J3.3 Operational Ensemble Cloud Model Forecasts: Some Preliminary Results

Kimberly L. Elmore[1], Steven J. Weiss[2], Peter C. Banacos[2] and Sarah K. Jones[3]

1. Introduction

As an extension of the Storm Prediction Center's (SPC) 2001 Spring Experiment, an ensemble cloud model (Elmore et al. 2002) was run on a daily basis from 15 July through 30 September (76 days) to determine if the ensemble could help SPC forecasters anticipate the nature of any severe weather that might occur in a selected region. Each day, the SPC forecasters selected a region of interest for the ensemble. This ensemble region was chosen on the preceding evening, based on the Day 2 outlook.

As in Elmore et al., (2002) the ensemble was run over a relatively small 160 x 160 km area. A significant difference between the previous cloud model ensemble work and this work is that the ensemble region is based on a forecast, instead of being run over regions known to have generated convection. Hence, this experiment is carried out in an operational context. The ensemble was run with the explicit goal of providing operational forecasters with a timely forecast product. This paper describes some results from this operational exercise.

2. Model

As in Elmore et al., (2002) the Collaborative Model for Mesoscale Atmospheric Simulation cloud model (Wicker and Wilhelmson 1995) is used for the operational exercise. Convection is always initiated with a 3.5K warm bubble, regardless of sounding characteristics. To meet the time constraints inherent in an operational setting, the Kessler (1969) autoconversion parameterization is used.

The ensemble runs on a Beowulf cluster that consists of 40 nodes, each with a 450 MHz Intel Pentium™ III processor, 66 MHz front-side bus, and 192 MB of memory. Hence, the hardware is pedestrian by current standards. Each ensemble member is run on a single node, which results in perfect scaling for each ensemble. Each ensemble consists of 39 members and two runs are made, each requiring about 3.5 h to complete on the cluster. Runs commence at roughly 0600 UTC and all output is typically available by 1400 UTC.

Unlike previous work, a supercell criteria is defined. Three conditions comprise the supercell criteria: 1) the modeled storm must last at least 40 min, 2) the correlation between positive vorticity and $w > 1$ m s^{-1} at the 5.3 km level must be at least 0.5, and 3) the correlation ≥ 0.5 must last for at least 20 min (Fig. 1).

3. Initial conditions and output

As in Elmore et al., (2002) initial conditions consist of 39 soundings extracted from a 5 x 5 AWIPS 212 grid covering a 160 x 160 km region valid at 1800, 2100 and 0000 UTC. Unlike previous work, three mesoscale models are used. They are the Operational Eta (OE), a locally-run version of the Eta using the Kain-Fritsch (Kain and Fritsch 1990) convective parameterization (KF), and a beta version of the rapid update cycle model (RUC, Bleck and Benjamin 1990) with 20 km grid spacing (RUC20).

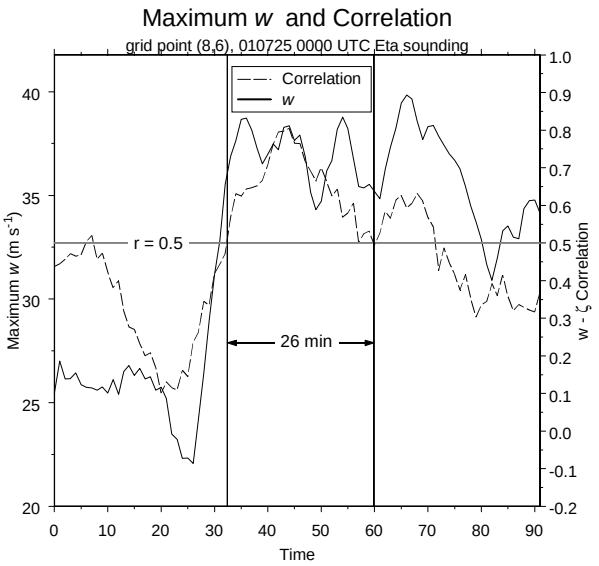

Figure 1. Example of a simulated storms that meets the supercell criteria. Updraft speed is shown by the solid trace and the scale on the left, while correlation values are shown by the dashed trace and the scale on the right.

Two 39-member ensembles are run. The first consists of a mixture of OE and KF while the second consists of a RUC20 KF mixture (Fig. 2). These two ensembles are presented separately but are also merged into a large, 78-member "super" ensemble. Because initial conditions from three models are dispersed throughout two separate ensembles, one of the models (KF) is over-represented.

