

**APPLICATION OF OBJECT-ORIENTED VERIFICATION TECHNIQUES
TO ENSEMBLE PRECIPITATION FORECASTS**

by

William A. Gallus, Jr.

Dept. of Geological and Atmospheric Science

Iowa State University

Ames, IA 50011

Submitted to Weather and Forecasting on January 28, 2009

Corresponding Author: William A. Gallus, Jr., 3025 Agronomy, Iowa State University, Ames, IA
50011, wgallus@iastate.edu

Abstract

Both the Method for Diagnostic Evaluation (MODE) and Contiguous Rain Area (CRA) object-oriented verification techniques have been used to analyze precipitation forecasts from two sets of ensembles to determine if spread-skill behavior observed using traditional measures can be seen in the object parameters. One set consisted of two 8 member Weather Research and Forecasting (WRF) model ensembles, one having mixed physics and dynamics with unperturbed initial and lateral boundary conditions (Phys) and another using common physics and a dynamic core but with perturbed initial and lateral boundary conditions (IC/LBC). Traditional measures found that spread grows much faster in IC/LBC than in Phys so that after roughly 24 hours, better skill and spread are found in IC/LBC. These measures also reflected a strong diurnal signal of precipitation. The other set of ensembles included 5 members of a 4 km grid spacing WRF ensemble (ENS4) and 5 members of a 20 km WRF ensemble (ENS20). Traditional measures suggested that the diurnal signal was better in ENS4 and spread increased more rapidly than in ENS20.

Standard deviations (SDs) of four object parameters computed for the first set of ensembles using MODE and CRA showed the trend of enhanced spread growth in IC/LBC compared to Phys that had been observed in traditional measures, with areal coverage of precipitation exhibiting the greatest growth in spread with time. The two techniques did not produce identical results, although they did show the same general trends. A diurnal signal could be seen in the SDs of all parameters, especially rain rate, volume, and areal coverage. MODE results also found evidence of a diurnal signal and faster growth of spread in object parameters in ENS4 than in ENS20.

Some forecasting approaches based on MODE and CRA output were also demonstrated. Forecasts based on averages of object parameters from each ensemble member were more skillful than

forecasts based on MODE or CRA applied to an ensemble mean computed using the probability matching technique for areal coverage, but differences in the two techniques were less pronounced for rain rate, volume, and displacement. The use of a probability threshold to define objects was also shown to be a valid forecasting approach with MODE.

1. Introduction

Mass et al. (2002), among others, have discussed several problems with using traditional point-to-point verification measures to evaluate precipitation forecasts from models using fine grid spacing. For instance, these measures can double penalize forecasts that may show a surprisingly accurate depiction of the shape and fine-scale pattern of the precipitation fields if a small displacement error is present (Ebert and McBride 2000; Baldwin and Kain 2006). Some of the scores have been shown to be inconsistent with the subjective impressions of forecasters (Chapman et al. 2004).

In an effort to provide more informative measures of forecast performance that better reflect the quality of these finer grid forecasts, several new spatial verification techniques have been proposed including neighborhood or fuzzy verification, scale decomposition, object-oriented verification, and field verification approaches (see Casati et al. 2008 and Gilleland et al. 2008 for reviews). Object-oriented approaches compare the properties of matched forecast and observed objects, where the object may be, for instance, a precipitation system determined using rainfall or reflectivity data. Object oriented techniques verify the location, size, shape, intensity, and other attributes of the object, and are therefore very intuitive in their interpretation (Ebert and Gallus 2009). One of the first object-oriented approaches developed was the Contiguous Rainfall Area (CRA) method (Ebert and McBride 2000) which was later used to explore systematic model biases in prediction of Central U.S. mesoscale convective systems (Grams et al. 2006). More recently, the Method for Object-based Diagnostic Evaluation (MODE; Davis et al. 2006a, b), was developed and included as part of a community verification system known as MET (Model Evaluation Tools; <http://www.dtcenter.org/met/users>).

