Operational Ensemble Cloud Model Forecasts: Some Preliminary Results

KIMBERLY L. ELMORE*

Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma

STEVEN J. WEISS

NOAA/Storm Prediction Center, Norman, Oklahoma

PETER C. BANACOS

NOAA/Storm Prediction Center, Norman, Oklahoma

12 December 2002 and 21 February 2003

ABSTRACT

From 15 July through 30 September of 2001, an ensemble cloud-scale model was run for the Storm Prediction Center on a daily basis. Each ensemble run consisted of 78 members whose initial conditions were derived from the 20-km Rapid Update Cycle Model, the 22-km operational Eta Model, and a locally run version of the 22-km Eta Model using the Kain–Fritsch convective parameterization. Each ensemble was run over a 160 km × 160 km region and was valid for the 9-h period from 1630 through 0130 UTC. The ensembles were used primarily to provide severe-weather guidance. To that end, model storms with lifetimes greater than 60 min and/or a sustained correlation of at least 0.5 between midlevel updrafts and positive vorticity (the supercell criterion) were considered to be severe-weather indicators. Heidke skill scores, along with the true skill statistic, are between 0.2 and 0.3 when long-lived storms or storms meeting the supercell criteria are used as severe-weather indicators. Equivalent skill scores result when modeled and observed storms are categorized by lifetime and supercell characteristics and compared with expertly interpreted radar data.

1. Introduction

From 17 April through 8 June 2001, the National Weather Service Storm Prediction Center (SPC) performed a spring research program aimed at exploring ways to improve short-term forecasts of convective initiation and evolution. Substantial improvements in convective forecasts should lead directly to increased lead time in SPC severe-local-storm watches. As an add-on to this program, on 76 days from 15 July through 30 September, an ensemble cloud model (the 2001 Ensemble) was run daily to determine whether the ensemble output could help SPC forecasters to anticipate the nature of any convection that might occur in a selected region. Soundings extracted from mesoscale numerical weather prediction models were used to initialize the cloud-scale model. By definition, the ensemble provides conditional forecasts: *if convection occurs,* the ensemble is designed to yield information about the range of storm behaviors that might be expected. The ensemble is based directly on Elmore et al. (2002a), which presents evidence that ensemble cloud models might have potential benefits for operational forecasters. However, the dataset used in Elmore et al. (2002a) was a priori limited to 18 days, and on those days convection was known to have occurred. Such a small dataset is insufficient to determine what, if any, benefit might result from supplying forecasters with ensemble model output on a routine, daily basis. We hope that any benefits an ensemble cloud model might provide to operational forecasters will become apparent in the operational test. The natural interaction between research and operational meteorologists during the test also provides operational forecasters the opportunity to guide applied research in a direction that best suits their particular needs.

Operational environments are arguably the most demanding ones in which to apply numerical models be-

* Additional affiliation: NOAA/National Severe Storms Laboratory, Norman, Oklahoma.

Corresponding author address: Kimberly Elmore, NSSL, 1313 Halley Circle, Norman, OK 73069.
E-mail: Kim.elmore@noaa.gov

cause interpretation is performed by many different meteorologists and users, and so any systematic peculiarity or deficiency is likely to be noticed eventually. Model developers may occasionally ignore such characteristics because they have become accustomed to seeing, and appropriately interpreting, them. Not so the operational forecaster, who is understandably less forgiving of poor or misleading guidance. Hence, any numerical model is subject to close scrutiny when placed in an operational environment. A numerical cloud model clearly should not be used for explicit guidance, as demonstrated in Elmore et al. (2002b).

The 2001 Ensemble also provides an opportunity to investigate how the cloud model responds to soundings extracted from different mesoscale models. For example, one mesoscale model may provide soundings that never result in deep convection while another model provides soundings that never fail to generate deep convection.

The next section briefly describes the operational system employed for the test. Section 3 discusses the various scoring criteria and methods. Section 4 presents the results and discusses some implications of these results, and section 5 draws conclusions and proposes future work.

2. Operational system

Similar to Elmore et al. (2002a), the operational exercise uses an ensemble that incorporates soundings extracted over a relatively small 160 km × 160 km area. The ensemble consists of 78 individual runs of the cloud-resolving Collaborative Model for Mesoscale Atmospheric Simulation cloud model (COMMAS; Wicker and Wilhelmson 1995) using a 1.25-km horizontal grid spacing and 43 levels on a vertically stretched grid. The horizontal extent of the COMMAS domain is 100 km × 100 km, which should not be confused with the size of the mesoscale model domain from which soundings are extracted. COMMAS is always initialized with 100 km × 100 km horizontally homogeneous conditions for a single sounding, and the convective process is always initiated with a 3.5-K warm bubble, regardless of sounding characteristics. A significant difference between the previous cloud model ensemble work and this work is that the ensemble region is usually chosen based on a forecast instead of being run over regions known to have generated convection. The forecast is based on the SPC day-2 outlook product. However, three cases are deliberately chosen over areas where convection was *not* expected. The SPC lead and outlook forecasters are queried about where they think the most likely region of severe convection will be or where they are most interested in the model results. For simplicity, either an airport or station identifier near the center of this region is chosen as the anchor location for the ensemble region. A modeled storm is defined when vertical velocity w is at least 8 m s^{-1} anywhere within the cloud model domain for at least 10 min, based on the work presented in Elmore et al. (2002a). To meet the time constraints inherent in an operational setting, the Kessler (1969) warm-rain-process parameterization is used. Hence, no ice processes, or the latent heat released by them, are included. This latent heat energy may be especially important when modeling or forecasting convection associated with low static stability and low relative humidity in the boundary layer (Wicker et al. 1997). Other implications are considered in Elmore et al. (2002a) and references therein.

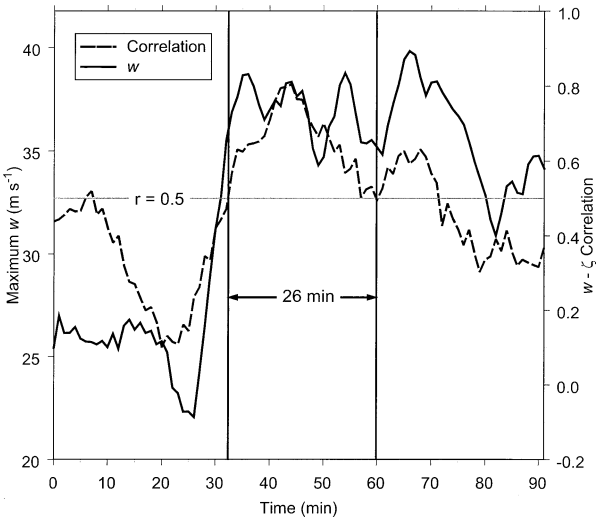

FIG. 1. Example of a simulated storm that meets the supercell criteria. Updraft speed is shown by the solid trace and the scale on the left; correlation values are shown by the dashed trace and the scale on the right.

