Final Report

Verification of WRF ensembles using object-oriented approaches

William A. Gallus, Jr

August 31, 2008

1. Introduction

The goal of this project was to use object-oriented (also known as entity-based) verification approaches on WRF ensemble precipitation forecasts. Two different sets of ensemble precipitation forecasts were evaluated with object-oriented verification approaches. The first set of ensembles included an 8 member 15 km ensemble that used unperturbed initial lateral boundary conditions with mixed physics and dynamic cores (Mix), and another 8 member 15 km ensemble using the same dynamic core and physics but perturbed initial and lateral boundary conditions (IC/LBC). Both of these WRF ensembles were run on a 64 node linux cluster at Iowa State University. These two ensembles were compared using traditional verification metrics in a paper by Clark et al. (2008). Thus, this set was chosen for the DTC project to determine if the behavior of object parameters as determined by both MODE (Davis et al. 2006a, b) and EMT (Ebert and McBride 2000) was comparable to that of traditional measures. Analysis with EMT was performed on a linux box at Iowa State University. MODE analysis was performed on a WRF-DTC machine at NCAR. Clark et al. found in the comparison of these two ensembles that the spread and skill of the two ensembles was initially comparable but that over time, the lack of perturbed lateral boundary conditions reduced the growth of spread in the Mix ensemble so that better spread and skill were found at later times in the IC/LBC ensemble.

The second set of ensembles evaluated included a 5 member 4 km WRF ensemble run by CAPS for the 2007 NOAA Hazardous Weather Testbed Spring Program and 5 members of a 15 member 20 km WRF ensemble run for the same cases on the 64 node linux cluster at ISU. These two ensembles have been compared in a paper being submitted to Monthly Weather Review (hereafter referred to as Clark et al. 2009), and thus were chosen for use in this DTC project to again determine if the behavior of object parameters was similar to that of traditional metrics for these cases. For these two ensembles, MODE was used alone (EMT was not run), and run on the DTC machine at NCAR. Clark et al. found in this comparison that the explicitly-resolved convection in the 4km ensemble led to a much better representation of the diurnal cycle, and thus the 4km ensemble performed better than the 20 km ensemble, even when the 4km ensemble's 5 members were compared to the full 15 members of the 20 km ensemble.

The project included two separate visits to the WRF-DTC. The first occurred in August 2007. During that time, analysis focused on the 8 member ensembles, and both EMT and MODE were run (this was the first time Gallus had run MODE). The second DTC visit occurred in August 2008. Because the mode_analysis tool was developed between these two visits, the tool was used on the 8 member ensembles' output at this time. In addition, during this most recent visit, an updated version of MODE was run on the 8 member ensembles, replacing output obtained during 2007, and this updated MODE also was applied to the 4 and 20 km ensembles, with mode_analysis used to evaluate all of the output from both of the 8 member ensembles and the 4 and 20 km ensembles. MODE and the analysis tool were also applied to probability forecasts from the 8 member ensembles, and to probability-matched ensemble mean forecasts from the 4 and 20 km ensembles.

In the report to follow, results for the comparison of the two 8 member ensembles are presented first, with the comparison of the 4 and 20 km ensembles following.

II. Overview of the two 15 km 8-member ensembles

For the two 8 member 15 km WRF ensembles, a total of 72 cases was available, and verification was performed for each case on the 10 6-h precipitation accumulations present during the first 60 hours of the forecasts. With 16 total WRF configurations for all periods and cases, 11,520 separate rainfall forecasts were evaluated using both EMT and MODE. A rainfall threshold of 6.25 mm (.25 inch) was used to define the objects in both techniques. All of the ensemble precipitation output had been remapped to a 10 km grid. Some limited sensitivity tests were done with MODE to determine configuration file parameters that would yield results reasonably similar to those obtained from EMT.

The ensemble precipitation data had been evaluated using traditional ensemble verification measures in Clark et al. (2008). As can be seen in Fig.1, Clark found for all rainfall thresholds evaluated that the skill as determined by the area under ROC curves was initially greater in the Mix ensemble, but after 30-36 hours, skill was greater in the IC/LBC ensemble. The better skill at later times appears to reflect a problem with lack of

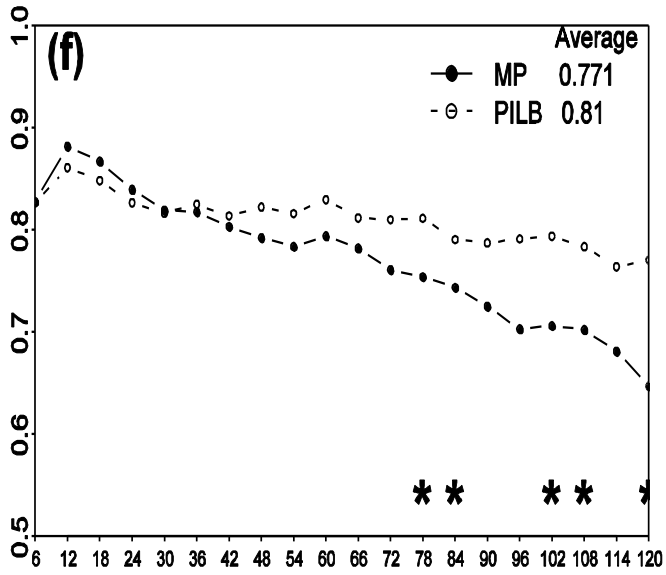

Figure 1: Area under the ROC curve as a function of time for the Mix (MP) and IC/LBC (PILB) ensembles (from Clark et al. 2008)

increasing spread in the Mix ensemble (Fig. 2). It can be seen that variance slowly grows with time throughout the 120 hours of integration examined in Clark et al. in the IC/LBC ensemble, but generally levels off after hour 30 in the Mix ensemble. Also of note is the diurnal cycle which results in peaks in both the mean square error of the ensemble means, and the variance at hours 6, 30, 54, etc, which are times when the rain area and rain

volume were maximized in the domain (not shown; see Clark et al. 2008). Rain areas tended to fluctuate by +/- 10% from the average during the day, while volumes fluctuated by +/- 30%. These results imply that rain rate also exhibited local maxima at these same times. Spread ratios (Fig. 3) show the difference between the 2 ensembles even more dramatically, with those for the IC/LBC ensemble growing rapidly at all times, while those for the Mix ensemble remain relatively constant. Thus, some differences exist in the results obtained using different traditional verification measures.

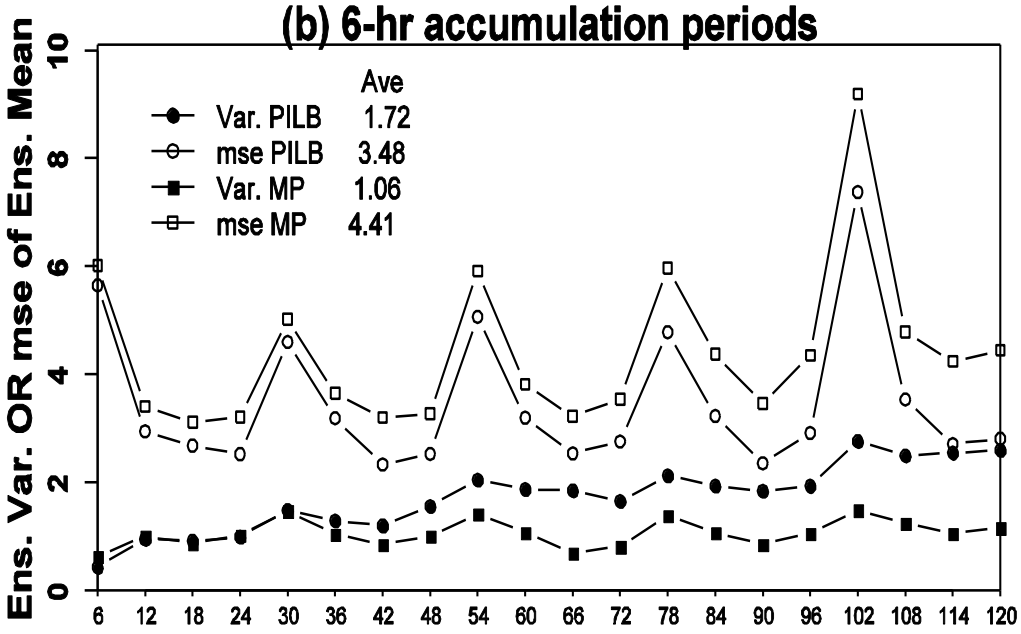

Figure 2: Ensemble variance (filled) and mean square error of the ensemble mean (open) for the two ensembles. Squares indicate the Mix ensemble, circles the IC/LBC one. From Clark et al. (2008)

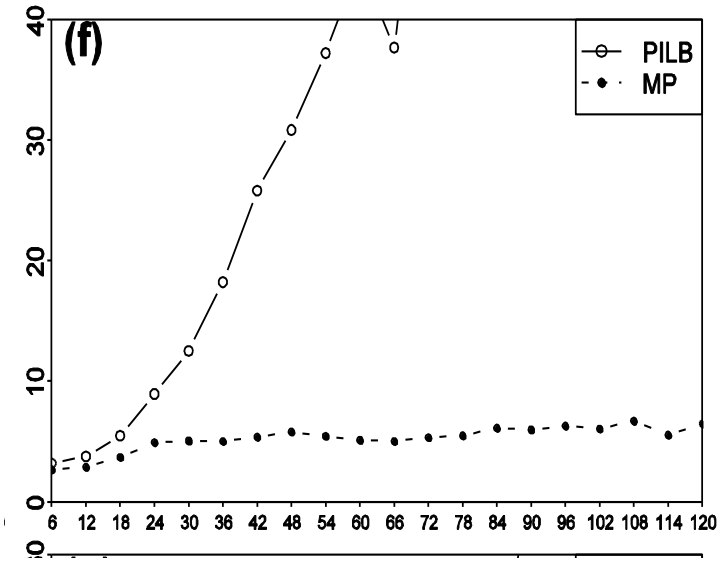

Figure 3: Spread Ratio for the IC/LBC ensemble (PILB) and the Mix ensemble (MP) as a function of time (from Clark et al. 2008)

In summary, the Clark et al. (2008) work shows that the lack of perturbed lateral boundary conditions results in major changes over time in the spread and skill behavior of the two ensembles. These changes become evident within the first 60 hours of the forecast, and thus the object-oriented verification approaches were applied on this portion of the Clark et al. data alone.

3. Comparison of Object-oriented verification parameters with Clark et al. (2008)

Using the first 60 hours of the forecasts from Clark et al. (2008), a comparison of standard deviations was performed between the results of the EMT and those of MODE for several parameters computed by the verification techniques. Standard deviations were used as a measure of spread, and data were plotted as a function of forecast hour from the first 6 hours through the 54-60 h forecast. The standard deviations were computed for parameters valid for individual objects. For both EMT and MODE, a system had to be depicted in at least 4 of the 8 ensemble members to be included in the analysis. EMT output files are automatically separated into different files for different CRAs (contiguous rain areas), making this analysis relatively easy. A check was performed to be sure the centroids of CRAs were within a few degrees latitude and longitude of each other to be sure the results were valid for the "same" system. For MODE output, such an analysis was more difficult since all of the object parameters end up in the same file. The mode_analysis tool was used to compute the standard deviations, and was run for those cases with multiple composite objects two or three times using latitude and longitude bounding boxes to try to restrict the standard deviation computations to the same system in each ensemble member. Information on composite objects was also taken into account, but since MODE does not necessarily assign the same numbering convention to the same systems in different ensemble members, latitude and longitude information for the centroids of objects was more useful. In the results to follow, mode_analysis was applied for simple objects where observed and forecasted objects were matched.

