits wavelike patterns with amplitudes comparable to the ensemble mean, precipitation-related spread at fine-grid spacing mimicked the corresponding ensemble mean (i.e., resembles storms) with comparable amplitude.

In general, forecasts started prior to the existence of convection had relatively lower skill (see exception below), although certain perturbed members (i.e., other than the control) still showed some predictive skill, though with very low probabilities. Experiments initiated directly from the background field, that is, without the assimilation of other data including those from the WSR-88D, also exhibited less skill. However, even in the absence of assimilated data and with relatively primitive methods for creating initial perturbations, ensemble forecasts of explicitly resolved convection exhibited potentially greater operational value than a single deterministic forecast by virtue of highlighting locally intense events. Forecasts having 1- to 2-h lead times, with the assimilation of new observations including Doppler radar data, generally performed the best.

We evaluated the impact of assimilating radar data into both the control run and ensemble members, which represents a departure from tradition in large-scale ensemble forecasting. Without radar data assimilated into all members, some did not contain perturbations of sufficient amplitude to trigger deep convection. This inevitably reduced spread and thus hindered the provision of meaningful storm-scale forecasts.

Finally, we examined the impact of initial perturbation amplitude and showed that a scaling factor of 0.5, relative to the value of unity used in the control run, leads to a more concentrated pattern of accumulated surface precipitation with higher probabilities but with significant underdispersion. A scaling factor of 1.5 produced a much broader (visually less optimal) precipita-

![Fig. 20. Comparison of ensemble skill scores for accumulated precipitation ≥ 2.54 mm with different perturbation scaling factors: (a) rmse and (b) ETS for the ensemble means, (c) BS, and (d) RPS. The reference factors are labeled next to the experiment identifier.](image-url)
tion pattern with lower peak probability values, but with spreads comparable to the corresponding rmse. This, however, should be interpreted with caution owing to the fact that small ensemble sizes tend to produce small spreads. Overall, it appears for this storm case that ensemble forecasts of explicitly resolved convection, regardless of how they are initialized, have the potential for greater skill and more operational value than a single deterministic forecast, represented here by the associated control experiment. This result supports other studies that note the value of an ensemble strategy for intense local weather (e.g., Brooks et al. 1992; Elmore et al. 2002a, b, 2003), and such strategies are now being explored as options for the yearly real-time spring prediction experiments conducted by CAPS in collaboration with the National Oceanic and Atmospheric Administration (NOAA) National Severe Storms Laboratory and Storm Prediction Center.

We have only begun in this two-part study to examine the many questions associated with fine-grid ensemble forecasts, here in the context of deep convection. In future work, the fundamental characteristics and growth of errors on fine scales must be examined as a means for understanding predictability and informing techniques for generating initial perturbations. The former is within reach for deep convection via the now decade-long archive of WSR-88D data, and consideration should be given to creating a U.S.-scale reanalysis, at 1-km grid spacing, to better understand the natural variability of high-impact local weather. Larger forecast domains, linkages among multiple grids, and the use of both coarser- and finer-grid spacing for ensembles must be examined as well, and a wide variety of physical scenarios ranging from strong to weakly forced must be studied. Additionally, comparisons of skill in the explicit prediction of deep convection between full dynamic models, such as the one here used, and statistically based approaches (e.g., Mueller et al. 1998) should be performed. Although prevailing notions suggest that the former are perhaps slightly better at early times during the forecast period (e.g., Wilson et al. 1998), the assimilation of radar and other finescale observations into dynamic models is expected to reduce model spinup and thus increase their relative performance. Perhaps most importantly, quantitative verification metrics more appropriate to intermittent phenomena are needed to assess both the skill and practical value of both ensemble and deterministic forecasts.

Acknowledgments. This research was funded by the National Science Foundation under Grants ATM99-81130 and ATM01-30396 for the second author. 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. Comments from two anonymous reviewers greatly improved the quality of the manuscript.

REFERENCES