New for this work is a supercell criterion, which is used to determine if modeled storms posses a midlevel rotating updraft. In general, a positive correlation between midlevel w and midlevel vorticity is used as a supercell indicator in numerical cloud models (Davies-Jones, 1984; Weisman and Klemp 1984). Three conditions compose the supercell criteria: 1) the modeled storm must last at least 40 min (this threshold prevents possible false alarms from simulated storms with brief lifetimes), 2) the correlation between positive vorticity and updrafts of at least 1 m s^{-1} at the 5.3-km level (an arbitrary choice representing the midlevel of the storm) must be at least 0.5 (Davies-Jones 1984; Weisman and Klemp 1984), and 3) the correlation must remain at least 0.5 for at least 20 min (Fig. 1). The last threshold prevents false alarms from random effects and modeled storms that exhibit very brief periods of rotation. If the ensemble contains at least one member that meet these criteria, it is considered a strong severe-weather indicator.

For the operational exercise, initial soundings are derived from a combination of three mesoscale models: the operational Eta Model (OE), a locally run version of the Eta Model that uses the Kain–Fritsch convective

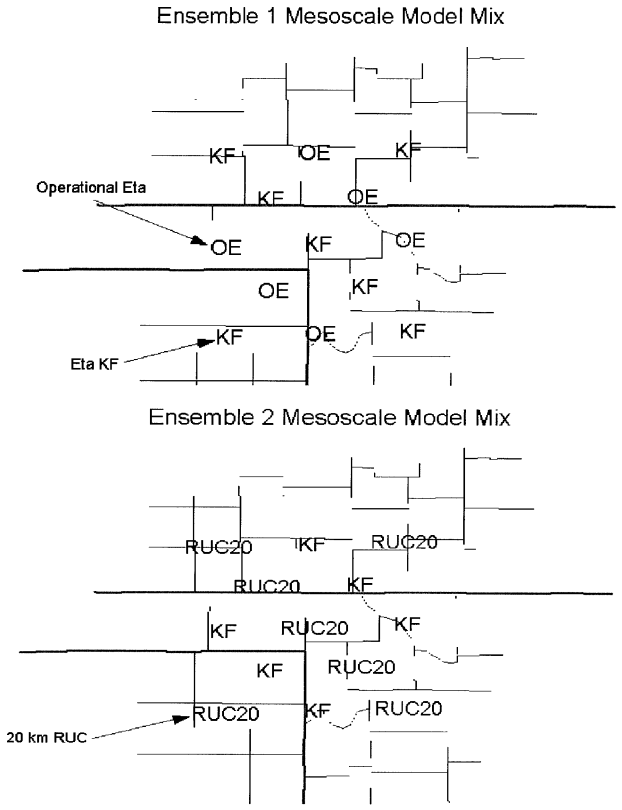

FIG. 2. Spatial distribution of the soundings taken from the different mesoscale models and used as initial conditions for the ensemble. Data are provided on the AWIPS 212 grid. Every other grid point is used, starting in the lower-left corner and progressing up each column, which explains the staggered appearance. Here, OE marks a sounding taken from operational Eta output, KF marks a sounding taken from the Eta KF output, and RUC20 marks a sounding taken from the 20-km RUC Model.

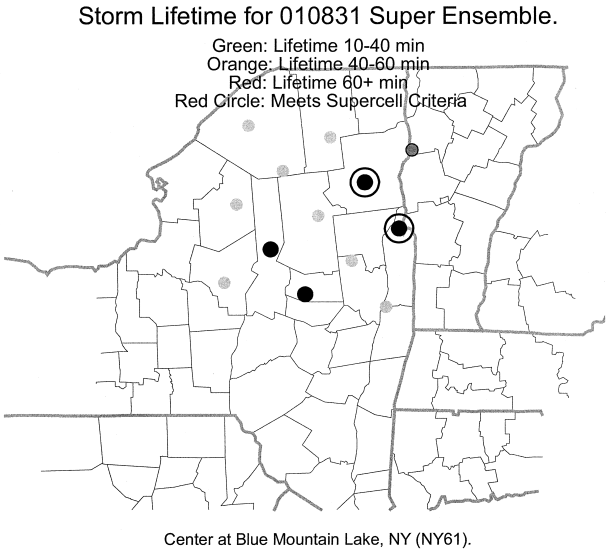

FIG. 3. Black-and-white example of the map-type ensemble output display. The large black dots with circles are displayed in red, the medium-gray dot represents orange, and the light-gray small dots are green. If no convection occurs at a grid point, an open circle is displayed.

parameterization (KF; Kain and Fritsch 1990), and a beta version of the Rapid Update Cycle Model (RUC; Bleck and Benjamin 1993) with 20-km grid spacing, which has subsequently become the operational RUC20. All model data are provided on the Advanced Weather Interactive Processing System (AWIPS) 212 grid (40-km horizontal spacing) with 25-hPa vertical spacing, and each sounding consists of all levels from a single grid point. Three sets of soundings, valid at 1800, 2100, and 0000 UTC of the day-2 outlook, are extracted from each model. The 78 separate runs are divided into two sets of 39 runs each, to facilitate sequential executions on a 40-node Beowulf (Sterling et al. 1999) cluster.

The first set of 39 members (ensemble 1) is constructed by spatially alternating soundings between the KF and the OE; the second set (ensemble 2) alternates soundings between the RUC20 and KF models (Fig. 2). Using this configuration, the KF soundings are oriented such that none are duplicated. This simple scheme results in an overrepresentation of the KF when compared with either the OE or the RUC20. Results from each ensemble are presented separately but are also combined into a single, ''super-'' ensemble consisting of all 78 members.

The ensemble runs on a Beowulf cluster that consists of 40 nodes, each with a 450-MHz Intel, Inc., Pentium-III processor, 66-MHz front-side bus, and 192 MB of memory. Hence, the hardware is pedestrian and far from state of the art. Each ensemble member is run on a single node, which results in perfect scaling for the ensemble. Each ensemble set requires about 3.5 h to complete on the cluster. Runs that use the OE and KF commence at roughly 0600 UTC; runs that use the RUC20 must wait until the 0600 UTC RUC is complete and usually start by 1000 UTC. All output is typically available by 1400 UTC. Processing is done by the first author manually, and processed information is available to forecasters by 1600 UTC.