Object-oriented techniques have traditionally been applied to deterministic forecasts where a set of forecasted objects is matched with observed ones. However, with the increasing use of ensembles at fine grid scales (e.g. Xue et al. 2007; Clark et al. 2009), an incentive exists to determine the best ways of using these techniques to both evaluate and provide better forecast guidance from ensemble forecasts, beyond simple application of the object-oriented methods to the ensemble mean forecast.

In this paper, both CRA and MODE are applied to two different sets of ensembles to examine how closely the behavior of the object parameters matches results found from traditional ensemble spread and skill measures applied to these two sets of ensembles. The first set was used by Clark et al. (2008) to compare the temporal evolution of skill and spread in an ensemble using mixed physics and mixed models along with no perturbation of initial conditions (ICs) or lateral boundary conditions (LBCs) with an ensemble having fixed physics but perturbed ICs and LBCs. The second set of ensembles, examined by Clark et al. (2009), was used to compare skill and spread between a relatively coarse grid spacing ensemble with 15 members and a finer grid spacing ensemble with only 5 members. Section 2 discusses the methodology and data used. Results for the comparison of the first set of ensembles are found in section 3. Section 4 describes the results for the second set of ensembles, and Section 5 discusses some methods for using object parameters derived from ensembles in forecasting. A discussion and conclusions follow in Section 6.

2. Data and Methodology

a. CRA and MODE

Two different object-oriented verification techniques, CRA and MODE, were used to evaluate ensemble precipitation output. The CRA method was developed to evaluate systematic errors in the prediction of rain systems (Ebert and McBride 2000, Grams et al. 2006). It measures errors in the predicted location of rain systems, and can be used to separate the total error into components due to incorrect location, incorrect amplitude, and differences in fine scale pattern. The CRA technique is described in detail in Ebert and McBride (2000), Grams et al. (2006) and Ebert and Gallus (2009), and only a brief overview will be provided here. In CRA, an entity finder is applied to isolate distinct contiguous rain areas (CRAs) in the merged field of observations and forecasts according to some minimum intensity threshold. For the 6 hourly rainfall evaluated in the ensemble forecasts examined in the present study, a threshold of 6.25 mm was used. Each CRA is assigned a unique ID, and a rectangular bounding box is fit to the CRA and then expanded by a certain distance on all sides to define a search area for the best forecast match. In the present study, the search distance was set to 300 km, a value only slightly larger than the 240 km used by Grams et al. (2006) to examine MCSs. Forecast rain features outside the search area are considered to be unrelated to the observed feature.

To match the forecast and observed entities within a CRA, the forecast is horizontally translated over the observations until a best-fit criterion is satisfied. The best-fit criterion can be minimum squared error (Ebert and McBride 2000), maximum correlation coefficient (Grams et al. 2006), or maximum overlap (Ebert et al. 2004). Because some recent studies suggest that the correlation matching is more successful than the minimum squared error matching (Grams et al. 2006, Tartaglione et al. 2005), this criterion was used in the present study. The present study examines the object parameters of system average rain rate, rain volume, areal coverage of rain, and displacement error.

MODE (Davis et al. 2006a, b) is a more recent object-oriented technique that allows more

user flexibility in determining how observed and forecast systems are merged and/or matched through the use of fuzzy logic. An interest parameter that can be a function of numerous other parameters such as distance between centroids or edges of the systems, agreement in angle of orientation of systems, areas, and intersecting region areas, among others, is specified by the user. Unlike CRA, the MODE technique smoothes the precipitation fields through the use of convolution before identification and matching of systems is performed. Also unlike CRA, forecasted systems can be matched to observed systems that are not contiguous, and may be separated by some distance. The user-specified formula for the interest parameter and the threshold value for that parameter determine how far apart the two systems can be and yet still be considered the same event. As in CRA, a rainfall threshold is used to define systems, and numerous object attributes are computed. The same threshold of 6.25 mm used in CRA was used in MODE with results again analyzed for rain rates, rain volumes, areal coverage, and displacement errors of the systems.