In Fig. 4, it can be seen that smaller standard deviations for rain rate were present in the MODE results, with the mixed ensemble starting with higher deviations than the IC/LBC one. Over time, the deviations are generally greater than in the first 6 hours, but any trends are very subtle. In MODE, again the mixed ensemble starts with higher SDs. A diurnal cycle is more evident in MODE than in EMT. Although not shown, it should be pointed out that observed rain rates evidenced a diurnal cycle with maxima in the 00-06, 24-30, 30-36, 48-54 and 54-60 hour periods, and minima during hours 06-12, 12-18, 36-42, and 42-48. The MODE standard deviations follow this same trend but the EMT results do not. The model rain rates miss the first diurnal peak, possibly evidence of spin up problems during the 00-12 h period (not shown). Both ensembles do have a maximum in rain rate in the 24-30 h period, and again in the 48-54 h period, but the amplitude of the signal is less than observed (peak variation in MIX of .79, and IC/LBC of .71 compared to observed variations of .81 and 1.24). Both EMT and MODE imply a very slow tendency for greater spread at increasing forecast times, and MODE shows a fairly

smooth trend for the rate of increase in spread to be greater in the IC/LBC ensemble than in the MIX one. It should be noted that the average rain rate determined using MODE (EMT) for both ensembles was around .4 in (.5 in).

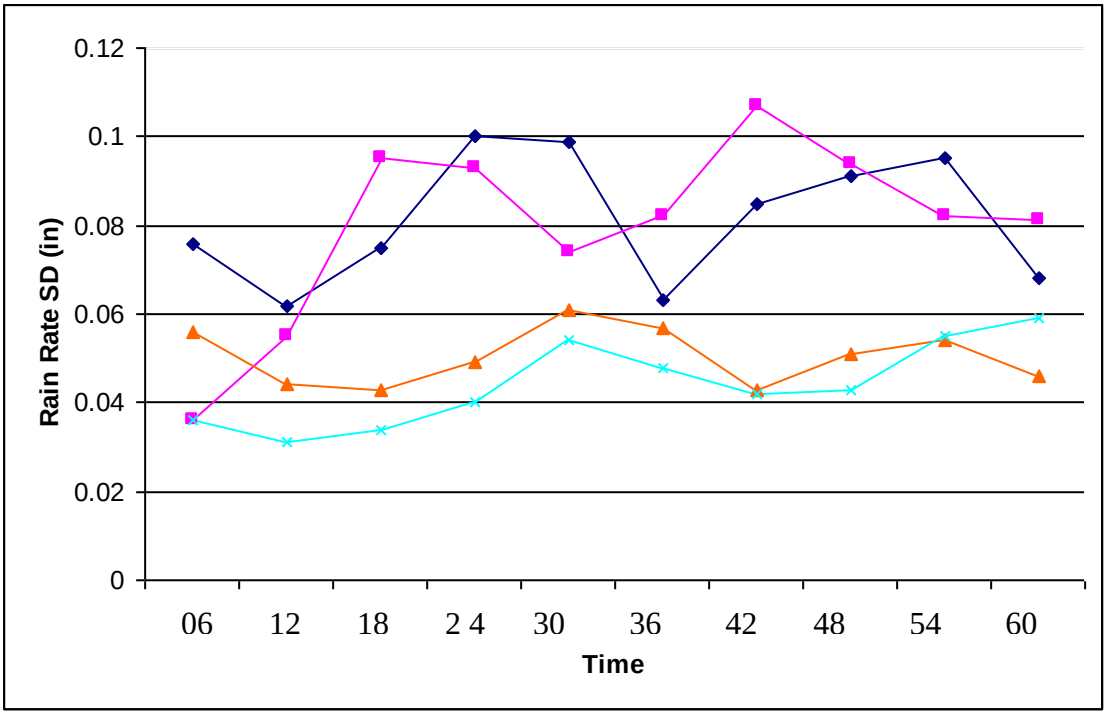

Figure 4: Standard deviation of 6-hr rainfall (in) among the 8 ensemble members of the mixed physics/dynamics (Mix) and perturbed initial/lateral boundary condition ensembles (IC/LBC) as evaluated via the EMT (blue and pink curves) and MODE (orange and cyan curves).

Standard deviations for rain volume are shown in Fig. 5. Systematic temporal trends are hard to discern in the EMT output, but the MODE output more clearly shows both impacts from the diurnal cycle (observed peak volume occurs at hours 00-06, 24-30, and 48-54) and a slow growth in deviations over time. In addition, both MODE and EMT show greater growth in the deviations for the IC/LBC ensemble than for the MIX one, although the EMT curves have more noise and do not show this trend as clearly. The mean at all times for all ensembles was roughly 1.0. The observed rain volume (not shown) evidences a large amplitude diurnal cycle with the peak volumes during the first part of the evening roughly twice those (.75 km^3 greater) of the minima (occurring in the 18-24 and 42-48 h periods corresponding to 18-00 UTC). As with rain rate, the maximum early in the forecasts is not well-depicted, likely due to spin-up problems, with MODE showing both ensembles having a peak in the 6-12 h period, a time where the observed volume has decreased by over .4 km^3 from the previous 6 hour period. Both ensembles do have a maximum in the 24-30 h forecast period, with a minimum in the 18-24 and 42-48 h periods. However, the difference in volume between the minima and maxima is only around .3 km^3 for the MIX ensemble and .14 km^3 for IC/LBC.

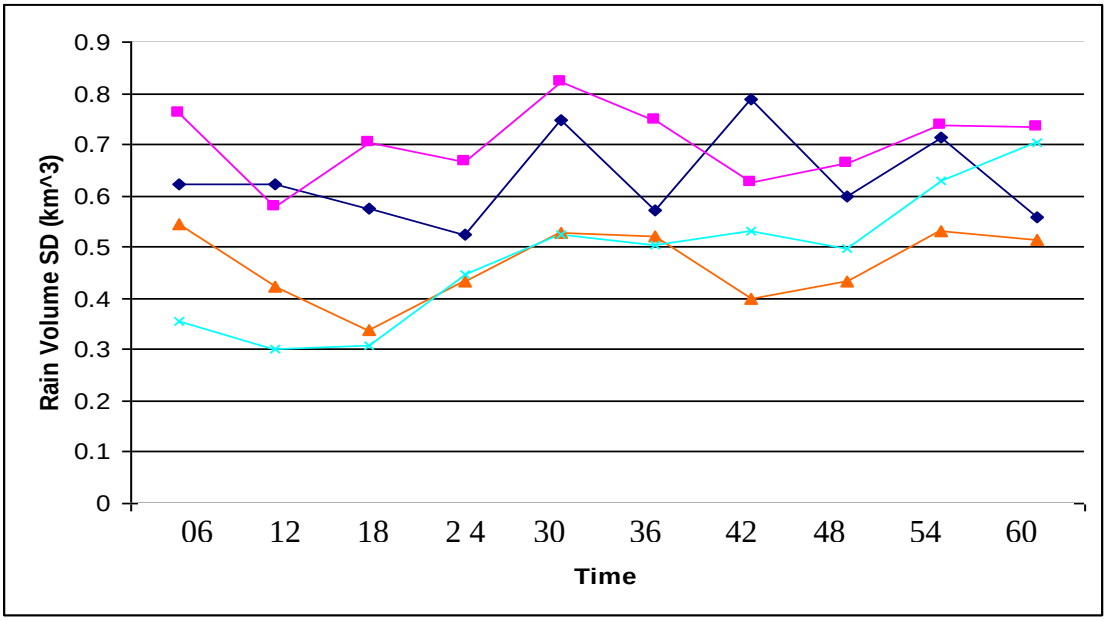

Figure 5: As in Figure 4 except for rain volume. Units for both are in km^3.

Standard deviation of areal coverage of rainfall is shown in Fig. 6. The MODE results differ noticeably from the EMT results in the first 18 hours, with the EMT showing generally increasing standard deviations with time for both ensembles while the MODE results depict a decrease. After the 12-18 h period, though, both techniques show a general increase with time, with the bigger growth in deviation happening in the IC/LBC ensemble. Keep in mind that there are differences in how the forecasted systems were defined. For the EMT, the forecasted rain areas had to be part of contiguous rainfall areas determined by the technique. For MODE, simple forecasted objects that were matched with observed ones were used in the analysis. For both techniques at least half of the 8 ensemble members had to show the system for it to be included in the analysis. It should be noted that the observed areas strongly showed a diurnal cycle with maxima/minima at roughly the same times as the peaks in rain rate. Maximum areal coverage was 2-3 times that of the minimum coverage. Similar to rain rate, and even moreso rain volume, the amplitude of the cycle in the forecasts was greatly damped, especially in the IC/LBC ensemble.

Figure 7 shows standard deviations of displacements (km) from the EMT and MODE. There may be a hint of increasing values with time in the EMT and especially MODE results for the IC/LBC ensemble that is not present in the Mix ensemble. This behavior would be consistent with the findings of Clark et al. (2008). MODE results indicate that displacement errors are a little larger for the IC/LBC ensemble at most times compared to the Mix ensemble (not shown). Typical displacement errors are on the order of 200 km. A diurnal cycle is difficult to see in the displacement error standard deviations. Also of note in the figure are the much larger values for the MODE results compared to EMT. It is possible that these larger values are related to the fact that systems are only matched in EMT when they are contiguous, whereas in MODE, systems are matched if the interest

parameter is greater than some threshold. Thus, some systems are matched in MODE even though they are not contiguous and may be separated by some distance. These differences in the way the schemes operate should lead to larger average displacement errors and standard deviations in MODE than in EMT.

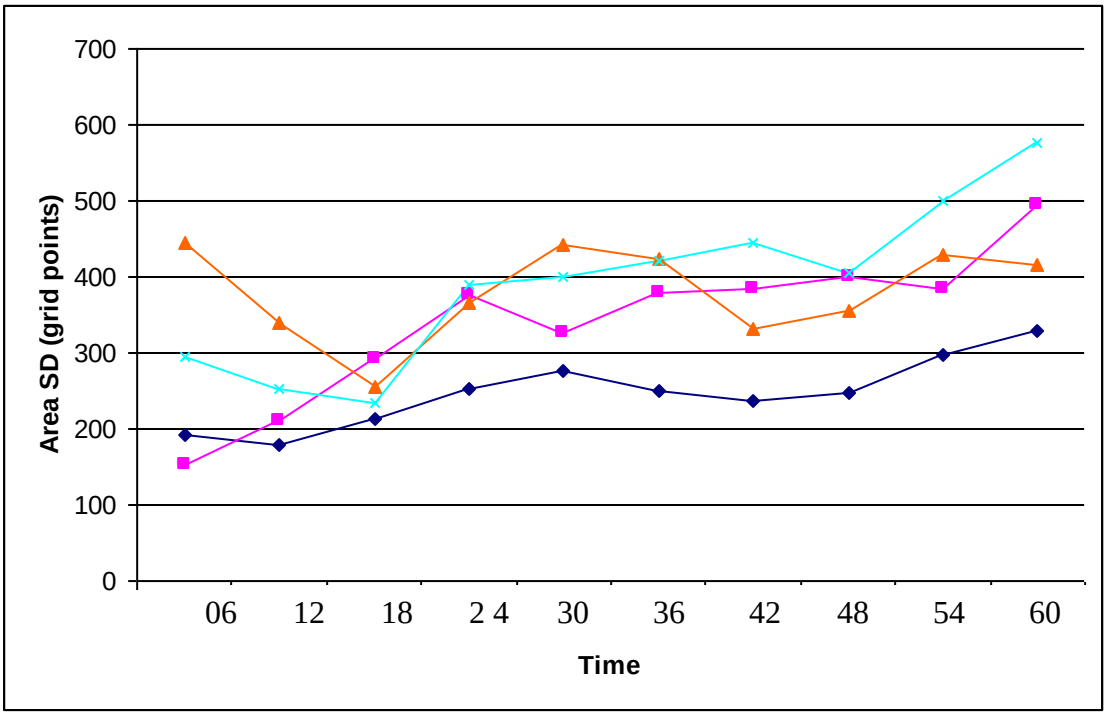

Figure 6: As in Fig. 4 except for areal coverage (number of grid points) of rainfall.

The standard deviation in number of cases from each ensemble and both verification systems is shown in Fig. 8. Both EMT and MODE show similar results, with the Mix ensemble having far larger values around hours 24 and 48, times just before the diurnal maximum of precipitation. Standard deviations for the IC/LBC ensemble are much smaller and more uniform with both verification systems. The MODE results for this ensemble show a tendency for a gradual increase to occur with time, potentially matching the Clark et al. (2008) results.