Output is placed on an Internet Web page to be viewed as needed by SPC forecasters at their convenience. This output is divided into three sets to show the results of the two separate ensemble runs and the combined superensemble. All output is displayed using a stylized map, developed in cooperation with SPC forecasters. Also provided is a display of the three main vertical-velocity time series modes based on principal components analysis (Elmore and Richman 2001), the raw maximum vertical velocity from each ensemble member (akin to a spaghetti plot), and a kernel density estimate (Silverman 1986) of the storm lifetime probability density function. Of these displays, the stylized map was most commonly utilized (Fig. 3). In this color display, open circles or colored dots are drawn at every grid point from which soundings are taken. Open circles show that no deep convection occurs within the ensem-

TABLE 1. Example text discussion for 31 Aug 2001.

010831 Cloud Ensemble Run Discussion
General
The ensemble region is centered on Blue Mountain Lake, NY (NY61). Each ensemble displays remarkably different characteristics. The operational Eta soundings produce five long-lived storms, and at least two of these meet supercell criteria. The other Eta KF and RUC20 soundings produce mainly short-lived storms. There is conflicting information between each ensemble concerning where storms may be most likely: ensemble 1 indicates storms are equally likely everywhere, while ensemble 2 indicates that storms are most likely over the southern half of the ensemble region.
Interpretation
Long-lived storms are generated from either the 2100 or 0000 UTC soundings. All long-lived storms display relatively weak vertical velocities (less than 20 m s^{-1}), including those that meet supercell criteria. The storms that meet supercell criteria also tend to have relatively low tops (roughly 12 km) and so could represent low-topped supercells or minisupercells. Only soundings from the operational Eta generate long-lived storms or storms that meet supercell criteria. The RUC20 produces only a few, weak, short-lived storms on the southern edge of the ensemble region. Hence, the RUC20 indicates storms will be most likely on the southern half of the region. Soundings from the Eta KF produce a few short-lived storms with weak vertical velocities. If the operational Eta forecasts valid at 2100 and 0000 UTC from the 0000 UTC run are considered plausible, then the occurrence of severe storms also appears plausible. Based on storms meeting supercell criteria and the existence of long-lived lifetime modes, combined, severe weather reports are expected in and around the ensemble region during the period from 1630 to 0130 UTC.

ble from any of the soundings at a given point; colored dots show the longest-lived storm generated by any sounding from that grid point. Green is used for lifetimes of less than 40 min, orange is used for lifetimes between 40 and 60 min, and large red dots are used for lifetimes greater than 60 min. A red circle around the dot indicates that the supercell criteria have been met.

A brief text discussion is also produced each day by the first author. This discussion describes the general behavior of the ensemble, notes any systematic differences between ensemble members that appear to be linked to the mesoscale model that supplies the initial conditions, and provides any relevant information about the ensemble run itself, such as if all mesoscale models are represented and if all ensemble members are present. The discussion also provides information about when the modeled convection occurs, for example, primarily 2100 UTC and after, or commencing with the 1800 UTC soundings. An example of the text discussion is shown in Table 1.

3. Verification

Based on the longest-lived model storm at any grid point within the superensemble, two different verification methods are employed, report based and radar based, each with different qualities. For report-based verification, any modeled storm that meets the supercell criteria or, as in Elmore et al. (2002a), has a lifetime of at least 60 min is considered to be a forecast for severe storms. The verification region consists of two boxes that contain the 160 km $\times$ 160 km region. The inner box is 240 km $\times$ 240 km centered on the ensemble region and so extends 40 km from the sides of the initial region; the outer box is 400 km $\times$ 400 km, also centered on the ensemble region (120-km extension from the sides). This verification process assumes that the 160 km $\times$ 160 km prediction grid represents a somewhat larger region for operational forecast purposes. The two verification boxes reflect slightly different approximations of this process. A ''hit'' is counted if a severe report derived from the SPC daily local severe report log (e.g., tornado, wind gust of at least 25 m s^{-1}, or hail of at least 1.9-cm diameter) is contained anywhere within the box of interest and the ensemble produces a long-lived storm or supercell-type storm. A ''miss'' is counted if a report of severe weather occurs in the box but no long-lived storms or storms meeting the supercell criteria are generated within the ensemble, and vice versa for a false alarm. A correct null is self-evident.

The radar-based verification method is based on storm lifetime by category. The categories are no thunder, thunder, medium lived, long lived, and supercell. There is no short-lived category, because it is implicit in the thunder category. Cloud-to-ground lightning strikes from the National Lightning Detection Network are used to determine the occurrence of deep convection within the verification domain. Radar-observed lifetime is that period for which the maximum composite reflectivity exceeds 40 dBZ_e. Radar-observed storms are categorized as medium lived if none lasts longer than 60 min but at least one lasts longer than 40 min. The long-lived category is for storms that last longer than 60 min, and the supercell category is for storms with supercell characteristics. Storms are qualitatively categorized based on radar data interpretation by SPC personnel using archived level-III Weather Surveillance Radar-1988 Doppler (WSR-88D) data. Similar categories are easily defined for the modeled storms within the ensemble.

First the criteria used to fill in the 2 $\times$ 2 contingency matrix is defined (Wilks 1995). Forecasts are divided into the following categories: no thunder, thunder, medium lived, long lived, and supercell. Table 2 defines the ''yes'' criteria for forecasts and observations. From this table, the ''no'' criteria become immediately apparent.

With these definitions, 2 $\times$ 2 contingency matrices are easily constructed from which the bias, probability of detection (POD), false-alarm ratio (FAR), critical success index (CSI), true skill statistic (TSS), and Heidke skill score (HSS) can be computed (Wilks 1995). The criteria expressed in Table 2 are chosen to yield the best HSS and TSS values. To obtain some insight into the reliability and stability of the various statistics, the data are resampled with replacement using 1000 replications (Efron and Tibshirani 1993). All values are shown with 95% confidence bounds for the median. To investigate

TABLE 2. Criteria used to construct the verification 2 × 2 matrix.