b. Model forecasts and observations

To examine the use of object-oriented verification methods on ensemble 6-hourly accumulated precipitation forecasts, two different sets of ensemble forecasts from the Weather Research and Forecasting (WRF; Skamarock et al. 2001) model were evaluated. The first set included an 8 member 15 km grid spacing ensemble that used unperturbed ICs and LBCs with mixed physics and dynamic cores (hereafter Phys), and another 8 member 15 km ensemble using a fixed dynamic core and physics package but perturbed ICs and LBCs (hereafter IC/LBC). Clark et al. (2008) compared the precipitation forecasts of these two ensembles, integrated for 120 hours for 72 cases, using traditional verification metrics and found that the spread and skill

of the two ensembles was initially comparable but after roughly 24 hours, the lack of perturbed LBCs reduced the growth of spread in Phys so that better spread and skill were found at later times in the IC/LBC ensemble. In addition, Clark et al. (2008) noted that a diurnal signal reflecting heavier nocturnal precipitation in the region could be seen in some of the traditional verification measures. In the present study, both MODE and CRA were applied to the first 60 hours of the forecasts from both ensembles to determine if the object parameters identified by both approaches behaved in a similar manner to the traditional spread and skill measures, and to examine how the diurnal precipitation cycle influenced the object parameters. As in Clark et al. (2008), both the WRF forecasted precipitation and Stage IV observations used for verification were remapped to a 10 km grid before being input to CRA and MODE.

The second set of ensembles evaluated included 5 members of a 10 member 4 km grid spacing WRF ensemble (hereafter ENS4) run by the Center for the Analysis and Prediction of Storms (CAPS; Xue et al. 2007) for the 2007 National Oceanic and Atmospheric Administration (NOAA) Hazardous Weather Testbed Spring Program, and 5 members of a 15 member 20 km grid spacing WRF ensemble (hereafter ENS20) run for the same 23 cases (Clark et al. 2009). All of these ensembles were constructed using mixed physics and perturbed ICs and LBCs. The 5 members analyzed from the 20 km ensemble were those members having the best statistical consistency, as in Clark et al. (2009). That study found in a comparison of these two ensembles using traditional measures that the explicitly-resolved convection in ENS4 led to a much better representation of the diurnal cycle than in ENS20 whose members used convective parameterizations. Possibly because of the better diurnal signal, ENS4 was more skillful than ENS20, even when the 4 km ensemble's 5 members were compared to the full 15 members of the 20 km ensemble. Spread was also found to increase more rapidly with time in ENS4 than in ENS20. MODE was used to evaluate the rainfall systems in these 33 h forecasts to determine if

object parameters also reflected the improved forecast of the diurnal signal in ENS4, and showed the same differences in spread growth. As in Clark et al. (2009), the comparison of the two ensembles was performed on a 20 km grid that was basically a 2000 x 2000 km subset of the coarser ensemble domain.

3. Comparison of the mixed physics ensemble and mixed IC/LBC ensemble

Using the first 60 hours of the forecasts from Clark et al. (2008), the standard deviations (SDs) within the 8 member ensembles for several object parameters computed by CRA and MODE were compared. SDs were used as a measure of spread, and data were examined 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 (systems). For CRA and MODE, a system had to be depicted in at least 4 of the 8 ensemble members to be included in the analysis. A Welch two-sided t-test was used within the R statistical package to test for statistical significance of differences in the SDs between the two ensembles. No differences identified by MODE were statistically significant, but for some parameters in CRA, the differences were significant.