In summary, the trends toward increasing spread with time in the IC/LBC ensemble with limited changes in the Mix ensemble seen in Clark et al (2008) can be seen in the object parameters computed in both MODE and EMT, but they are not as obvious as in traditional ensemble measures. The diurnal cycle with convective precipitation being more abundant during the overnight hours influences the object parameters. Standard deviations for areal coverage of objects evidence the greatest increase over time in both the MODE and EMT results. Overall, the MODE results seem to show the increased rate of spread growth in IC/LBC versus Mix more clearly than the EMT results.

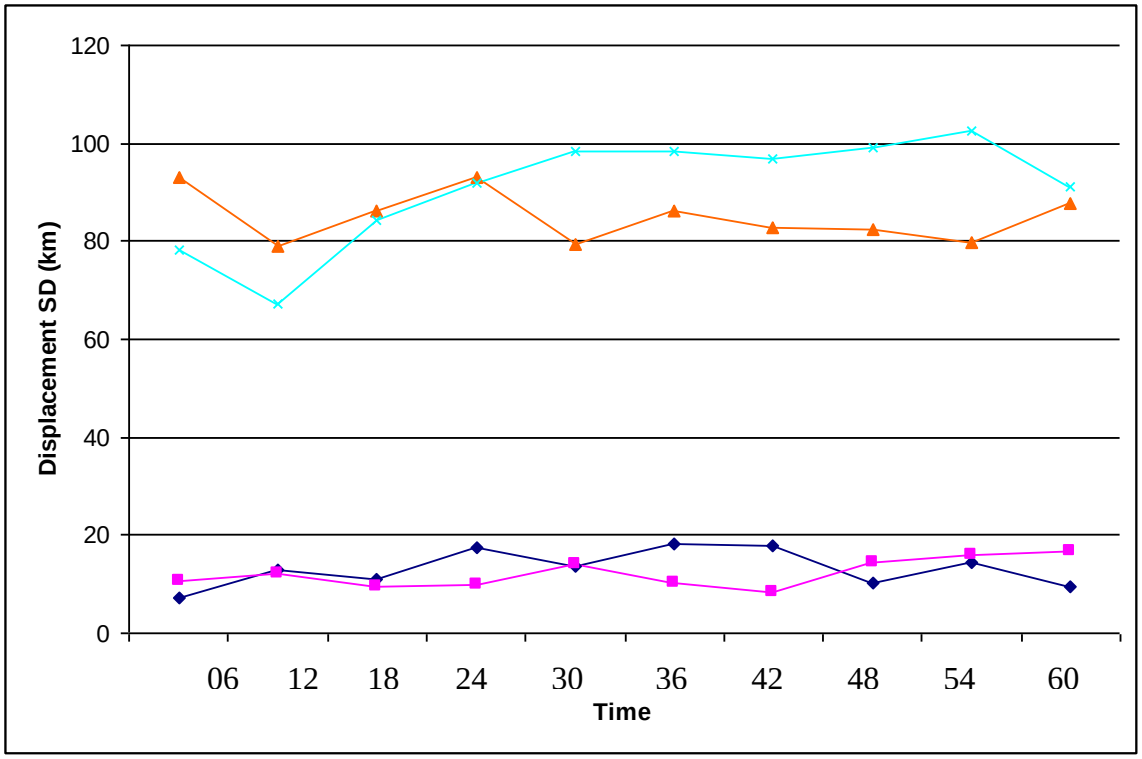

Figure 7: Standard deviations of displacement error (km) for the Mix (blue –EMT, orange - MODE) and IC/LBC ensembles (pink –EMT, cyan - MODE) as a function of time.

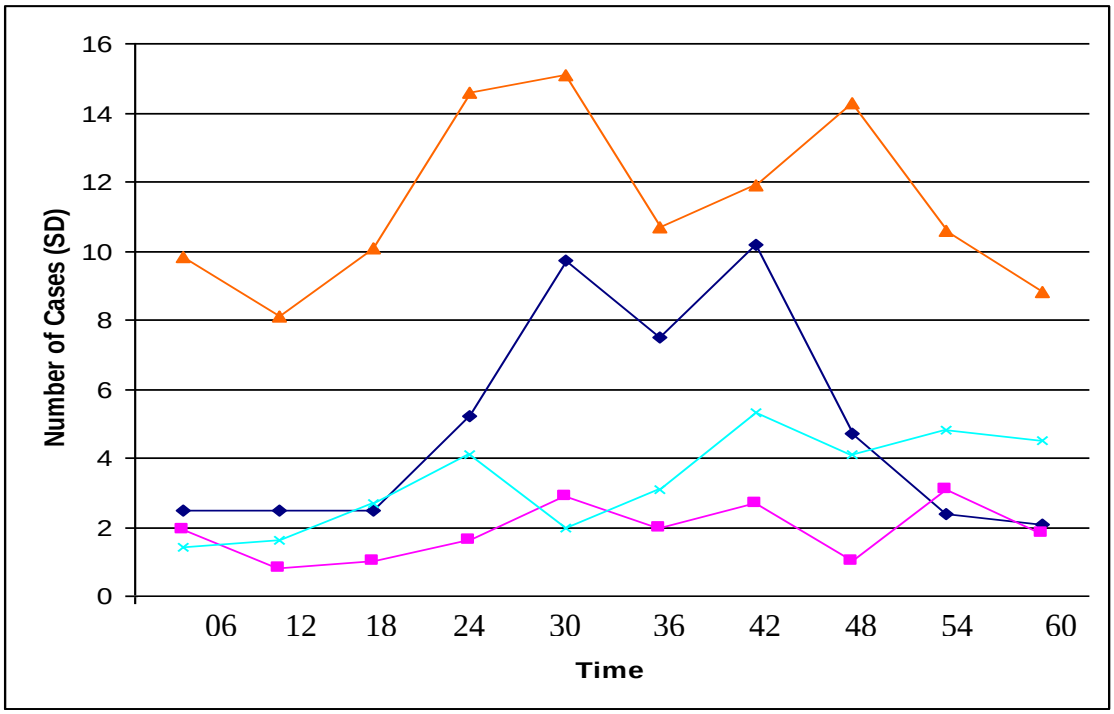

Figure 8: As in Fig. 7 except for number of cases identified in each ensemble.

4. Forecasting based on ensemble object-oriented output

In addition to the comparison with Clark et al. (2008) results, the 8 member ensemble output was used in EMT and MODE to explore ways that the output could be used to assist forecasters.

A. Comparison of Probability-Matched Ensemble Mean object parameters with averages of object parameters from all ensemble members using EMT

In one test, observed values of parameters were compared with both the means of the distribution of object parameters from each of the 8 ensemble members in both ensembles and output from the EMT applied to the deterministic forecast created using the probability matching (PM) technique. In other words, could a better forecast be obtained by applying object-oriented approaches to individual ensemble members and then using the mean value for object parameters as the forecast instead of using a more traditional ensemble mean forecast?

Averages for all 72 cases by forecast hour for rain rate (Fig. 9) showed that objects in the probability-matched forecast tended to have slightly lower rain rates at all hours than the average rain rate from objects identified in each of the 8 ensemble members of the Mix ensemble. However, since the observed value was sometimes greater than both of these forecasts and sometimes smaller, the better forecast was split evenly between these two approaches. For the IC/LBC ensemble, the objects from the probability-matching had smaller rain rates only up to hour 36, wth the exception of the 18-24 h period. From hours 36-60, the PM approach had larger values. Unlike the Mix ensemble, at almost all times, both forecasts had too small rain rates compared to observations. For the IC/LBC ensemble, the PM approach had the same or smaller errors than the averaging of the EMT results from all 8 members about 70% of the time. Differences were usually small, however. An examination of individual cases from a subset of 20 of the 720 total 6-h periods suggests the PM results are better a much larger fraction of the time. Thus, from the EMT results alone, it appears that if a forecaster wants guidance on an average rain rate, the PM forecast will do a better job than applying the EMT code to each member and then determining an average from all of the ensemble members. It can also be seen from Fig. 9 that the spread in rates in the Mix ensemble is larger than that of the IC/LBC ensemble, but it does not appear to grow with time. There may be some hint of this growth in the IC/LBC ensemble. A high bias exists in the Mix ensemble at most times, but seems to trend lower toward hour 60. At hours 12 and 18, all members had too high of rates compared to observations. The IC/LBC ensemble has a low bias at most times, with all members too light at hours 00-06, and from 18-60.

Similar analysis using the averages of all 72 cases has been performed for rain volume (Fig. 10) but not other parameters. Rain volume errors show a diurnal trend with a tendency for overestimates prior to the convective maximum and underestimates around the time of the maximum. As with rain rate, no tendency for growth in spread with time

appears, and the PM forecast appears comparable in skill to one based on an average of the parameter values from individual ensemble members. In fact, as with rain rate, the forecast from the average of the parameters and that from the PM approach was very similar. Thus, at least for these two parameters, there would be little reason to average the output from object-oriented approaches run on every ensemble member. However, there may still be value in using the object-oriented approaches on each member to create distributions of possible scenarios (e.g., probabilities of rates or volumes exceeding thresholds). A separate comparison of 20 individual times out of the 720 total continued to find that a PM forecast did a better job of predicting rain volume and areal coverage than a forecast made using the mean of the EMT results from the 8 ensemble members in either ensemble.

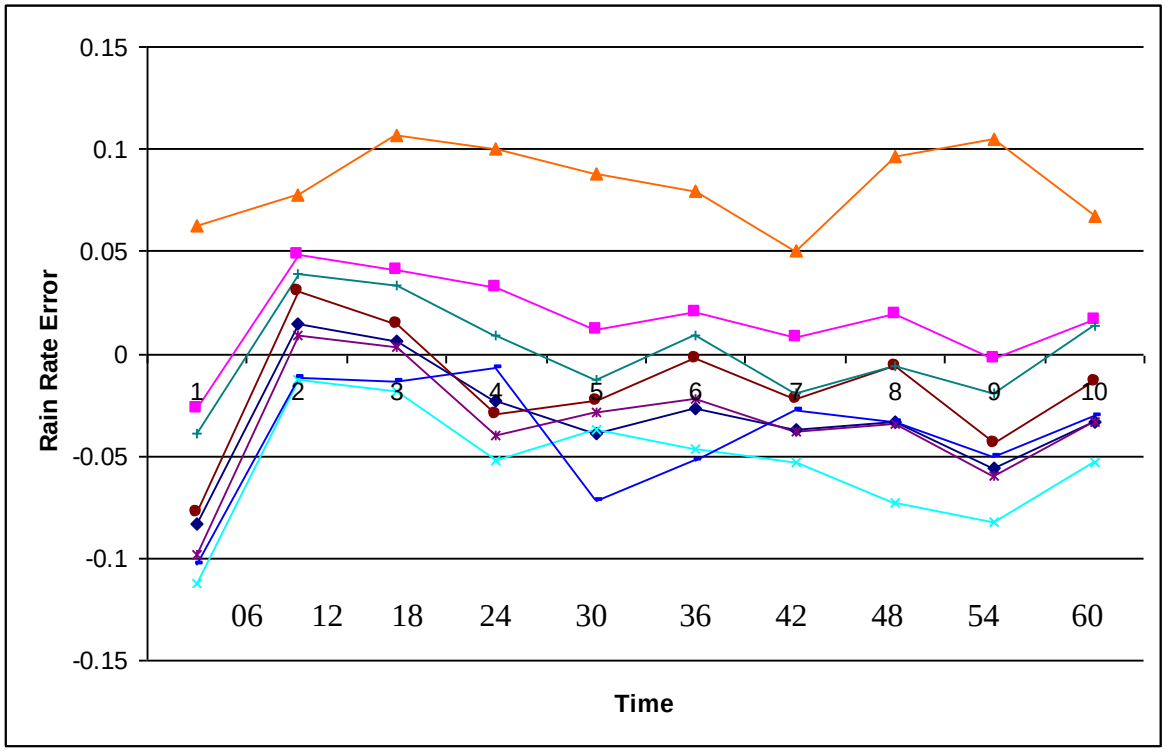

Figure 9: Difference between the forecasted rain rate and the observed as a function of time. Curves are plotted to show the maximum value by any ensemble member (orange-Mix, brown-IC/LBC), the mean value (pink-Mix, purple-IC/LBC) and the minimum value (dark blue –Mix, light blue-IC/LBC). The PM forecast is shown in green (Mix) and medium blue (IC/LBC).