Category	Forecast yes	Observed yes
No thunder	No grid points produce storms	No lightning in ensemble region
Thunder, lifetime < 40 min	One or more grid points produce any category of storm	Lightning and echoes > 40 dBZ_e in ensemble region
Medium lived, 40 min < lifetime < 60 min	One or more grid points produce medium-lived storms	Lightning and echoes > 40 dBZ_e lasting less than 60 min
Long lived, lifetime > 60 min	One or more grid points produce long-lived storms	Lightning and echoes > 40 dBZ_e lasting longer than 60 min
Supercell, meets supercell criteria	One or more grid points meet the supercell criteria	Lightning and echoes > 40 dBZ_e displaying supercell characteristics

how sensitive the ensemble behavior is to soundings from the different mesoscale models, results for individual mesoscale models are extracted and analyzed separately. The same scores computed for the superensemble are also computed for ensembles generated from each mesoscale model. Ensemble scores that vary widely between the different mesoscale models suggest sensitivity to the soundings produced by the models. For example, if the OE provides much better scores for long-lived storms then are achieved by the superensemble, then most of the useful characteristics within the initial conditions are due to the OE. If, on the other hand, scores based on individual mesoscale models do not differ greatly from the superensemble scores, or tend to be worse, then the contributions from all the mesoscale models are necessary for optimal performance.

4. Results

a. Report-based verification

Using modeled long-lived storms as an indicator of severe weather is moderately successful. The TSS is slightly better for the 40-km extension than the 120-km extension, HSS is practically identical for the two, POD and FAR both decrease, and CSI increases with increasing verification domain size (Fig. 4). Modeled storms that meet the supercell criteria, although less common than long-lived storms, have slightly higher TSS and HSS scores as severe-weather indicators (Fig. 5). The CSI for long-lived storms as a severe-weather indicator is slightly higher than for supercell storms. The increased TSS and HSS scores for supercells are due to a decreased FAR in the face of a slightly lower POD when compared with long-lived storms. The values of these statistics compare favorably with similar scores computed for watches issued in 2001 by the SPC. These values also compare favorably with National Weather Service (NWS) warning verification statistics from 2001.

Long-lived storms used as severe-weather indicators are associated with a positive bias that depends on the verification region (Fig. 6). In particular, there are nearly 2 times as many days with modeled long-lived storms as days with severe weather within the small region (which results in a bias value of almost 2), mirrored by the higher FAR. However, the number of days with long-lived modeled storms and number of days with severe weather within the large region are nearly equal, which yields a bias value of nearly 1. In contrast, there are fewer days with modeled supercell storms than days with severe reports. There is reason to suspect that the number of observed supercells is negatively biased because of limitation of the radar data products used in the verification process. Even so, this result seems reasonable because, even if the supercell criteria lead to roughly the same proportion of supercells in the model world as in the real world, not all severe-weather reports come from supercell storms. Hence, days on which severe weather occurs are more numerous than days with supercell storms.

b. Radar-based verification

When the ensemble output is verified against radar data, slightly different scores and interpretations result. A radar-based storm is defined as a cell with at least 40-dBZ_e maximum reflectivity accompanied by cloud-to-ground lightning. Using the ensemble results as an indicator that thunderstorms will occur results in deceptively good scores (Table 3). These scores are deceptive because the sample is biased toward an unusually high likelihood for convection because, except for three cases, all ensemble regions are chosen based on the expectation of convection. Hence, if there is any skill in anticipating convection by SPC forecasters, this skill will act to bias results toward a high likelihood of convection. This seems to be the case because, out of 76 days, only 12 days produce no convection in the ensemble region. One way to gauge this score is simply to assign a forecast of thunder to each day, regardless of the ensemble results. In this case only the HSS and TSS indicate problems because both are 0. The POD is, of course, 1, while FAR is only 0.079 and CSI *increases* to 0.921. Even the bias remains reasonable. These results are reminiscent of the Finley affair presented in Murphy (1996). A better measure of ensemble skill at anticipating convection should result if the ensemble were run over a single region (or many single regions simultaneously) and results were compared with the climatological values for that region. There is no

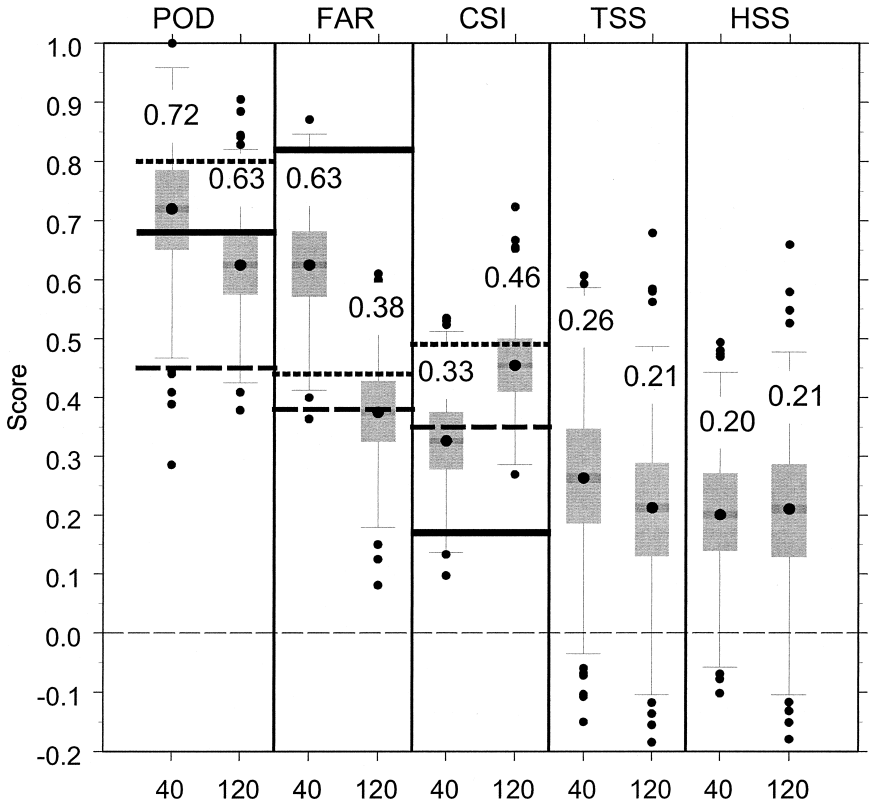

FIG. 4. Verification statistics for long-lived (lifetime > 60 min) modeled storms as severe-weather indicators. Numbers along the *x* axis indicate the scoring region used: 40 indicates values for the region that extends 40 km out from each side of the ensemble grid, and 120 indicates a scoring area that extends out 120 km from each side of the ensemble grid. The box-and-whisker plot indicates the stability of each value for 1000 bootstrap resamples. The large black dot and associated number show the median (almost identical to the mean) value, the dark gray bar shows the 95% confidence interval for the median value, the light gray box encompasses the inner quartile, and the whiskers extend out ±1.5 times the inner quartile range. Individual dots show outliers that are beyond 1.5 times the inner quartic range. For reference to more familiar processes, the horizontal, heavy solid lines represent verification values for NWS 2001 tornado warnings; the horizontal, heavy dotted lines represent verification values for NWS 2001 severe-thunderstorm warnings; and the horizontal, heavy dashed lines represent SPC 2001 watch verification values for all severe-weather and tornado watches, combined. TSS and HSS values are not available because the NWS and SPC do not consider ''correct null'' forecasts.

current evidence that the ensemble skillfully identifies when convection is likely.