Standard deviations for rain rate show a more consistent tendency for increases with time in IC/LBC than in Phys (Fig. 1). In fact, the nearly 11% difference in the relative rate of increase (slope normalized by the average rain rate in all curves at all times) of a best-fit line between the two ensembles in the CRA data is the largest difference among the four parameters examined in the present study. During at least the first five forecast periods, both CRA and MODE indicate greater SDs in Phys than in IC/LBC, but by the last forecast period, both techniques show a larger SD in IC/LBC than in Phys. Some of these differences in the CRA results were

statistically significant at the 95% confidence level. This result implies that a mixture of different physical schemes is necessary to result in more variability in rain rates until enough time has passed that differences in LBCs likely affect the atmospheric conditions contributing to precipitation systems, and hence rain rates, in the IC/LBC members. Although not shown, it should be pointed out that observed rain rates evidenced a diurnal cycle with maxima in the 00-06, 24-36, and 48-60 hour periods, and minima during hours 06-18, and 36-48. The model forecasts (not shown) missed the first diurnal peak, possibly evidence of spin up problems during the 00-12 h period, but did capture the other extrema in the diurnal cycle, albeit with less amplitude than that observed (e.g., peak variation in Phys of 20.1 mm, and IC/LBC of 18.0 mm compared to observed variations of 20.6 mm and 31.5 mm). The diurnal signal is most apparent in the SDs for Phys from CRA (Fig. 1), and is stronger in the Phys ensemble than in the IC/LBC ensemble using both techniques.

Standard deviations for rain volume from both CRA and MODE do show some diurnal signals in both Phys and IC/LBC (Fig. 2). In addition, the MODE output clearly shows a faster rate of growth for SDs in IC/LBC than in Phys. Unlike with rain rate, no differences between IC/LBC and Phys were statistically significant. Clark et al. (2008) showed a strong diurnal cycle in observed 3-hourly rain volume (maxima at hours 00-06, 24-30, and 48-54) that also occurred with much smaller amplitude in both ensemble forecasts. That study also showed that the Phys ensemble members had larger 3-hourly rain volumes than the members of IC/LBC and that Phys possessed larger spread for rain volumes compared to IC/LBC. The larger volumes in Phys in Clark et al. (2008) correspond with larger SDs for Phys found in the present study.

Standard deviation of areal coverage of rainfall in terms of 10 x 10 km grid boxes is shown in Fig. 3. The MODE results differ from the CRA results in the first 18 hours with CRA showing relatively constant SDs with time 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 standard deviation happening in the IC/LBC ensemble. In the CRA results, some of the differences between the two ensembles are statistically significant at the later times. Overall, SDs for rain area increase more rapidly with time than for the other parameters examined. It should be noted that the observed areas strongly reflected a diurnal cycle with maxima/minima at roughly the same times as the peaks in rain rate (not shown). 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 (not shown). The SDs in Phys depict more of a diurnal cycle than those of IC/LBC. It is possible the cycle in IC/LBC is hidden somewhat by the faster rate of growth of SDs with time in that ensemble than in Phys.

As with the other parameters, SDs of displacements from both techniques (Fig. 4) suggest a faster rate of growth for IC/LBC than for Phys. In addition, at most times, the SDs are larger in IC/LBC than in Phys. Perhaps most noticeable in Fig. 4 is the large difference in the magnitudes of the SDs between the two techniques. It is likely that these large differences in values are due to the fact that systems are only matched in CRA 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. Analysis of the first and last 6-h forecast periods from the 72 cases for a few of the ensemble members suggests that nearly half of the matched objects in MODE do not exhibit overlap. Average displacements for the two techniques (not shown) support this theory with CRA values generally between 100-150 km, and MODE values between 200-250 km at all times. The differences in the way the schemes operate should lead to larger average displacement errors and standard deviations in MODE than in CRA.

In summary, the trends toward increasing spread with time in the IC/LBC ensemble with less increase in the Phys ensemble seen in Clark et al (2008) are observable in the four object parameters examined, and are especially noticeable in both the MODE and CRA results for SDs of areal coverage. This result suggests that spread in areal coverage of forecasted precipitation systems is more sensitive to whether or not perturbed LBCs are used than spread in average rain rates, rain volumes, or displacement errors. For all four parameters, MODE finds spread to be greater in Phys at the early times and greater in IC/LBC at the later times; CRA only shows this for two parameters. Although the two techniques do not produce identical results, they do show the same general trends. CRA better shows differences between the two ensembles for the SDs of rain rate, while MODE shows more of a difference for SDs of rain volume. The diurnal cycle with precipitation being more abundant during the overnight hours has some influence on the SDs of these parameters, especially for rain rate, volume, and areal coverage.