In another test of how the object-oriented output might be used to help forecasters, the percentage of time that the observed rate, volume, and areal coverage fell within the spread of the 8 ensemble members was examined during all 60 hours. Figure 11 shows that this approach worked much better for predicting rain rate and volume than for areal coverage. In addition, for all three parameters, the IC/LBC ensemble seemed to perform better as the forecast time became greater.

A final test was performed to see if spread-skill relationships often examined for ensemble forecasts could be applied to object-oriented ensemble guidance. Fig. 12 shows the time evolution of mean absolute error for volume, rate, and areal coverage for those systems that had large standard deviations among the members and those that had small values. Large and small were defined to be greater than 150% of the average standard

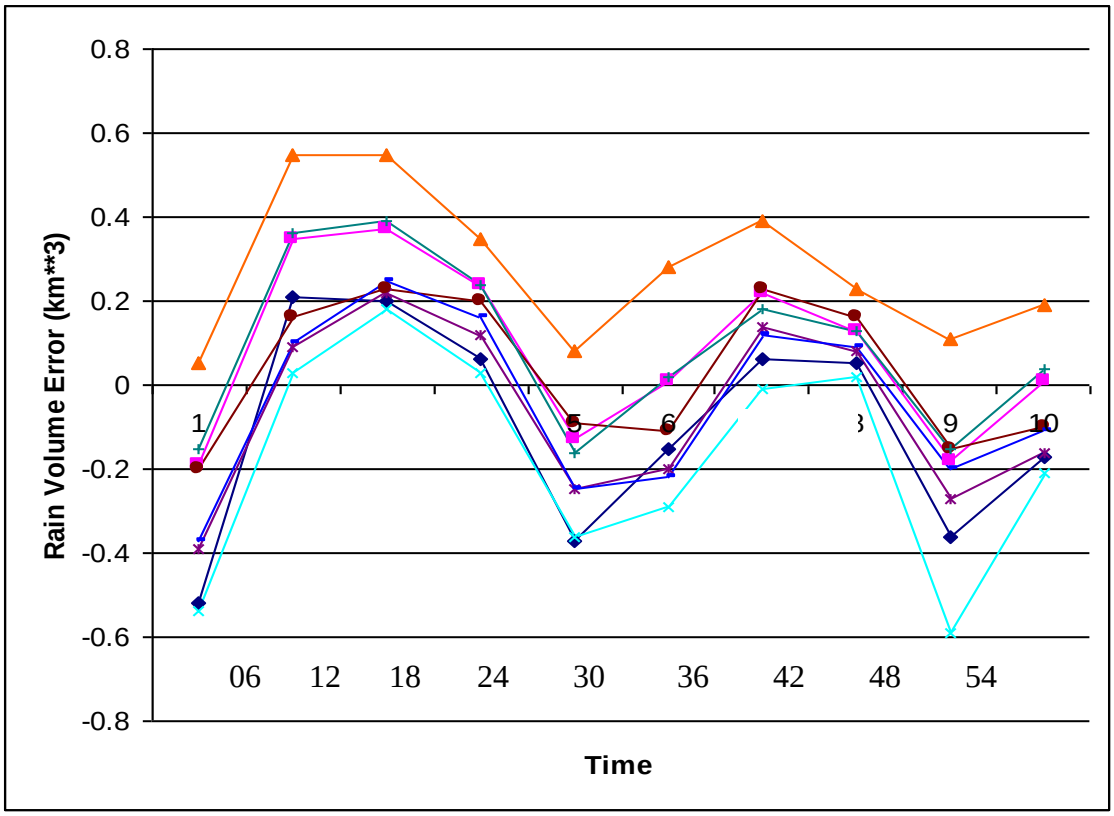

Figure 10: As in Fig. 9 except for rain volume error in km^3.

deviation or less than 50% of the average. Rain volume and areal coverage showed a clear separation with much more accurate forecasts in the events where spread was relatively small. Rain rate did not show the relationship as definitively, although for a majority of the time, it still applied. However, at hours 00-06, 18-24, and then 48-60, the forecast skill did not vary much with standard deviation, or the more accurate forecast had the bigger standard deviation. It is not entirely clear why rain rate would behave differently. These results in general imply that forecasters may be able to establish a confidence level for their forecasts of some object parameters using the standard deviations.

In another test of how the object-oriented output might be used to help forecasters, the percentage of time that the observed rate, volume, and areal coverage fell within the spread of the 8 ensemble members was examined during all 60 hours. Figure 11 shows that this approach worked much better for predicting rain rate and volume than for areal coverage. In addition, for all three parameters, the IC/LBC ensemble seemed to perform better as the forecast time became greater.

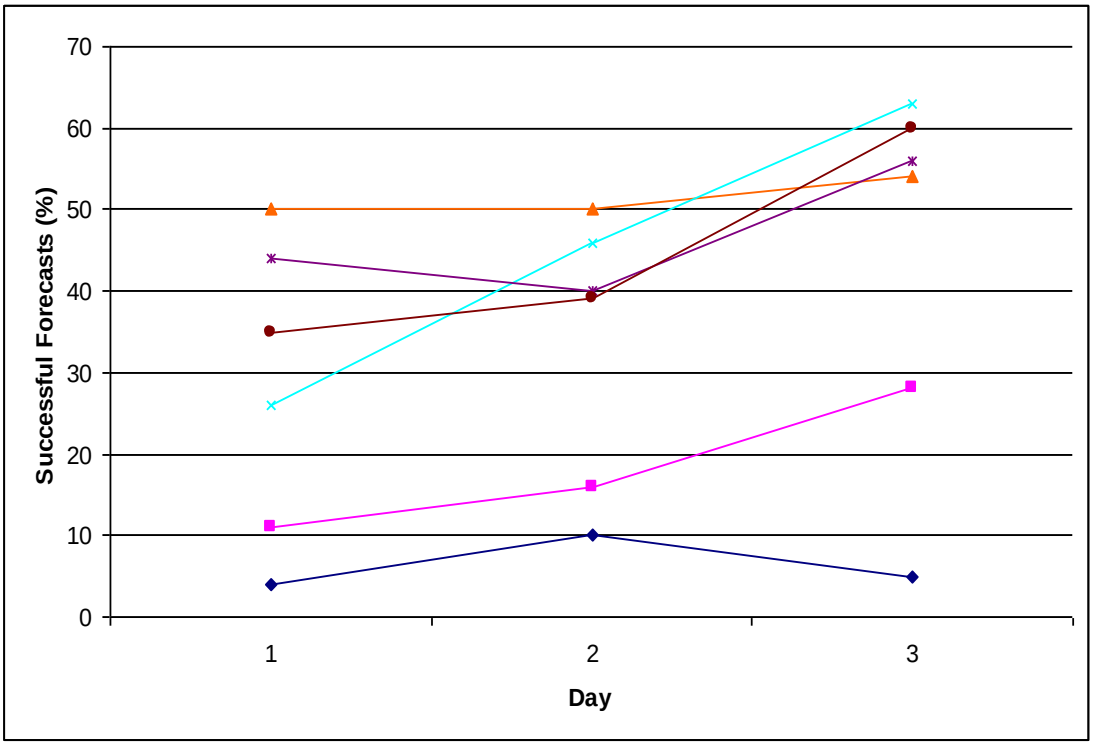

Figure 11: Percentage of time that observed values of volume (mix-purple, IC/LBC-brown), rate (mix-orange, IC/LBC-cyan), and areal coverage (mix-dark blue, IC/LBC-pink) fell within the minimum and maximum predicted by the Mix ensemble, as determined using the EMT. Day 1 refers to 6h forecast periods in the first 18 hours of the forecast, day 2 refers to those in the 18-42 h period, and day 3 those in the 42-60 h period.

A final test was performed to see if spread-skill relationships often examined for ensemble forecasts could be applied to object-oriented ensemble guidance. Fig. 12 shows the time evolution of mean absolute error for volume, rate, and areal coverage for those systems that had large standard deviations among the members and those that had small values. Large and small were defined to be greater than 150% of the average standard deviation or less than 50% of the average. Rain volume and areal coverage showed a clear separation with much more accurate forecasts in the events where spread was relatively small. Rain rate did not show the relationship as definitively, although for a majority of the time, it still applied. However, at hours 00-06, 18-24, and then 48-60, the forecast skill did not vary much with standard deviation, or the more accurate forecast had the bigger standard deviation. It is not entirely clear why rain rate would behave differently. These results in general imply that forecasters may be able to establish a

confidence level for their forecasts of some object parameters using the standard deviations.

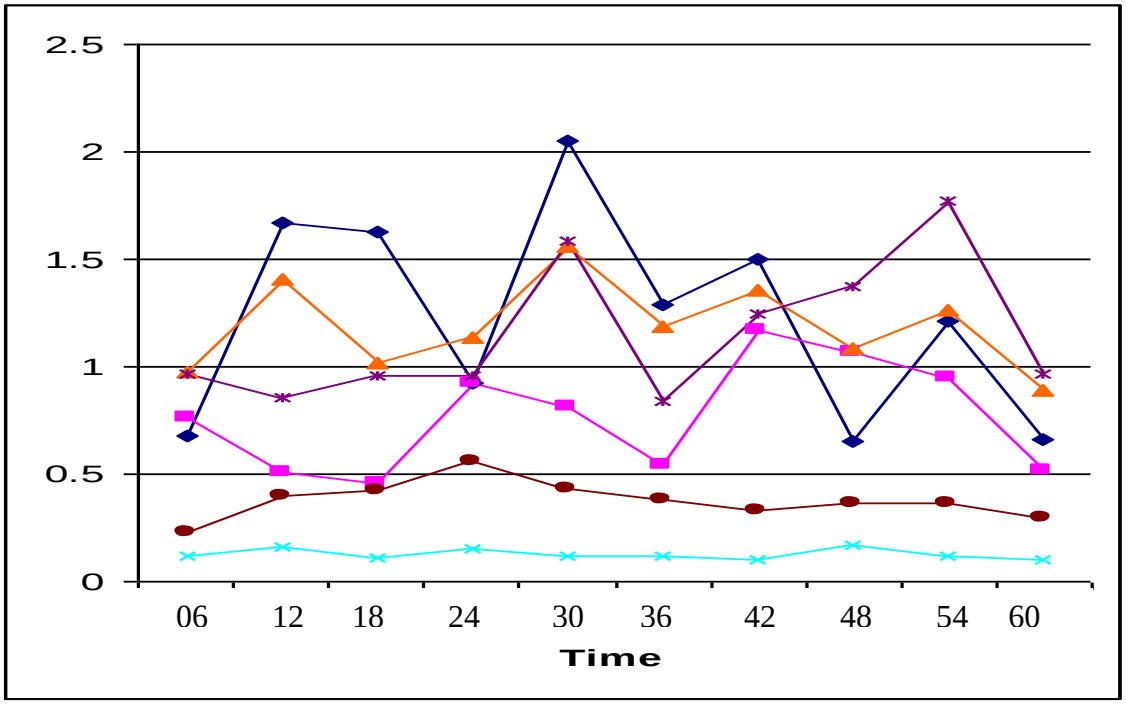

Figure 12: Mean absolute error for volume (km^3), rate (x 0.1 in) and areal coverage (*1000 points) of precipitation as a function of time for cases with standard deviations over 150% of the mean (rate-dark blue, volume-orange, area-purple), and less than 50% of the mean (rate-pink, volume-brown, area-cyan).