An alternative forecast is one for no convection, which is defined as an ensemble run for which none of the grid points generate deep convection. Forecasts for no convection may be informative even in the face of the obvious sample bias. Hence, the ability to determine that storms will not occur requires more skill. Used this way, the ensemble may have skill as an indicator that convection is unlikely (Table 3).

Scores for radar-based verification of the long-lived and supercell categorical forecasts are shown in Fig. 7. The FAR is lower and the CSI is considerably higher for long-lived and supercell storms than for medium-lived storms. The skill scores for long-lived and supercell storms hover between 0.2 and 0.3. The CSI is slightly better, and the HSS and TSS scores are slightly worse for long-lived storms than for supercell storms. Thus, the skill of the ensemble in identifying days on which supercells will occur is slightly better than in identifying days on which long-lived cells will occur. However, the scores for medium-lived storms is poor, which indicates that the ensemble has no skill in identifying days on which medium-lived storms occurred. This result is similar to the results reported in Elmore et al. (2002a), in which forecasts for medium-lifetime storms displayed the least skill.

The bias scores for medium-lived, long-lived, and supercell storms are 0.5, 1.53, and 1.19, respectively, meaning that the ensemble tends to produce far too few medium-lived storms, too many long-lived storms, and only a few too many supercells as compared with the observed frequency. The verification dataset is certainly biased toward areas that contain convection, but any

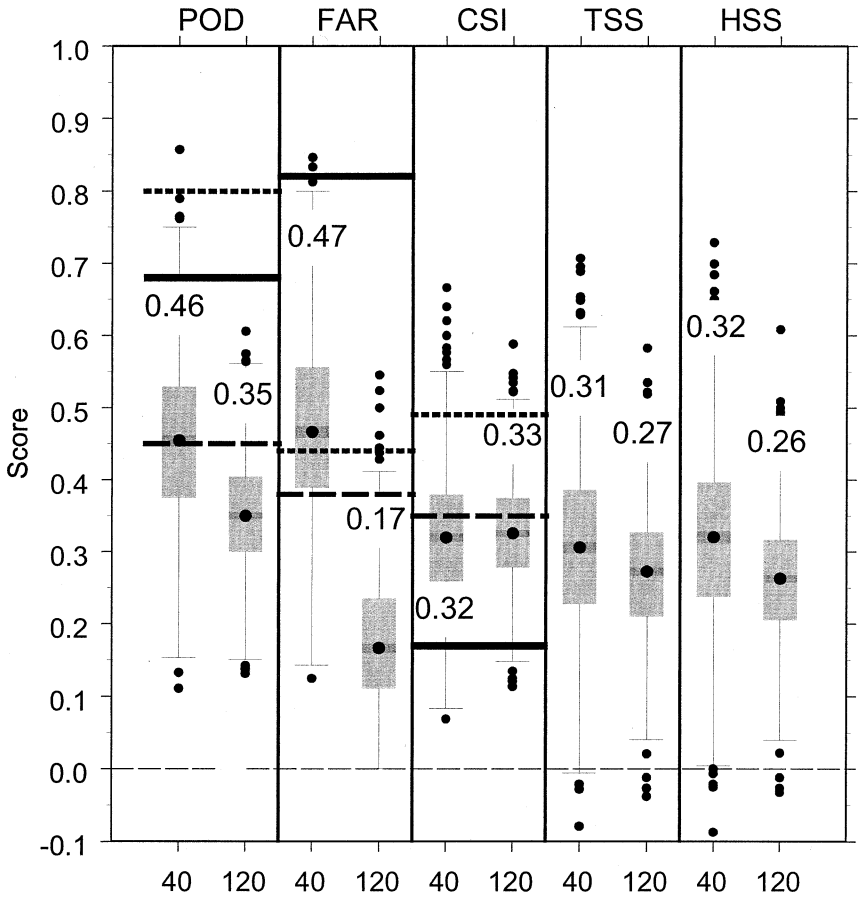

FIG. 5. Same as in Fig. 5 but for modeled storms that meet the supercell criteria.

bias for longer-lived storms versus shorter-lived storms or supercells versus nonsupercells is far less obvious. Hence, it seems likely that, when conditioned on the existence of convection, the cloud-scale ensemble skillfully depicts when long-lived storms and supercells are likely.

c. Mesoscale model dependency

All results presented so far are derived from ensembles that use the combined initial conditions from three different mesoscale models. Reasonable questions arise: How dependent are these results on the mesoscale models used for the initial soundings? Are initial conditions from all three models needed for the best scores, or is all of the needed information contained in the initial conditions from only one or two of the models? Which model(s) provide the most information? To address these questions, the radar-verified results are stratified by mesoscale model.

Stratifying results by mesoscale model is straightforward. However, because of the way the ensemble is constructed, the KF model is overrepresented. For each day, 39 members are based on the KF while only 21 members are based on the RUC and 18 are based on the OE. Hence, the effect of the individual models is likely weighted in some (possibly nonlinear) way based on these proportions.

None of the models performs particularly well on medium-lived storms, which is consistent with results from Elmore et al. (2002a). Performance is better for long-lived storms, for which the POD is highest for the superensemble, FAR is lowest for the RUC, and the OE provides the best CSI, TSS, and HSS (Fig. 8). Clearly, mixing initial conditions from the KF or the RUC with the OE has an overall detrimental effect.

When results for supercells are examined, the OE clearly contains most, if not all, of the information available to the ensemble. None of the soundings from the RUC ever produces a modeled storm that meets supercell criteria, and only two soundings from the KF produce storms that meet supercell criteria. The KF unfortunately produces these soundings on days on which no supercells are observed within the ensemble region. Hence, the POD is 0 and the FAR is 1 for the KF (aside from POD = 0, no scores can be computed for the RUC). Thus, the scores for the superensemble are lower than for the OE alone.