4. Comparison of 4 km and 20 km grid spacing ensembles

To determine whether or not object parameters show the trends found in Clark et al. (2009) in a comparison of two ensembles using different grid spacing, 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, for the cases evaluated in that study. The comparisons used five members for each ensemble, with the rainfall input to MODE on a 20 km grid. Because the results discussed in section 3 showed that both object-oriented verification techniques depicted the same general trends, the CRA method was not used on ENS4 and ENS20 output.

The rain area (for amounts exceeding 6.25 mm) in terms of grid boxes (20 x 20 km) averaged

for all objects identified in all ensemble members and SD of rain area, from both ENS4 and ENS20, and the observed rain area for each 6 hr period are shown in Fig. 5. The diurnal minimum during the 12-18 UTC period can be seen in the observations, with higher values during the 00-12 UTC period. Both ensembles incorrectly show a peak during the 18-00 UTC period, and both show an overestimate (high bias) at all times. However, ENS4 has less of a high bias and does show a minimum during the 12-18 UTC period, unlike the 20 km ensemble. Both ensembles disagree most with observations during the daylight hours (12-00 UTC). SDs for the ensembles generally follow the same trends as the average rain area, with SDs for ENS20 larger than those for ENS4. During the last two periods, the rate of growth of spread is slightly larger in ENS4 than in ENS20. Additionally, the slope of the best fit line applied to the data from all 5 periods is slightly greater for ENS4 than for ENS20. This faster growth of spread is consistent with Clark et al's (2009) finding of faster growth in ensemble variance in ENS4 compared to ENS20. Unlike the Equitable Threat Score (ETS) results discussed in Clark et al., the biggest improvements in the 4 km depiction of rain area relative to ENS20 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 SDs are shown in Fig. 6. Both ensembles tended to predict the rain rate to within 10% of the observed value, much better agreement than found for rain area (Fig. 5). At all times except for the diurnal minimum (12-18 UTC), the two ensembles slightly overestimated the rates. The models correctly depicted the times of maxima and minima. The 4 km results were closer to the observed rates during the 00-06 UTC period, and then again in the last 12 hours of the simulation. At all times, ENS4 had more spread than ENS20, with a hint of faster growth of spread during the last 6-12 hours of the forecast. However, the slopes of the best fit lines for all of the data indicated less growth with time than for the SDs of rain area, and slightly faster growth in spread for ENS20 than for ENS4,

a result opposite to that for rain area (Fig. 5) and what was found in Clark et al. (2009). SDs were no more than 25% of the magnitude of rain rates, a much smaller fraction than for rain area (Fig. 6) where the SDs always exceeded 50% of the magnitudes of the average areas.

For rain volume (Fig. 7), as with rain area, both ensembles showed a high bias at all times, with particularly large errors during the 12-00 UTC period. The diurnal cycle is apparent in the observations, and ENS4 does a better job of showing the diurnal cycle, although both ensembles are too quick to increase the rain volume during the afternoon (18-00 UTC). The SDs follow the same trends as with rain area, with ENS20 having a greater SD at all times until the last 6 hours, when ENS4 may be evidencing the faster error growth discussed in Clark et al. (2009). The slopes of the best fit lines show the biggest difference in rate of growth with time for this parameter, with ENS4 having a noticeably larger slope. Clark et al. also noted that in Hovmoller diagrams averaged over the model domain, the ENS20 mean computed using probability matching (PM; Ebert 2001) 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 ENS20 data between 18 and 00 UTC (Fig. 7). Best agreement between both ensembles and the observations does occur during the 06-12 UTC period, in agreement with ETS values for the PM means in Clark et al. (2009). Because errors and SDs were relatively small for rain rate (Fig. 6), the larger values occurring for rain volume indicate that areal coverage of rainfall is a more troublesome forecasting challenge for the ensemble members.