B. Comparison of results using MODE applied to ensemble probability forecasts and MODE applied to ensemble member forecasted precipitation amounts

MODE allows for a user to compare objects generated from two different fields. In a second test, object parameters were obtained from both the MIX and IC/LBC ensembles using probability forecasts created using equal weighting of the 8 members (thus POP values could be 0, 12.5%, 25%, 37.5%, etc). A probability threshold of 30% to define an object was tested in detail, meaning that at least 3 of the 8 members had to show precipitation above 0.25 inches in 6 hours to be considered a system. A sensitivity test was also performed raising this threshold to 50%, which would mean that at least 4 members had to show an object.

Displacement errors using MODE applied to the probability forecasts are compared to the errors when MODE was applied to QPF amount and the results for all individual members were averaged in Fig. 13. It can be seen that in general displacement errors grow with time. Also, the results using a 30% probability threshold do appear to be

generally a bit better than those based on QPF amount (orange versus dark blue for MIX; cyan versus pink for IC/LBC). In the sensitivity test using a 50% threshold (not shown), displacement errors in the first 30 hours average around 20 km worse for both ensembles, making them slightly worse than for the technique using QPF amount. After hour 30, results are mixed with the 50% threshold performing about the same as the 30% threshold, and actually being the best during the 54-60 h period. Future work should explore how the results change when different probability thresholds are used.

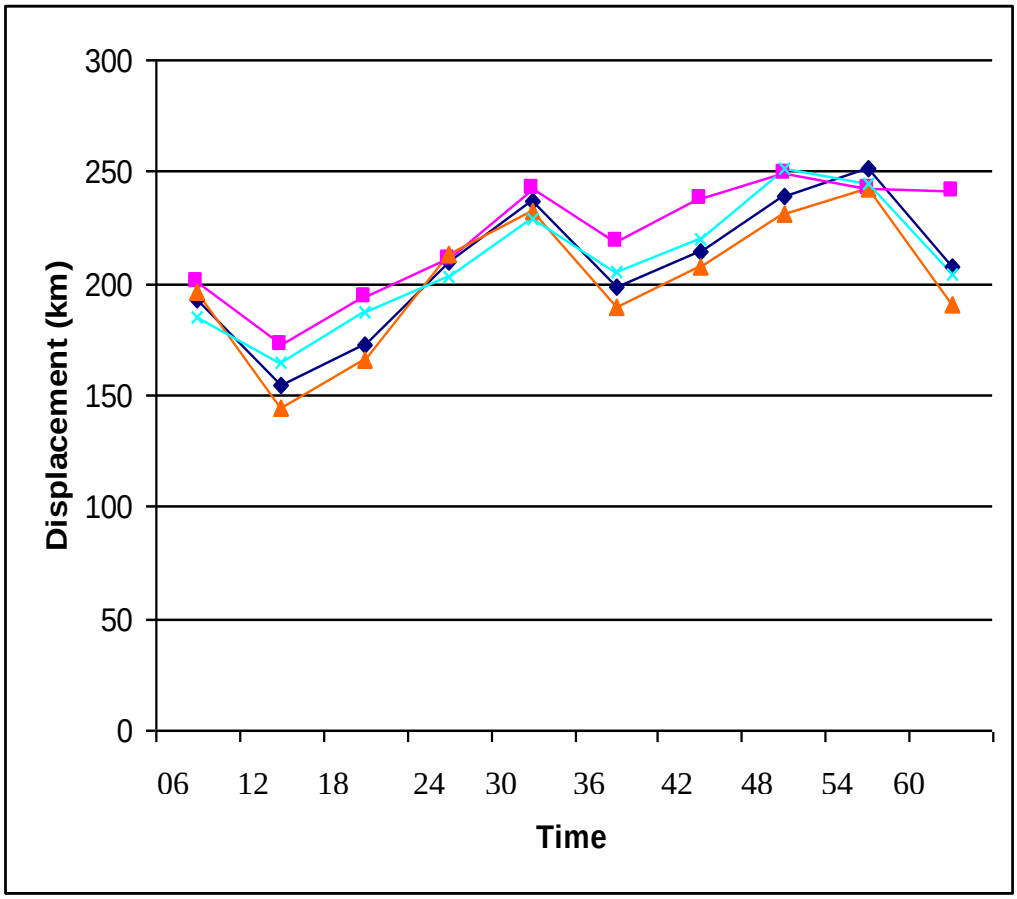

Figure 13: Comparison of displacement errors (km) by time between the 2 ensembles using the average of the MODE output applied to QPF amount for all individual members (dark blue for MIX, pink for IC/LBC) and using MODE applied to probability forecasts with a 30% threshold (orange for MIX, cyan for IC/LBC).

Average rain areas for the two ensembles using both techniques with MODE are shown in Fig. 14, along with a curve representing the rain areas of the observed systems. It should be noted that since not all of the same observed areas get matched with forecasted systems, the observed area differs between the two ensembles. The purple curve plotted in Fig. 14 is an average of the MODE results for objects found using the MIX and IC/LBC ensembles. As mentioned earlier, none of the forecasts has as strong a diurnal signal as the observations. Obviously the threshold used for the probability forecast will greatly affect the rain areas depicted in the forecasted objects. Using 30% probabilities yields better results than those obtained applying MODE to QPF amounts around the

times of the diurnal maximum (note orange versus dark blue curves and cyan versus pink curves at 00-06, 24-30, 48-60 h), but this is likely primarily a result of the probability approach always yielding a much larger rain area (by 200 – 500 grid boxes at most times). The amplitude of the diurnal cycle is also a little stronger when the probability approach is used. When the probability threshold was increased to 50%, areas decreased dramatically, as would be expected (not shown), and the curves looked more like those from the technique based on QPF amount (pink and dark blue curves). The diurnal signal was almost completely removed by changing the probability threshold in this manner. It should be noted that since more agreement is needed to have a 50% POP forecast, one might expect the numbers of systems identified to decrease as a higher POP threshold was used. However, a lower threshold could result in what had been several objects combining into one larger object. An examination of the numbers of objects found by MODE with the two thresholds did show that at most times, substantially more objects were present when the 30% threshold was used (usually 10-40% more), although for at least one time period, the two numbers were nearly identical.

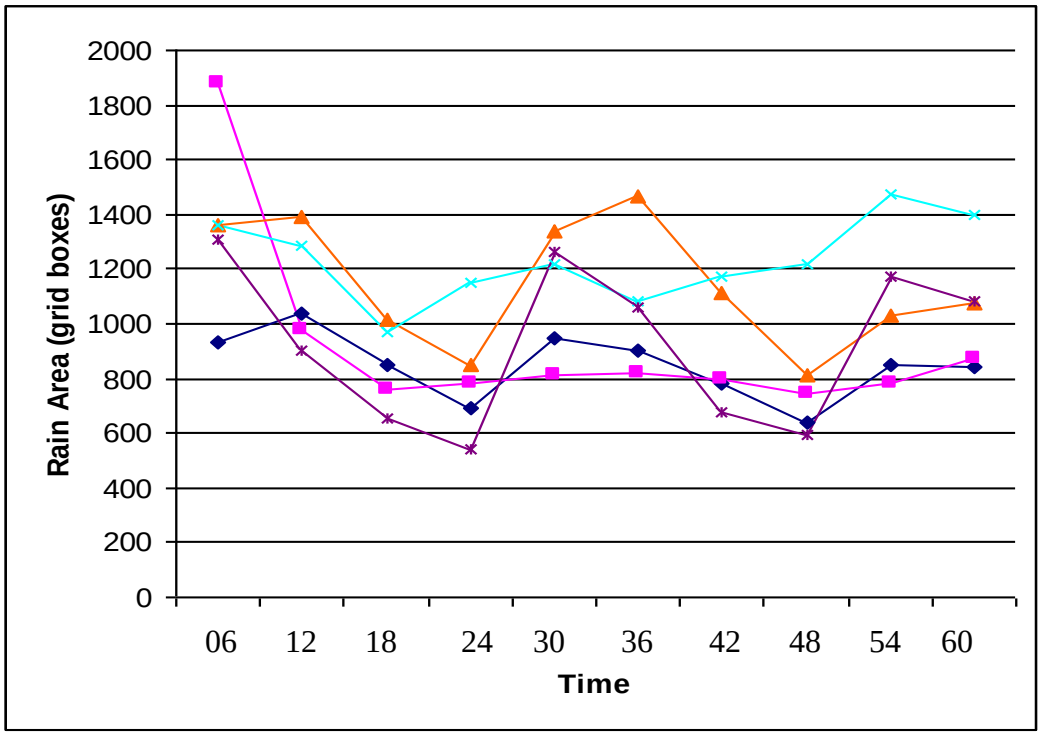

Figure 14: As in Fig. 13 except for rain area (in 10 x 10 km grid boxes), and with the observed areas shown in purple.

Although not shown, one other interesting parameter computed using MODE on probability forecasts was the average probability value for the forecasted objects. During the first 24 hours, the MIX ensemble has probability values roughly 5-10% higher than those in IC/LBC. After 24 hours, the differences increase with MIX typically 15-20% greater than IC/LBC. This result is consistent with IC/LBC having larger spread, such that probability forecasts based on its members would have lower values. The same trends are apparent in the results based on a 50% probability threshold (not shown).

5. Background on 4 km versus 20 km ensembles

For the 2007 NOAA Hazardous Weather TestBed Spring Experiment at SPC, CAPS ran a 10 member 4 km ensemble over a roughly 3000 x 2500 km domain. Five of these members used mixed physics, while the other five used both mixed physics and perturbed initial and lateral boundary conditions. Simulations were initialized at 21 UTC and run for 33 hours. Clark et al. (2009) chose the same 22 cases used by CAPS and repeated the simulations using a 30 member 20 km WRF ensemble. For the comparison of the two ensembles, Clark et al. chose the 5 4 km members having both mixed physics and initial and lateral boundary conditions, and 15 of the 20 km members also making use of all of these variations. The precipitation output was put onto a 20 km grid that was a roughly 2000 x 2000 km subset of the 20 km and 4 km model domain. The primary purpose of the comparison was to determine if a fine grid ensemble restricted to just a few members due to computational costs would perform better or worse than a coarser ensemble that could be run with more members. Comparisons were made with the full 15 20 km members and with a subset of just 5 to allow an apples-to-apples comparison with ensembles having equal numbers of members. It was found that in general, the 4km ensemble performed better than the 20 km ensemble, even when the 5 members were compared against the 15 members of the 20 km ensemble. The improved performance seemed to be related to better depiction of the diurnal cycle of convection in the central United States, and a faster error growth that occurred in the 4km members, increasing the spread. Among the findings in Clark et al. (2009) was that the probability matched ensemble mean of the 4 km ensemble had its greatest improvement in Equitable Threat Score over the 20km ensemble during the 06-12 UTC period, near the time of a diurnal maximum in precipitation. Also, rank histograms for the 4km ensemble became flatter than those of the 20 km ensemble after 18 UTC, implying better depiction of forecast uncertainty at those later times.

To determine whether or not the object parameters showed some of the same trends, MODE was used on 6 hour accumulation periods covering 00-06, 06-12, 12-18, 18-00 and 00-06 UTC, or the 3-9, 9-15, 15-21, 21-27, and 27-33 hour forecast periods. The comparisons used five members for each ensemble, with the rainfall input to MODE on a 20 km grid.

The average rain area in terms of grid boxes (each grid box is 20 x 20 km) and standard deviation of rain area from both the 4km and 20km ensembles and the observed rain area at each 6 hr period are shown in Fig. 15. The diurnal minimum during the 12-18 UTC period can be seen in the observations, with higher values during the 00-12 UTC period.