Some further investigation reveals that a combination of characteristics inherent to the OE and the nature of

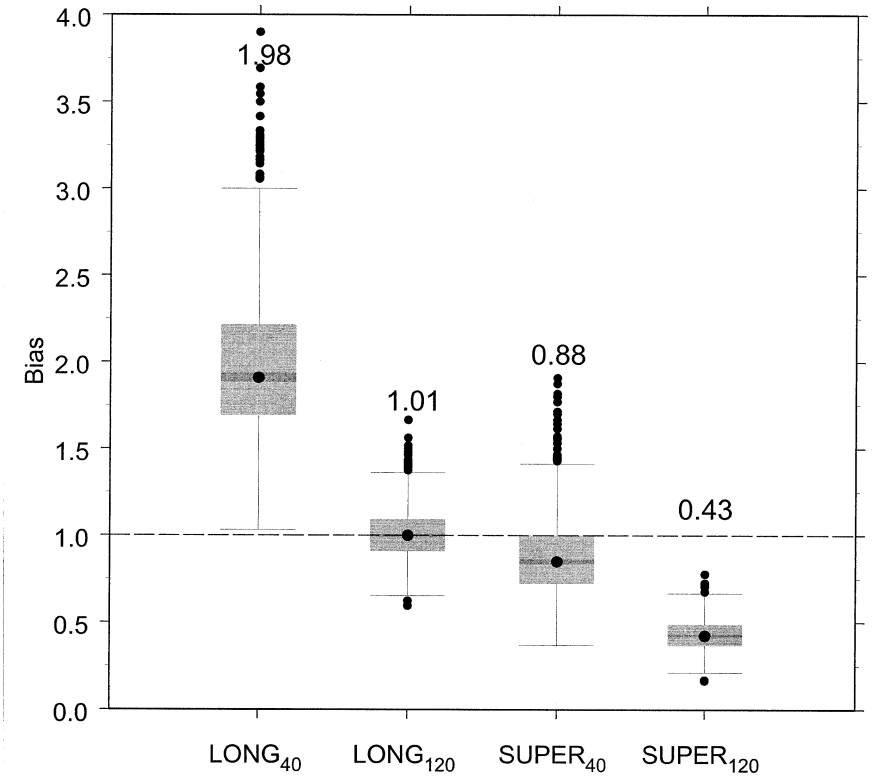

FIG. 6. Same as in Fig. 6 but showing bias values for modeled long-lived and supercell storms as severe-weather indicators, based on severe-weather reports. ''Long'' and ''super'' indicate values for modeled long-lived and supercell storms, respectively.

the cloud model are together at least partly responsible for this result. In almost all cases, soundings from the OE differ from KF and RUC soundings in one obvious, significant way: because of the differing ''shallow convection'' schemes in the mesoscale models (Baldwin et al. 2002), a low-level inversion is almost always eradicated within OE soundings but tends to be maintained in both the KF and RUC.

A typical example is shown in Fig. 9. Both soundings have similar kinematic characteristics, and both contain very large surface-based convective available potential energy (CAPE) values (5800 J kg^{-1} for the operational Eta vs 4500 J kg^{-1} for the KF). Clearly, the warm-bubble initiation scheme used in the COMMAS numerical cloud model cannot overcome the inversion contained in the KF sounding. Even if a bubble warm enough to overcome the inversion evident in the KF sounding were used to initiate deep convection, the resulting modeled storm would probably not continue for very long in the face of such a strong inversion.

The other primary difference between soundings produced by the OE and the KF is that the low-level moisture tends to be significantly richer in the OE soundings (Baldwin et al. 2002). Moisture effects are only mildly apparent in this example but can lead to significantly higher CAPE values for Eta soundings. Hence, although storms may form in soundings from all three mesoscale models, storms will tend to be stronger and longer lasting in soundings that come from the OE.

TABLE 3. Scores for the cloud model ensemble when interpreted as either forecasts for convection or forecasts for no convection. Numbers in parentheses for thunder indicate the score resulting if all days were treated as forecasts for convection.

Event	POD	FAR	CSI	Bias	HSS	TSS
Thunder	0.929 (1.000)	0.015 (0.079)	0.915 (0.921)	0.943 (1.086)	0.762 (0)	0.584 (0)
No thunder	0.833	0.500	0.455	1.67	0.762	0.584

5. Summary and conclusions

Verification scores based on severe reports appear similar to those based on radar data. When severe reports are used for verification, modeled storms that are long lived and, in particular, modeled storms that meet the supercell criteria tend to be good indicators that severe weather is likely should convection occur. In addition, when long-lived storms and supercells occurred in the ensemble, they were likely to be observed in the radar data.

Two different-size regions are used for report-based verification. The resulting skill scores display little dependence on the region size, even though the outer re-

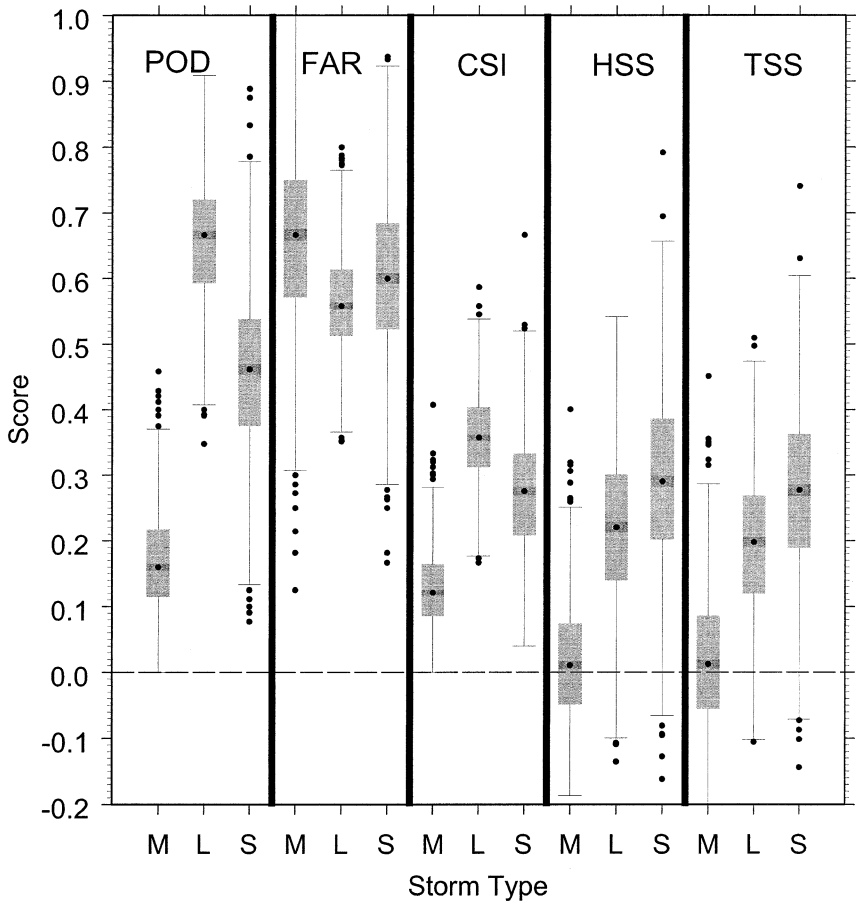

FIG. 7. Resampled verification statistics for radar-verified storm types: M is medium lifetime, L is long lived, and S is supercell.