Output was placed on a web page to be viewed as needed and when convenient by SPC forecasters. Output was displayed using a stylized map, developed in cooperation with SPC forecasters. Also provided was a display of the three most similar vertical velocity time series based on PCA analysis (Elmore and Richman 2000), the raw vertical velocity from each ensemble member, and a kernel density estimate (Silverman 1986) of the storm lifetime probability density function. Of these displays, the stylized map was most commonly used (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 ensemble from any of the soundings at

1. Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma/National Severe Storms Laboratory
Corresponding Author Address: 1313 Halley Circle, Norman, OK 73069, elmore@nssl.noaa.gov.
2. Storm Prediction Center, Norman, OK.
3. University of Oklahoma, Norman, OK

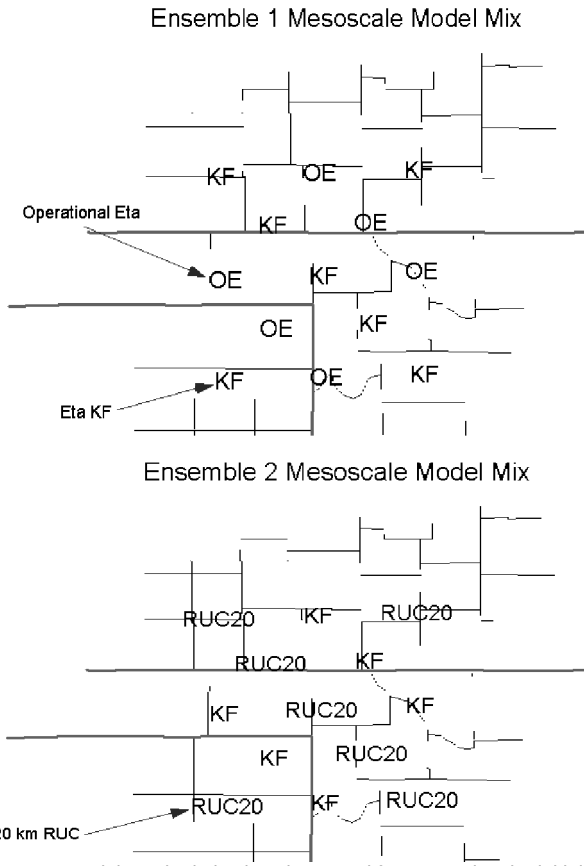

Figure 2. Models and relative locations used for generating the initial condition soundings. OE is the operational Eta model, KF is the Eta model using the Kain-Fritsch convective parameterization, and RUC20 is the 20 km RUC model.

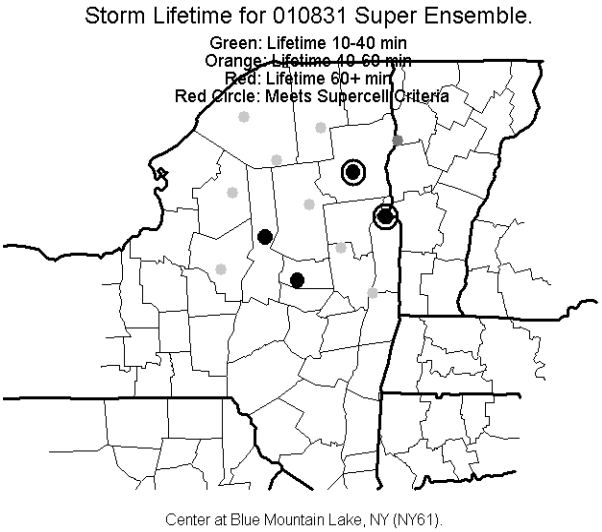

Figure 3. Black-and-white example of the map-type ensemble output display. The large, black dots with circles are displayed in red, the medium-gray small dot is typically orange and the light-gray small dots are green. If no convection occurs at a grid point, an open circle is displayed.

a given point. The characteristics of the 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 for lifetimes between 40 and 60 min, and large red dots for lifetimes greater than 60 min. A red circle around the dot indicates that the supercell criteria have been met.

In addition, a brief text discussion was produced each day. The text described the general behavior of the ensemble, noted any systematic differences between ensemble members that appeared linked to the mesoscale model supplying the initial conditions, and provided information about when the modeled convection occurs, i.e., primarily 2100 UTC and after, commencing with the 1800 UTC soundings, only occurring within soundings from the OE, etc.

4. Verification

Based on the longest-lived storm at any grid point within the super ensemble, two different verification methods are employed, each with different qualities. In the first case, storms that meet the supercell criteria or, as in Elmore et al. (2002), storms with lifetimes of at least 60 min are used as indicators of severe storms within two boxes containing the 160 x 160 km region. The first box extends 40 km from the sides of the initial region while the second extends 120 km from the sides. The same scoring is then performed for any forecast storms meeting the supercell criteria. For these, a “hit” is counted if a severe report is contained anywhere within the box of interest. A “miss” is counted if a severe weather report 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 second verification method uses archived Level III WSR-88D data. For this method, the data are used to categorize storm type and approximate lifetimes. Storm type is either supercell or non-supercell, and storm lifetimes are No Thunder, Thunder (convection of any type), Medium for storms with 40-60 min lifetimes, Long for storms with lifetimes greater than 60 min, and Supercell for those modeled storms that meet the supercell criteria.