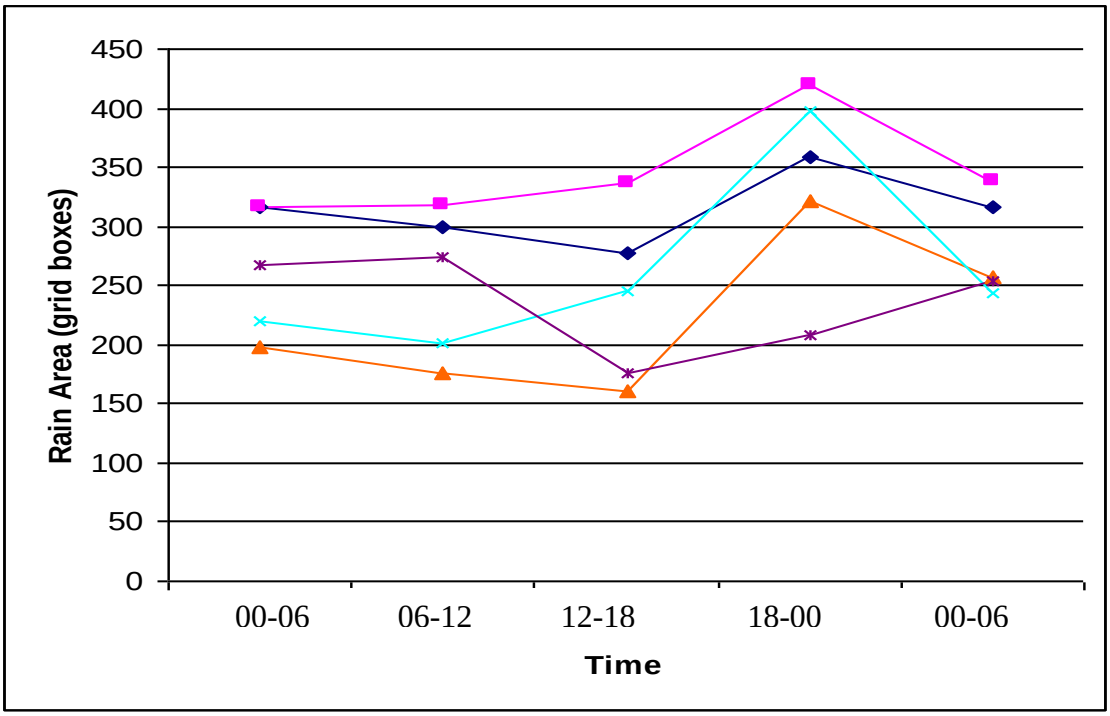

Figure 15: Average rain area (in 20x20 km grid boxes) as a function of time for the 4km (dark blue) and 20 km (pink) ensembles, along with the observed value (purple) and the standard deviations for the 4 km (orange) and 20 km (cyan) ensembles.

Both ensembles incorrectly show a peak during the 18-00 UTC period, and both show an overestimate (high bias) at all times. However, the 4 km ensemble has less of a high bias and does show a minimum during the 12-18 UTC period, unlike the 20 km ensemble. Standard deviations for the ensembles generally follow the same trends as the average rain area. During the final 6 hours of the forecast, the spread of the 4km ensemble becomes larger than that of the 20 km ensemble. This is a bit later than in the histograms. discussed in Clark et al. (2009) where increasing spread was apparent in both the 18-00 and 00-06 UTC periods. Also, unlike the ETS results discussed in Clark et al., the biggest improvements in the 4 km depiction of rain area relative to the 20 km ensemble occurred during the 12-18 and 18-00 UTC periods, and not in the 06-12 UTC period.

Average rain rates for the ensembles and observations, along with the standard deviations are shown in Fig. 16. Both ensembles tended to predict the rain rate to within 10% of the observed value. At all time except for the diurnal minimum (12-18 UTC), the two ensembles overestimated the rates. The models correctly depicted the times of maxima and minima. At all times, the 4km ensemble had more spread than the 20 km ensemble, with a hint of faster growth of spread during the last 6-12 hours of the forecast. The 4km results were closer to the observed rates during the 00-06 UTC period, and then again in the last 12 hours of the simulation.

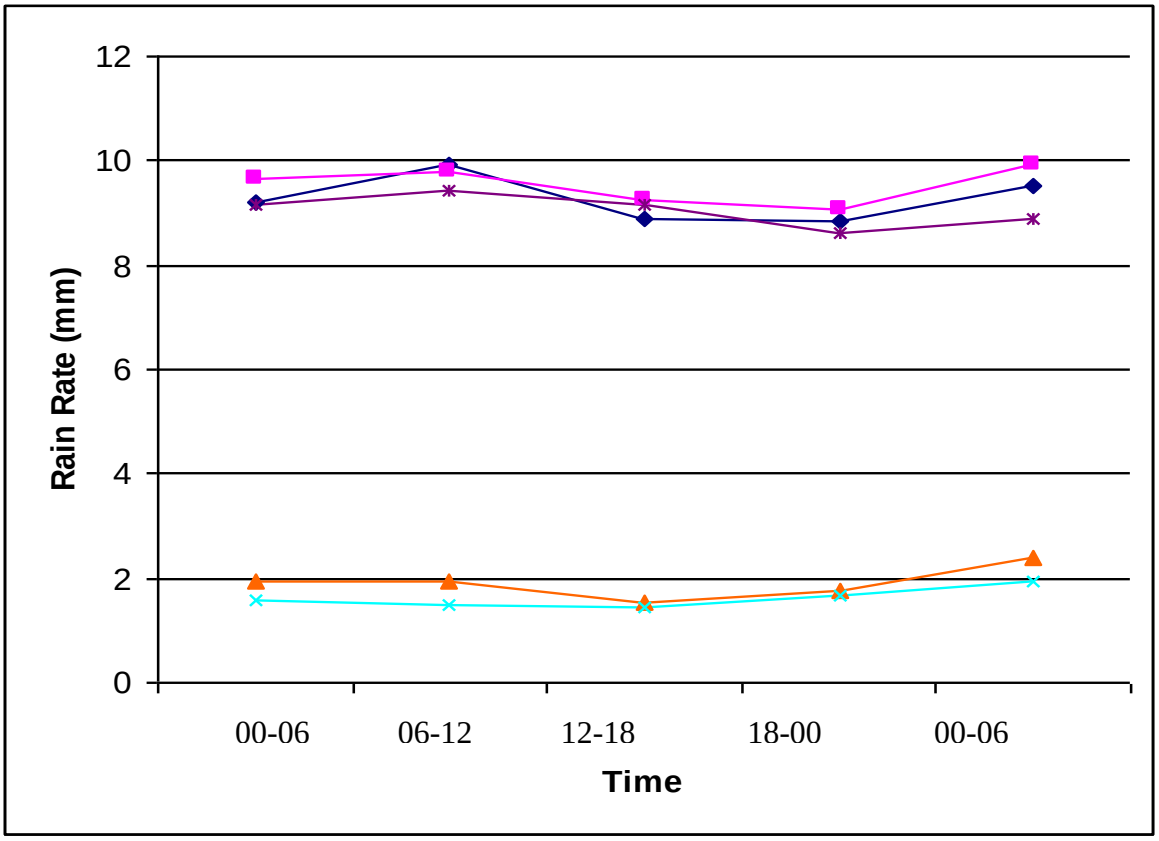

Figure 16: As in Fig. 15, except for rain rate (mm).

Rain volume is plotted in Fig. 17. As with rain area, both ensembles show a high bias at all times. The diurnal cycle is apparent in the observations, and the 4km ensemble does a better job of showing the diurnal cycle, although both ensembles are too quick to ramp up the rain volume during the afternoon (18-00 UTC). The standard deviations follow the same trends as with rain area, with the 20 km ensemble having a greater standard deviation at all times until the last 6 hours, when the 4 km ensemble may be evidencing the faster error growth discussed in Clark et al. (2009). Clark et al. also noted that in Hovmoller diagrams averaged over the model domain, the 20km ensemble probability matched mean appeared to generate the diurnal maximum of day 2 too early and too intensely, and this result may be reflected in the peak in volume occurring in the 20km data between 18 and 00 UTC (Fig. 17). Best agreement between both ensembles and the observations does occur during the 06-12 UTC period, in agreement with ETS values for the probability matched means in Clark et al. (2009).

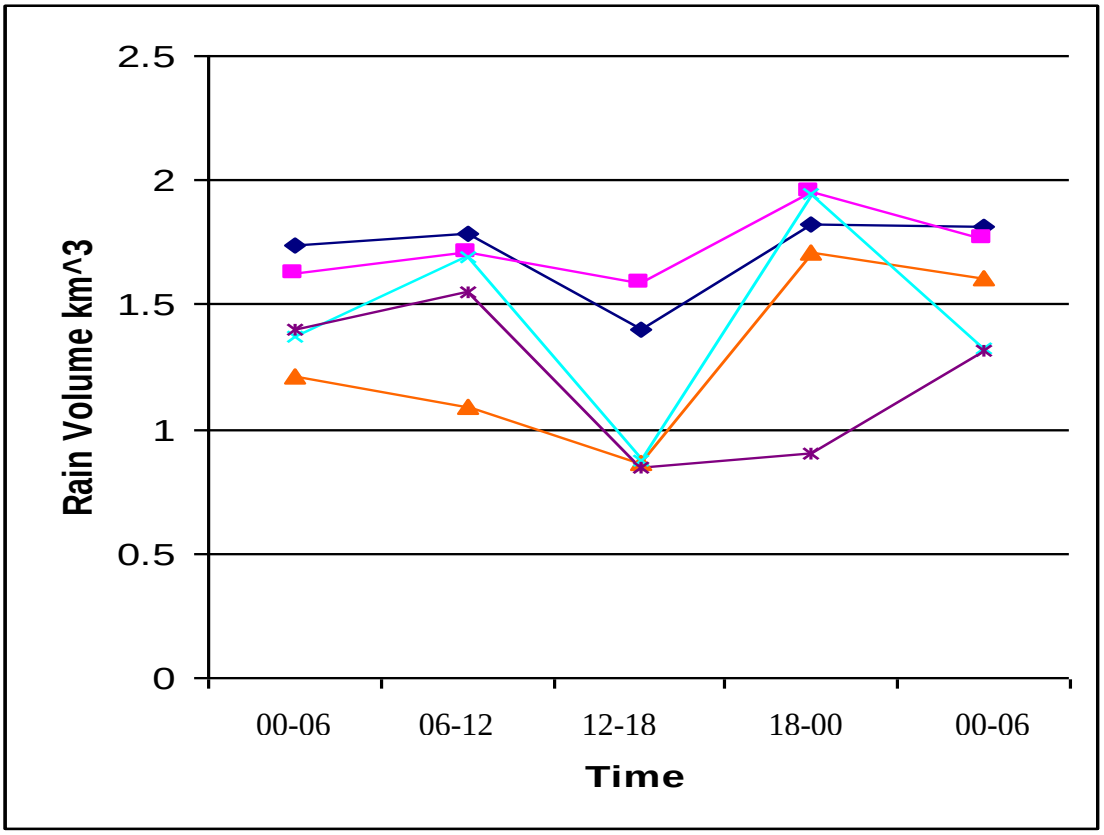

Figure 17: As in Fig. 15, except for rain volume in km^3.

Average displacements and standard deviations of displacement for the two ensembles are shown in Fig. 18. The 4km ensemble has less displacement error at all times. Both ensembles have comparable standard deviations. The displacement errors do not appear to reflect the diurnal cycle. Displacements in the 4 km ensemble were usually in the S or SW direction through the 27 h forecast (00 UTC) and then toward the NNE after that time (i.e., mean position of forecasted objects usually was SW of the observed one prior to hour 27). For the 20 km ensemble, there were no systematic trends in the displacement direction.

In summary, the object parameters evaluated from MODE often supported the conclusions based on traditional verification approaches applied to the 4 and 20 km ensembles. The object parameters indicated the 4km ensemble did a better job with depicting the diurnal cycle, and generally had smaller errors at most times for most parameters than the 20 km ensemble. There was also some evidence of the faster error growth found by Clark et al (2009) with standard deviations growing more rapidly in the 4 km parameters than in the 20 km ones, although this faster growth often only appeared in the final 6 hour period evaluated, and a longer integration period would be needed to determine if this trend is more than just a short term feature.