Average displacements and SDs of displacement for the two ensembles are shown in Fig. 8. ENS4 has less displacement error at all times. Both ensembles have comparable SDs, roughly 50% of the magnitude of the displacements, and unlike the other three parameters examined, growth in SDs with time is negligible. The displacement errors do not appear to reflect the diurnal cycle. Displacements in ENS4 were usually in the S or SW direction (not shown) 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 ENS20, 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 that ENS4 better depicted the diurnal cycle, and generally had smaller errors at most times for most parameters than ENS20. There was also some evidence of the faster error growth found by Clark et al (2009) with SDs growing more rapidly in the ENS4 parameters than in the ENS20 ones. The analysis of object parameters suggests that average rain rate is forecasted better than area, location and volume. Also, variability in rain rate among members is less than it is for the other parameters.

5. Use of object-oriented verification parameters in ensemble forecasting

Because object-oriented verification techniques can provide information on system parameters, it is possible that such techniques could be used to provide useful forecast information. As a preliminary exploration of how such information could be used to benefit forecasters, two tests were performed. First, CRA and MODE were applied to the ensemble mean precipitation forecast generated using PM applied to the various ensembles discussed earlier, and the skill of these forecasts of object parameters was compared to that of a forecast created by averaging the CRA and MODE output for parameters in each individual ensemble member. In a second test, forecast skill was compared between a forecast for object parameters based on an average of MODE output where objects in each ensemble member were defined using the 6.25 mm precipitation amount threshold (same technique used earlier in the present study), and a

forecast where MODE objects were defined using a probability threshold applied to a probabilistic forecast created using equal-weighting of the ensemble members.

a. Comparison of PM Ensemble Mean object parameters with averages of object parameters from all ensemble members

To investigate if a more skillful forecast could be achieved by applying object-oriented techniques to a PM ensemble mean forecast or by applying the techniques to each individual ensemble member and then averaging the object parameters, CRA was used on the two 8 member ensembles discussed in section 3, and MODE was used on the two 5 member ensembles discussed in section 4.

Averages for each 6h forecast period from CRA for the 72 cases for rain rate (Fig. 9) showed that objects in the PM 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 Phys ensemble (Fig. 9a). However, since the observed value was sometimes greater than both of these forecasts and sometimes smaller, the more accurate forecast was split evenly between these two approaches. For the IC/LBC ensemble (Fig. 9b), the objects from the PM forecast had smaller rain rates only up to hour 36, with the exception of the 18-24 h period. During hours 36-60, the PM approach had larger values. Unlike the Phys 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 average of the CRA results from all 8 members about 70% of the time. Differences were usually small, however. A high bias existed in the Phys ensemble (Fig. 9a) at most times, but trended lower toward hour 60. At hours 12 and 18, all members had too high of rates compared to observations. The IC/LBC ensemble (Fig. 9b) had a low bias at

most times, with all members too light at hours 00-06, and from hours 18-60.

Similar analysis was performed using the averages of all 72 cases for rain volume and displacement error (not shown), and as with rain rate, the PM forecast was comparable in skill to one based on an average of the parameter values from individual ensemble members. Thus, at least for these three 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, volumes, or displacements exceeding thresholds or falling within binned ranges).

A comparison of areal coverage errors between forecasts using the PM mean and forecasts using averaging of individual ensemble member areas (Fig. 10) reveals somewhat different behavior. At most times, the errors are more positive for the PM approach. A strong diurnal signal can be seen in the area errors with underestimates near the times of diurnal maxima in precipitation and overestimates at other times. Because the PM mean forecasts result in larger areas of precipitation, the agreement with observations is generally worse except during the times of the diurnal maxima.