If 12 or more grid points produce no modeled convection and none occurs within the ensemble region, it is considered a correct No Thunder forecast. A forecast for convection occurs if three or more grid points produce any modeled convection, and a hit is counted if any convection occurs within the ensemble region. If at least one grid point produces a medium-lived modeled storm and at least one storm with a lifetime greater than 40 min occurs within the ensemble domain, a hit for medium-lived convection is counted. A similar strategy is used for long-lived and supercell storms. The existence of supercell storms is deduced qualitatively based on expert interpretation by SPC personnel.

With these definitions 2 x 2 contingency (or confusion) 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) are all computed (Wilks 1995). The data are bootstrap resampled with replacement using 1000 replications to obtain some insight into the reliability and stability of the various sta-

tistics (Efron and Tibshirani 1993). All values shown lie within the 95% confidence bounds for both the mean and the median.

5. Results

Typically, the OE tends to provide soundings that were more likely to support deep convection within the cloud model, followed by the RUC20. Soundings from the KF tended to be the least likely to support deep convection within the cloud model. Cursory examinations reveal that the OE tends to completely eradicate any low level inversions while the KF often maintained a low-level inversion over the 1800-0000 UTC forecasts. The RUC20 was generally somewhere in between, sometimes eliminating any low level inversions and sometimes retaining them. The KF seldom supported long-lived or supercell modeled convection. The OE often produced soundings that supported long-lived convection and was responsible for most of the modeled supercells. The RUC20 occasionally generated soundings leading to long-lived modeled storms, but seldom provided soundings that lead to modeled supercells.

Using modeled long-lived storms as indicators for severe weather is moderately successful (Table 1).

Table 1: Report-based skill scores for modeled long-lived and supercell storms as severe report indicators.

Region	POD	FAR	CSI	TSS	HSS	Bias
Long 40 km	0.71	0.63	0.33	0.26	0.20	1.91
Long 120 km	0.63	0.38	0.36	0.21	0.21	1.00
Supercell 40 km	0.45	0.47	0.32	0.31	0.32	0.85
Supercell 120 km	0.35	0.18	0.33	0.27	0.29	0.42

There appears to be little difference between TSS and HSS for the 40 km extension and the 120 km extension, and POD, FAR and CSI all behave predictably. Using long-lived storms as an indicator of severe weather displays a positive bias that is strongly dependent on the verification region used. There are nearly twice as many days with modeled long lived storm than days with severe reports within the 40 km extension region, mirrored by the higher FAR. The number of days with long-lived modeled storms and number of days with severe reports within the 120 km extension region are nearly equal, since the bias values are very nearly 1.

Modeled storms meeting the supercell criteria, while less common than long-lived storms, have slightly higher TSS and HSS scores as severe weather report indicators (Table 1). The CSI for long-lived storms as a severe report indicator is slightly higher than for supercell storms. The increased TSS and HSS scores is due to a decreased FAR in the face of a slightly lower POD compared to long-lived modeled storms. There are fewer days with modeled supercell storms than days with severe reports. If the supercell criteria leads to even roughly the same proportion of supercells in the model world as in the real world, this result is reasonable because not all severe weather reports come from supercell storms. These statistics compare quite favorably with similar scores that can be computed for the current suite of SPC products.

Using radar data for verification statistics leads to slightly different scores and interpretations. For the radar-derived verification, scores are also computed for forecasts of No Thunder, Thunder, Medium and Lon-Lived storms, and Supercells (Table 2). Apparently, an excellent indicator of whether or not any convection will take place within the ensemble region is obtained when three or more soundings at different locations generate deep convection, which is defined as convection that lasts at least 10 min. Also, when used this way, the number of days on which convection is observed and the number of days when convection is indicated by the ensemble model are nearly equal. Three grid points yields optimum values for all of the scores used here.

An alternative forecast is one for no convection, which is defined as an ensemble run for which at least 12 grid points (the optimal value) generate no deep convection. Used this way, the ensemble is an excellent indicator of when convection is not likely, though the number of days for which the ensemble indicates no convection is biased high compared to observations. Even so, these results alone are compelling.

Table 2: Radar-based skill scores for modeled long-lived and supercell storms as severe report indicators.

Region	POD	FAR	CSI	TSS	HSS	Bias
No Thunder	0.83	0.50	0.46	0.76	0.58	1.67
Thunder	0.91	0.02	0.90	0.75	0.54	0.93
Medium	0.76	0.26	0.60	0.26	0.26	1.02
Long	0.65	0.56	0.35	0.21	0.19	1.50
Supercell	0.47	0.59	0.28	0.30	0.29	1.13

Scores are not as impressive, but are still respectable, for the various categories of storms. A forecast for medium or longer lived storms is counted if the ensemble produces at least one modeled medium-lived storm. Aside from CSI, skill scores are significantly reduced for this category. A high CSI in the face of low TSS and HSS indicates that the ensemble does not correctly identify days when there will not be medium lived storms.