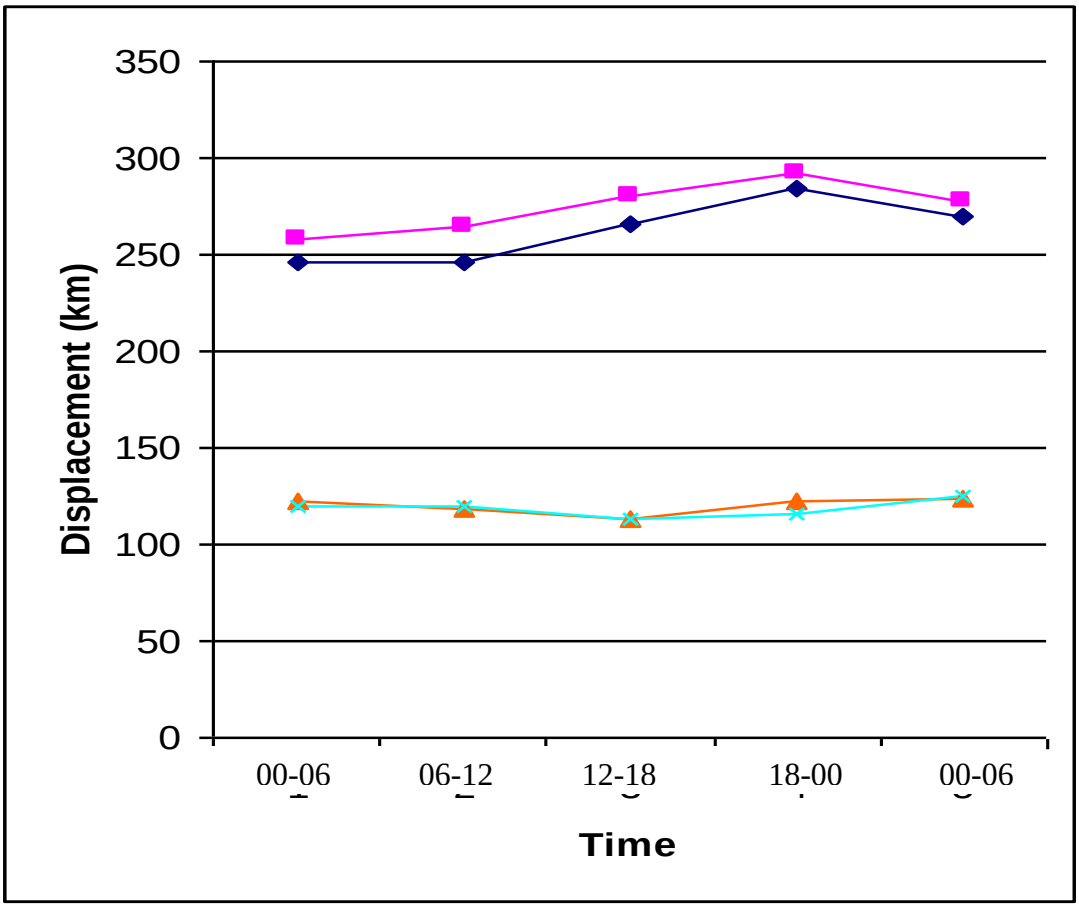

Figure 18: As in Fig. 15, except for displacement in km.

6. Comparison of Probability Matched ensemble mean values to averages of individual members

As was done for the two 8-member 15 km ensembles, a few tests were performed to see if the average of a parameter value from the set of ensemble members would be a better forecast than the parameter value from the probability matched ensemble mean. In Fig. 19, the results for displacement can be seen. For the 20 km ensemble, the probability matched ensemble mean precipitation field always yields a smaller displacement error than the average of the ensemble members (pink versus cyan curves). For the 4 km ensemble, the probability matched ensemble mean value is better in the first 4 evaluation periods (dark blue versus orange curves), but shows much faster error growth over time, so that the average of the individual members' displacements becomes a better forecast during the final 6 hour period. Such a trend is not as apparent in the 20 km ensemble output.

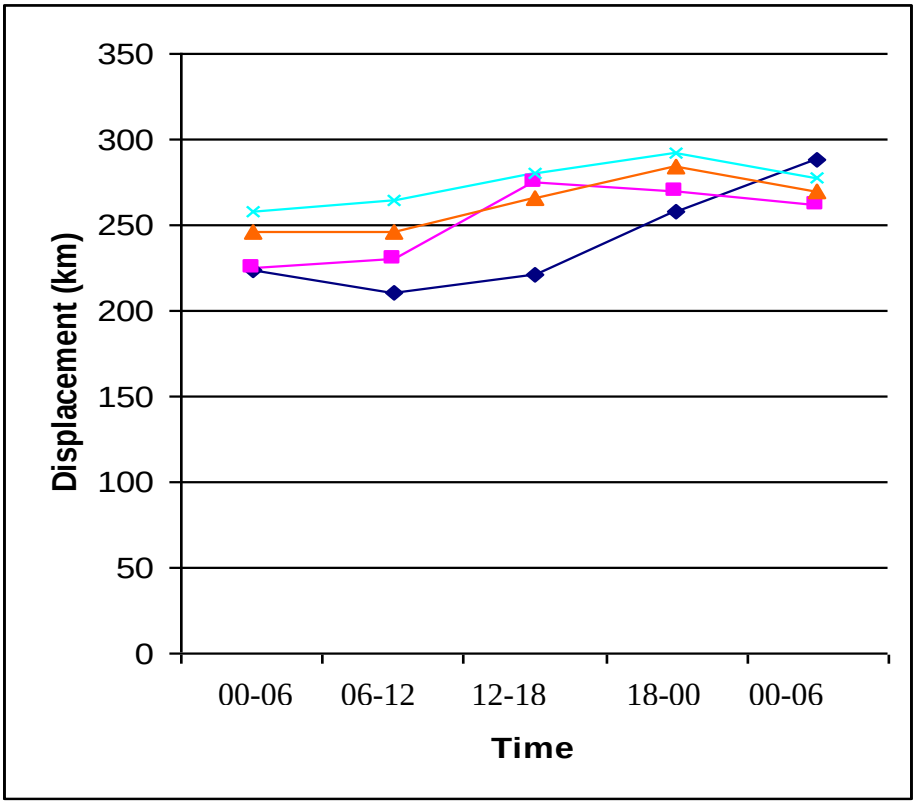

Figure 19: Displacement error (km) for the average of 4km ensemble members (orange) and the probability matched ensemble mean (dark blue), and for the average of the 20 km ensemble members (cyan) and the 20 km probability matched ensemble mean (pink).

Interestingly, different behavior is noted for the rain areas (Fig. 20). For both ensembles, the forecasted rain areas are much closer to the observed values when the average of the areas from the individual ensemble members are used compared to the areas determined from the probability matching ensemble mean. The temporal trends in the average values of area still differ substantially from the observed diurnal cycle but do appear to be slightly more realistic.

For rain rate (Fig. 21), the behavior is similar in both ensembles to rain area, with the average rate determined from the individual members being closer to the observed values than the probability matched ensemble mean. During the last 6 hours of the forecast, there is some convergence of the curves so that the probability matched approach yields comparable results to the averaging of individual members' rain rates. Note that the high bias for precipitation found by Clark et al. (2009) is apparent in both the rain area (Fig. 20) and rain rate (Fig. 21) results.

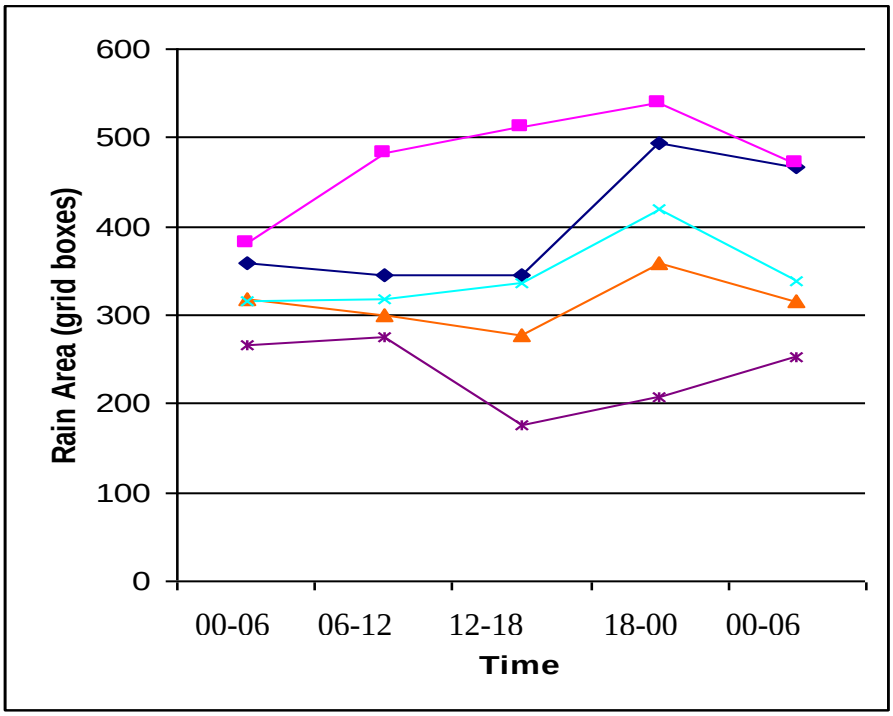

Figure 20: As in Fig. 18 except for rain area (20 x 20 km grid boxes) with observations also shown (purple).

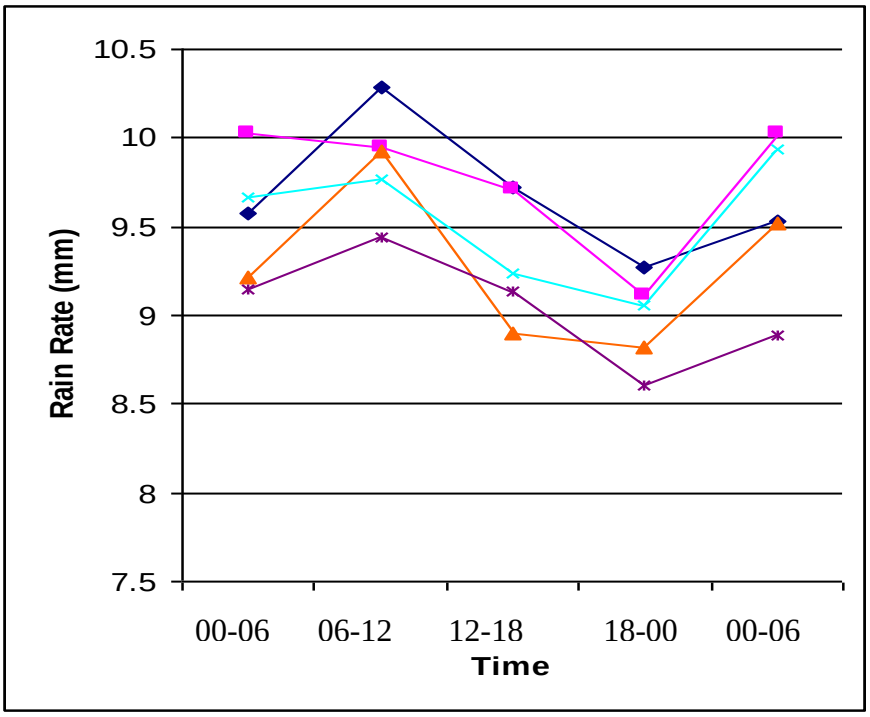

Figure 21: As in Fig. 19 except for rain rate (mm).