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 two 8 member ensembles was examined during all 6-h forecast periods, which were grouped into day 1, 2 and 3 forecasts. Figure 11 shows that this approach worked much better for predicting rain rate and volume than for areal coverage. For rate and volume, the two ensembles captured the observed values roughly 50% of the time; for areal coverage the figure was closer to 10%. For all three parameters, increasing skill with time was more pronounced in IC/LBC than in Phys, a result likely influenced by the faster growth in spread in IC/LBC (Clark et al. 2008).

An additional test was performed to see if a spread-skill relationship where large spread was associated with less skillful forecasts existed in the object parameters from the ensembles. Fig. 12 shows the time evolution for the Phys ensemble of mean absolute error for rain volume, rate, and areal coverage for those systems that had large SDs among the members and those that had small values. Large and small were defined to be greater than 150% of the average SD or less than 50% of the average, respectively. Rain volume and areal coverage show a clear separation with much more accurate forecasts in the events where spread is relatively small. Rain rate does not show the relationship as definitively, although for a majority of the time, it still applies. However, at hours 00-06, 18-24, and 48-60, the forecast skill either does not vary much with SD or the more accurate forecast is associated with larger SDs. It is not entirely clear why rain rate would behave differently, although it should be pointed out that the SDs for rain rate are a much smaller fraction of the average magnitudes than those for the other parameters (not shown). Average rain rates at all times for both ensembles are around 12 mm, so that the SDs (Fig. 1) are roughly 10% of the average values. Rain volumes are typically around 1 km³ (not shown), so that SDs (Fig. 2) are often over 50% of these values. Likewise, SDs for areal coverage (Fig. 3) and displacement (Fig. 4) are often 50% or more of the typical values (not shown) of around 700 (MODE) – 800 (CRA) points and 100 km (CRA) - 200 km (MODE). These results in general imply that forecasters may be able to establish a confidence level for their forecasts of some object parameters using the SDs from ensembles. The same general tendencies were found for IC/LBC (not shown), with clearly better forecasts of rain volume and areal coverage when spread was small, but mixed results for rain rate.

As was done for the two 8-member 15 km ensembles, a few tests were performed using MODE with ENS4 and ENS20 output 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 PM ensemble

mean. In Fig. 13, the results for displacement can be seen. For ENS20, the PM ensemble mean precipitation field always yields a smaller displacement error than the average of the ensemble members. For ENS4, the probability matched ensemble mean value is better in the first 4 evaluation periods, 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.

Different behavior is noted for the rain areas (Fig. 14). 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 PM ensemble mean. As was the case with Phys and IC/LBC, the areas in the PM mean forecast were larger than the average areas from the ensemble members. Because the ENS4 and ENS20 forecasts had a more persistent problem with overestimates of rainfall coverage, the PM forecasts of area were always worse than those obtained from averages of the individual ensemble members. 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 than the trends associated with the PM forecast.

For rain rate (Fig. 15), the behavior was similar in both ensembles to rain area, with the average rate determined from the individual members being closer to the observed values than the PM ensemble mean. Note that this trend differs from that found for the two 8 member ensembles (Fig. 9) where the rates determined from the PM approach and those from averaging all ensemble members were equally accurate. During the last 6 hours of the forecast, there is some convergence of the curves so that the PM 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. 14) and rain rate (Fig. 15).

Because rain volume should be equal to the rain rate multiplied by the rain area, the behavior

of the curves in Fig. 16 is similar to that in Figs. 14 and 15. Once again, the average of rain volumes from individual ensemble members yields a value closer to that observed than what is obtained from the PM 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 PM results (dotted curves) but not as much 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 ENS20 than ENS4.

Overall, it appears that for displacement error using MODE with the 4 km and 20 km ensemble output, the best forecast can be obtained by using the PM ensemble mean. However, for the parameters of areal coverage, rate, and volume of rainfall, better forecasts are possible by applying object-oriented verification tools to all ensemble members, and then averaging the parameter values from all members. Differences between the two approaches are less noticeable in the CRA results for the Phys and IC/LBC ensembles, although the averaging of individual member forecasts of areal coverage is more likely to yield a better forecast than that from the PM mean.