gion is 2.8 times as large as the inner region. There seems to be no strong evidence of the appropriate area over which the ensemble results are valid, but more work is obviously needed. Overall, the ensemble may be used over an area at least as large as 400 km $\times$ 400 km.

For medium-lifetime storms, there appears to be a mild dependence on the mesoscale model from which soundings are extracted because some models produce soundings that result in better scores than others. Even so, for all but TSS, the best scores for medium-lived storms are obtained when ensemble members resulting from all three mesoscale models are combined. Differences between mesoscale models are much more obvious for long-lived storms, because almost all of the information resulting from the ensemble comes from members that use OE soundings. Within this dataset, only the OE provides information about supercell storms. In fact, for supercells, removing the KF and RUC results in better overall scores.

The mesoscale model dependence may be interpreted various ways. One interpretation is that the KF/RUC soundings more properly represent large-scale environments, because large uncapped regions east of the Rocky Mountains that do not ultimately host deep convection are seldom observed. The OE soundings may be interpreted as more representative of small-scale conditions for which convection initiates and can be maintained. These results are serendipitous, but useful information is still contained in the cloud-scale ensemble. These results also suggest that the ensemble is best used as conditional guidance.

Forecaster comments on the ensemble were overwhelmingly positive. Forecasters used the ensemble output mainly to confirm their own expectations; no outlooks or forecasts were changed based on the ensemble output, but forecasters felt much more confident of their interpretations in light of the ensemble results. With more experience, and if the ensemble could be based on the day-1 outlook instead of the day-2 outlook, some forecasters believed they would use the ensemble to adjust their convective outlooks.

One recurring comment from forecasters concerns timeliness. In the ideal case, forecasters want to run the ensemble interactively and on demand over arbitrary regions throughout the day. Forecasters desire such a

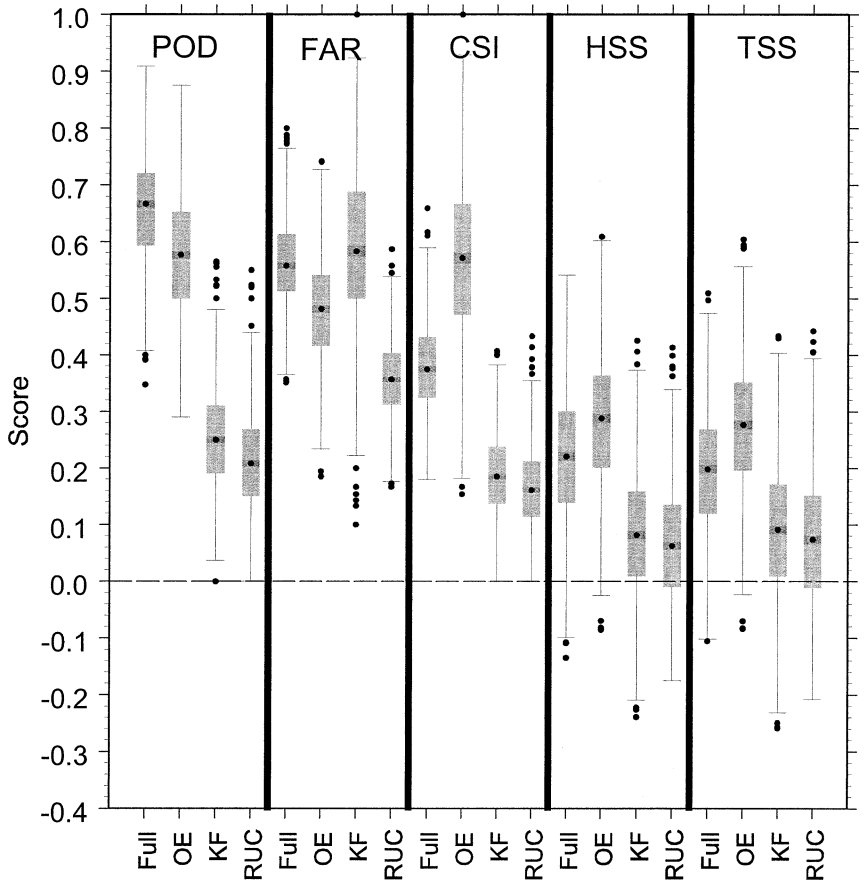

FIG. 8. Verification scores decomposed by mesoscale model for long-lived storms. The *x* axis yields the initial condition source: ''Full'' indicates the score for the superensemble, ''OE'' shows the score for the operational Eta Model, ''KF'' shows the scores for the Eta KF Model, and ''RUC'' shows the score for the 20-km RUC.

product to be available interactively and more frequently, with less latency (defined as the period between the submission of the ensemble run and the availability of results). Latency is always attributable to hardware limitations. The hardware used for this experiment was left over from the 2000 U.S. census activities and is far from optimal for the computational task. However, processed output from the ensemble could be available within 1 h after the ensemble run is submitted if state-of-the-art compilers merged with fast, commercial, off-the-shelf hardware configured as a LINUX cluster are used. Hence, the desire for frequent, interactive ensemble runs at arbitrary locations is currently achievable.

There remains, however, the issue of initial soundings. Both the RUC and KF tend to maintain low-level inversions while the OE tends to eradicate them. The ensemble is intended to be used for conditional convection forecasts, and if the warm-bubble initiation continues to be used, soundings with low-level inversions appear to have limited value. How best to approach this problem is an area of ongoing research.