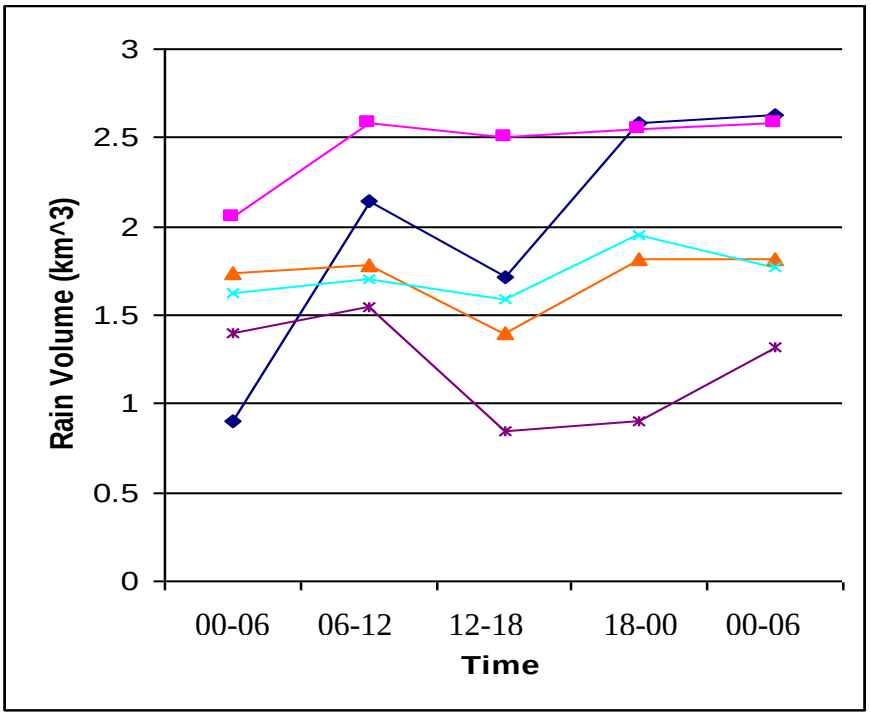

Figure 22: As in Fig. 20 except for rain volume (in km^3).

Because rain volume should be equal to the rain rate multiplied by the rain area, the behavior of the curves in Fig. 22 is similar to that in Figs. 20 and 21. Once again, the average of rain volumes from individual ensemble members yields a value closer to that observed than what is obtained from the probability matching approach. It is interesting to note that a spin up delay observed in the 4 km ensemble output in Clark et al. (2009) does show up in the probability matched results (dark blue curve) but not in the average of rain volumes from all members. Differences between the two ensembles are amplified when using the average from all members during the first 3 forecast periods. These averages are much higher for the 20 km ensemble than the 4 km ensemble.

Overall, it appears that for displacement error, the best forecast can be obtained from using the probability matched ensemble mean. However, for parameters relating to area, rate, and volume of rainfall, better forecasts may be possible by taking object-oriented verification tools and applying them to all ensemble members, and then determining the average for the parameters.

7. Conclusions

Although this DTC project involved extensive use of object-oriented techniques on a large amount of ensemble data, it really just scratched the surface and much additional work should be performed in the future to understand the best ways of applying object-oriented techniques for useful forecast guidance and verification of ensembles. This study showed that although MODE and EMT are both object-oriented, there are some differences in the results that they show. In general, the main trends determined from both techniques were consistent with each other, and at least somewhat resembled those

found from traditional ensemble verification metrics. MODE at present provides output for more parameters than EMT, which could be beneficial for forecasters and those verifying ensembles. EMT outputs information on its systems (CRAs) as individual files, which allows for easy computation of ensemble system statistics. MODE output files contain information on all objects present in the forecasted field, which could make ensemble analysis more difficult. However, the mode_analysis tool and its polygon feature help to create one possible path to facilitate computation of ensemble object parameters.

The primary goal of this study was to determine if object parameters from MODE and EMT behave in a similar way to traditional metrics. Clearly, much additional work can be done on methods of displaying the information to benefit forecasters or help with understanding model shortcomings. This study did investigate three approaches of applying object-oriented techniques. One was to apply techniques like MODE to each ensemble member, and then compute the average value for a given parameter. Another was to apply the technique to an ensemble mean of precipitation. EMT results suggested little difference in the skill of these two approaches, while MODE results suggested that the probability matched approach might work better for determining position of the systems, while the average of all of the individual members' MODE output may work better for parameters like rain area, rate and volume. Further work should be done to determine if these results apply on even larger datasets. A third approach that was tested in this study was to use MODE on probability information from an ensemble. Time constraints prevented a detailed study of sensitivity to choice of probability threshold, but the two tests performed here, using a 30% threshold and a 50% threshold for probability, did imply that better skill might result for some parameters when probabilities were used, and results for some parameters were very sensitive to the probability threshold used.

One of the advantages of using an ensemble to forecast is that distributions of possible outcomes can be seen. Although the last approach discussed used probability information as input to MODE, this study did not explore using object-oriented verification guidance in a probabilistic manner. Instead, average values of parameters were computed. Fig. 23 shows the variability occurring for one 6 hour period for one case (April 2, 2006) from the 8 member 15 km mixed physics ensemble, along with a schematic of a type of forecast that might be issued based on the averages of parameters determined from MODE. The star in the lower right panel represents the centroid location, and the ellipse is drawn to roughly represent the mean area and axis angle. As can be seen in the other panels, several difficulties have to be considered with such a forecast. Some of the members show one large object in the Midwest while others show two or more. The ensemble members for which MODE depicts multiple objects would contribute some objects having very small areas and axis orientations quite different from the larger objects. This can be seen in the schematic where the average area for the object for this time period was only 60% of the area found in the probability-matched mean shown in the panel just to its left. Likewise, the average axis angle of -18 degrees is more E-W oriented than that of the larger object found in the other forecasts shown. Additional work is needed to find an automated way to best convey the information from MODE if applied in this manner.

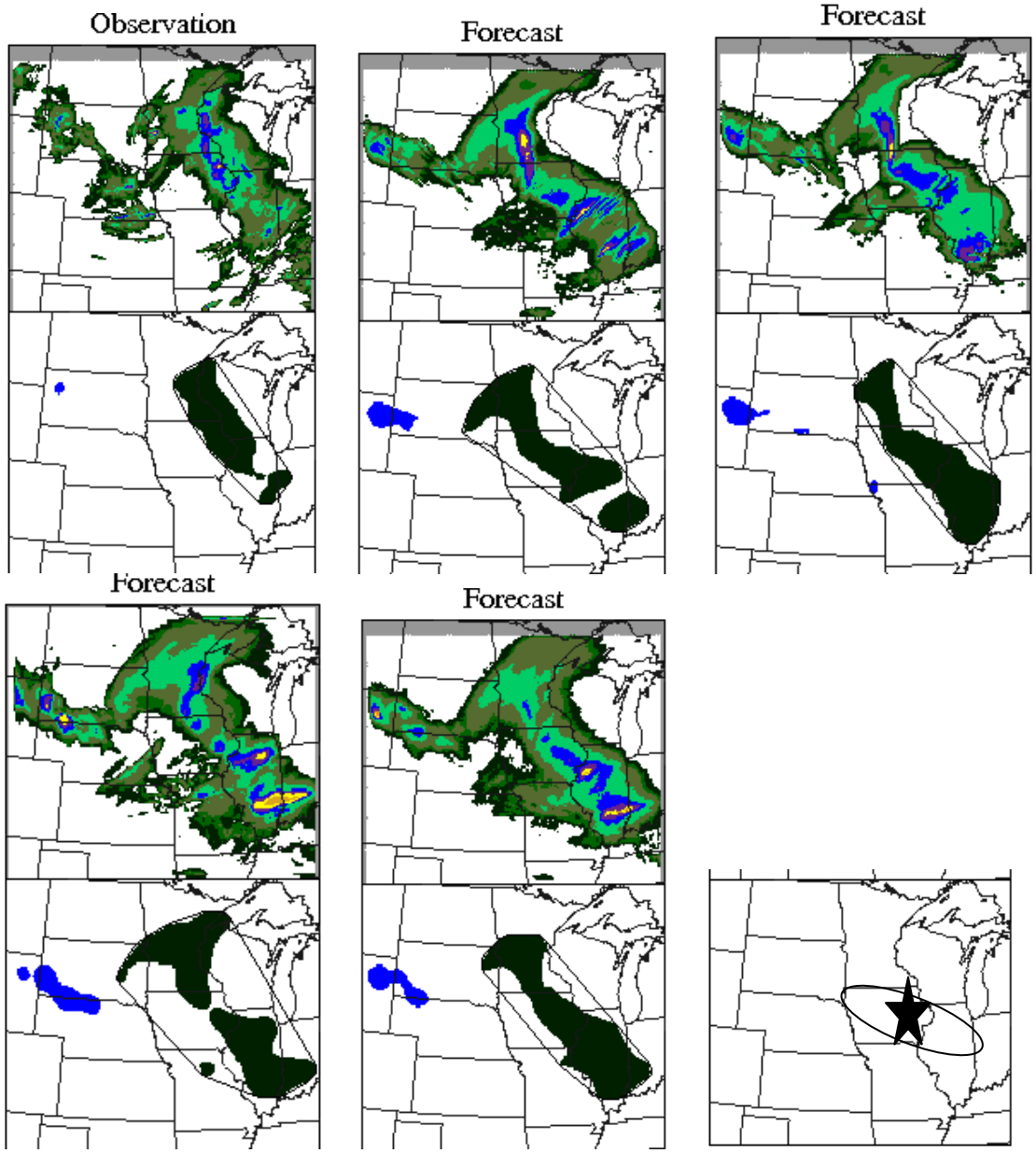

Figure 23: Variations in predictions of 6-hourly rainfall and MODE-determined objects for the 36-42 h forecast period ending at 18 UTC April 2, 2006. Observations shown in top left, with BMJ forecast having Ferrier microphysics in top center, KF forecast with Ferrier microphysics in top right, BMJ with Lin microphysics in lower left, probability-matched mean for mixed physics ensemble in lower center, and schematic depicting centroid, angle axis and area based on the average of the 8 mixed physics members in the lower right.

It is likely that the output of object-oriented techniques applied to ensembles might be better used in a probabilistic manner. Perhaps MODE output could be used to issue forecasts of precipitation areas, average rates, or volumes having sizes within some range, or exceeding thresholds. Once again, care would have to be taken since small objects identified near larger ones in some members may adversely impact the results, or at the very least, complicate interpretation of the parameter data. MODE could also be used to plot expected locations and tracks of objects during a forecast. PDFs could be created showing the distribution of object parameters in the ensemble forecast. These could be used in a forecasting mode to issue probability forecasts, or in a verification mode where the distributions would be compared to observed ones for the time period analyzed.

8. ACKNOWLEDGEMENTS

The work performed for this DTC project would not have been possible without the assistance of many individuals, both at NCAR and ISU. Special thanks are given to John Halley Gotway who frequently helped to get both MODE and the MODE Analysis tool working with the precipitation data used in this study, and taught me how to use the statistical package R which was used in the analysis. Randy Bullock also assisted in some of these same areas. Louisa Nance and Pam Johnson helped to answer important questions and ensure good access to computational resources during the DTC visits and throughout the project. Interesting discussions with Chris Davis and Barb Brown were appreciated. At ISU, thanks are given to Adam Clark for providing all of the ensemble output, Jimmy Correia for assisting with a needed NCL conversion script, Jon Hobbs for some R assistance, and Daryl Herzmann for assistance with computer issues.

9. PRESENTATIONS AND PUBLICATIONS BASED ON WORK:

Gallus, W. A, Jr., 2008: The CRA technique applied to "fake" cases. Presentation, *Intercomparison Project Workshop*, Boulder, CO, Apr. 14.

Gallus, W. A., Jr., 2007: Object-oriented verification methods applied to ensemble forecasts. *ISU Meteorology Program Seminar*, Oct. 30.

One or more conference papers on the work are planned for the next Amer. Meteor. Soc. Weather Analysis and Forecasting/Numerical Weather Prediction Conference, presumably in 2009.

A manuscript describing the work and results will likely be prepared for submission to *Weather and Forecasting*, with submission tentatively planned before January 2009.

10. REFERENCES

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

Clark, A. J., W. A. Gallus, Jr., and M. Xue, 2009: A comparison of precipitation forecast skill between small convection-resolving and large non-convection-resolving ensembles. *Mon. Wea. Rev.* (to be submitted).

Davis, C., B. Brown, and R. Bullock, 2006a: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev*., **134**, 1772-1784.

Davis, C., B. Brown, and R. Bullock, 2006b: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. *Mon. Wea. Rev.*, **134**, 1785-1795.

Ebert E. E., and J. L. McBride, 2000: Verification of precipitation in weather systems: determination of systematic errors. *J. Hydrology*, **239**, 179-202.