The CSI, HSS and TSS scores are slightly worse for long-lived storms, but improve a bit for supercell storms. Thus, the skill of the ensemble in identifying days on which supercells will occur is slightly better than identifying days when long-lived cells will occur. Except for the general category of Thunder, the ensemble tends to produce too many of each kind of storm compared to the observed frequency, though the high bias for medium and supercell storms is practically negligible.

Overall, the ensemble was favorably received by SPC forecasters. Ideally, the ensemble region should be located within the region that SPC forecasters are most concerned about. But, because the ensemble region is

chosen using the Day 2 outlook, it was often somewhat removed from the ideal location, and sometimes provided no utility to the forecasters because the primary threat region changed from the Day 2 expectations. On a few occasions, the ensemble was run over an area for which there was clearly no threat. On two occasions, forecasters asked that a special run, using the OE/KF mix, be produced for the 0000, 0300, 0600 UTC time frame

6. Conclusions

Based on these results, an ensemble of cloud models initialized with soundings from mesoscale models appears to be particularly skillful in identifying when convection will and will not occur over a small (roughly 160 x 160 km) region. Additionally, the ensemble seems to have some skill at identifying if severe reports will be generated in and around the ensemble region. Scores for two different-sized regions are, except for FAR, examined and seem relatively insensitive to the region size. As might be expected, FAR is considerably lower for the larger-sized region. This may indicate that the ensemble results are valid for an area larger than the region used to generate the initialization, though no upper or lower bound is apparent with only two different choices. This is not in itself surprising, but there is little indication of how the scores degrade as the region shrinks or expands. The relative insensitivity to the different-sized regions may also indicate that the ensemble does not provide much information about the likely location of convection within the region, though no statistics have been compiled to examine this aspect.

Equal to the skill in identifying days for which severe reports will be generated is the skill shown when identifying days on which supercell storms are likely within the region of interest, based on a subjective interpretation of archived radar data. The ensemble is somewhat less skillful in identifying the primary storm lifetime, though still shows positive skill even here.

There were significant, systematic differences in soundings produced by the three mesoscale models employed in this study. The KF and RUC20 produced soundings deemed more representative of the general environment, which were typically capped by a weak inversion. These soundings were less likely to produce deep convection within the model. In contrast, the OE tended to generate soundings that might be more characteristic of conditions where storms occur, which are typically not capped. In the future, each model will be separated from the ensemble whole and scored independently. Doing so will help determine if most of the information extracted by the ensemble comes primarily from one of the mesoscale models. For example, the scores may be insensitive to excluding the KF.

There is some uneasiness about how the cloud model ensemble elements are initialized. Warm bubbles do not represent natural processes very well and work well primarily in uncapped environments. Yet, a strong argument can be made that uncapped environments are rare over the resolvable scales available within current mesoscale models. An equally strong argument can be made that deep convection typically initiates and continues within uncapped environments. How best to use a mesoscale model that maintains a scale-appropriate cap is no t yet clear.

While no statistics have yet been compiled, there was a clear signal within the ensemble regarding the timing of when convection was most likely during the day. Seldom did the ensemble generate storms from any 1800 UTC sounding. More often, the ensemble produced deep convection from the 2100 and 0000 UTC soundings. How useful the ensemble might be in providing reliable timing information is a topic for future work.

Mixing the soundings, as was done here, makes it difficult to remove initial conditions from a particular model if that model is deemed to contain significant errors. Having this capability would allow forecasters to more easily examine the contribution of each model in the final ensemble result and, when appropriate, results from initial conditions provided by a particular mesoscale model could legitimately be excluded from the ensemble. It also would allow forecasters to examine systematic differences between different mesoscale models.

Acknowledgements.

The authors wish to thank Dr. L. J. WIcker for generously allowing the use of his cloud model in this work. We also thank the SPC forecasters for their feedback and the SPC support staff for access to archived radar data and support for its display. Support for this work was provided by the National Severe Storms Laboratory.

References.

Bleck, R. and S. G Benjamin, 1990: Regional weather prediction with a model combining terrain-following and isentropic coordinates. Part I: Model description. *Mon. Wea. Rev.,* **121,** 1770-1785.

Efron, B. and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap.* Chapman and Hall, New York, NY, 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.

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

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

Kessler, E., 1969: On the distribution and continuity of water substance in atmospheric circulation. *Meteor. Monogr.,* **10,** no. 32, 83 pp, Amer. Meteor. Soc.

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

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.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction.* Academic Press, New York, NY, 467 pp.