There is no evidence that information about the likelihood of convection is available from the ensemble, which is hardly surprising. There is, however, some evidence that the ensemble may provide guidance about when convection is particularly *un*likely. Convection initiation remains poorly understood, and warm bubbles are unlikely to represent properly the natural processes. Yet, the cloud-scale ensemble does appear to provide useful conditional guidance about the nature of storms that may be expected should deep convection occur within the ensemble region. In particular, interpreting storm lifetime as an indicator for severe convection appears to be a useful way to distill the ensemble information. The existence of modeled storms that meet the supercell criteria also appears to be a useful indicator that severe convection is likely should convection occur.

Future storm-scale ensembles will derive initial conditions from an ensemble of mesoscale models. Doing so means that mixing in soundings from different times may no longer be necessary. Initial conditions for the ensemble will probably continue to be obtained from an area, so an areawide interpretation of the output will continue to be used. However, faster hardware allows the ensemble to be run interactively by forecasters using Web-based applets. Thus, forecasters will be able to

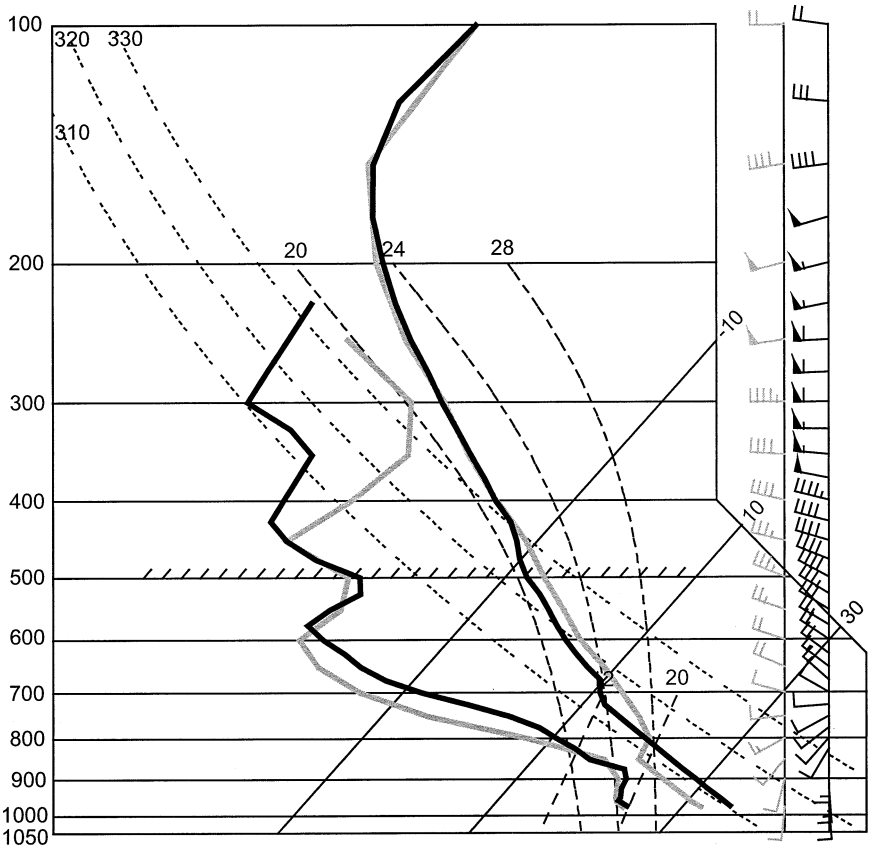

FIG. 9. Skew *T*–log*p* plots of soundings separated by 53 km from the operational Eta (black) and the Eta KF (gray) for 2100 UTC 17 Jul 2001. The operational Eta sounding results in a modeled storm that lasts 92 min, whereas the Eta KF sounding results in no convection.

examine ensemble results in the most interesting regions only a few hours before storms are expected to form.

Acknowledgments. The authors thank three anonymous reviewers whose comments and suggestions materially enhanced this work. We also thank Dr. Louis J. Wicker for making his COMMAS cloud model available for this work. We thank Ms. Sarah K. Jones and Ms. Shanna J. Sampson for their help in analyzing these data. We thank the Storm Prediction Center staff for accommodating this experiment and for their carefully considered comments, and we especially thank Paul Janish for his help in setting up the Web-based system for displaying results. Many thanks to Michael Baldwin and Steven Fletcher for helping to set up the data acquisition and ensemble execution scripts. Thanks are also given to Brett Morrow for his able administration of the Beowulf LINUX cluster and for his enthusiastic help in utilizing it.

REFERENCES

Baldwin, M. E., J. S. Kain, and M. P. Kay, 2002: Properties of the convection scheme in NCEP's Eta Model that affect forecast sounding interpretation. *Wea. Forecasting,* **17,** 1063–1079.

Bleck, R., and S. G. Benjamin, 1993: Regional weather prediction with a model combining terrain-following and isentropic coordinates. Part I: Model description. *Mon. Wea. Rev.,* **121,** 1770–1785.

Davies-Jones, R., 1984: Streamwise vorticity: The origin of updraft rotation in supercell storms. *J. Atmos. Sci.,* **41,** 2991–3006.

Efron, B., and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap*. Chapman and Hall, 436 pp.

Elmore, K. L., and M. B. Richman, 2001: Euclidean distance as a similarity metric for principal component analysis. *Mon. Wea. Rev.,* **129,** 540–549.

——, 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.

Kain, J. S., and J. M. Fritsch, 1990: A one-dimensional entraining/detraining plume model and its application in convective parameterization. *J. Atmos. Sci.,* **47,** 2784–2802.

Kessler, E., 1969: *On the Distribution and Continuity of Water Substance in Atmospheric Circulation. Meteor. Monogr.,* No. 32, Amer. Meteor. Soc., 83 pp.

Murphy, A. H., 1996: The Finley affair: A signal event in the history of forecast verification. *Wea. Forecasting,* **11,** 3–20.

Silverman, B. W., 1986: *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 175 pp.

Sterling, T. L., J. Salmon, D. J. Becker, and D. F. Savarese, 1999: *How to Build a Beowulf*. MIT Press, 239 pp.

Weisman, M. L., and J. B. Klemp, 1984: The structure and classi-

fication of numerically simulated convective storms in directionally varying wind shears. *Mon. Wea. Rev.*, **112,** 2479–2498.
Wicker, L. J., and R. B. Wilhelmson, 1995: Simulation and analysis of tornado development and decay within a three-dimensional supercell thunderstorm. *J. Atmos. Sci.*, **52,** 2675–2703.
——, M. P. Kay, and M. P. Foster, 1997: STORMTIPE-95: Results from a convective storm forecast experiment. *Wea. Forecasting*, **12,** 388–398.
Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction.* Academic Press, 467 pp.