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

Because MODE allows a user to compare objects generated from two different fields, a test was performed in which object parameters were obtained from both Phys and IC/LBC using probability of precipitation (POP) forecasts created using equal weighting of the 8 members discussed in section 3 (thus POP values could be 0%, 12.5%, 25%, 37.5%, etc). A probability threshold of 30% to define an object was tested, meaning that at least 3 of the 8 members had to show precipitation above 6.25 mm 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.

Figure 17 compares the displacement errors using MODE applied to the POP forecasts to the errors when MODE was applied to QPF amount with the results for all individual members averaged. In general, displacement errors grow with time. Also, the results using a 30% probability threshold are generally 5-10% better than those based on QPF amount. 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, particularly for larger ensembles or more complex methods of determining the probabilities.

Average rain areas for the two ensembles using both techniques with MODE are shown in Fig. 18, 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 observations curve plotted in Fig. 18 is an average of the MODE results for objects found using Phys and IC/LBC. 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 (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 (solid lines). 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 when a higher threshold was used combining into one larger object when the lower threshold was used. 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.

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

6. Conclusions

The use of object-oriented verification approaches to evaluate and enhance ensemble forecasts was tested by using both CRA and MODE on two sets of ensembles. The first set included an 8 member ensemble with mixed physics and dynamic cores with unperturbed ICs and

LBCs (Phys) and an 8 member ensemble with fixed physics and perturbed ICs and LBCs (IC/LBC). Clark et al. (2008) found using traditional spread and skill measures that spread increased much faster in IC/LBC compared to Phys so that both spread and skill were better in IC/LBC than in Phys after roughly 24-30 hours. The second set of ensembles included 5 members of a 4 km ensemble (ENS4) and 5 members of a 20 km ensemble (ENS20). Clark et al. (2009) found that a diurnal signal in precipitation was better depicted in ENS4, and this may have partly accounted for better skill measures for ENS4 compared to ENS20. In addition, spread grew faster with time in ENS4.

Both CRA and MODE showed that in four object parameters studied, rain rate, volume, areal coverage and displacement error, greater increases in spread with time occur in IC/LBC than in Phys, agreeing with Clark et al. (2008). This trend was particularly noticeable in areal coverage of precipitation systems, suggesting that spread in areal coverage of forecasted precipitation systems is more sensitive to whether or not perturbed lateral boundary conditions are used than spread in average rain rates, rain volumes, or displacement errors. Although the two object-oriented techniques do not produce identical results, they do show the same general trends. The diurnal cycle with precipitation being more abundant during the overnight hours has some influence on the SDs of the object parameters examined, especially rain rate, volume, and areal coverage.

In a comparison of object parameters derived from MODE for ENS4 and ENS20, ENS4 was found to better depict the diurnal cycle, and generally had smaller errors at most times for most parameters than ENS20. There was also some evidence of the faster error growth found by Clark et al (2009) with standard deviations growing more rapidly in the ENS4 parameters than in the ENS20 ones. Standard deviations were much smaller for rain rate than for other parameters, and agreement with observations was also best for rain rate.

Several tests were performed also to examine methods of using object-oriented verification output to assist forecasters. It was found that predictions of areal coverage of precipitation systems are more accurate when based on the average of the predicted areas from each ensemble member as opposed to using a PM ensemble mean forecast input into the object-oriented techniques. For the other parameters, rain rate, volume, and displacement, differences in the skill of both approaches were less substantial. It was also found that object-oriented approaches making use of probability thresholds to define precipitation systems could yield forecasts of object parameters equally skillful to those based on a precipitation amount threshold.

The present study is a preliminary look at how object-oriented approaches could be applied to ensemble forecasts, and much additional work is needed. Most of the cases examined here consisted of convective systems in the central United States. Analysis should be performed on larger-scale cold season events covering other areas. In addition, the present study emphasized deterministic information obtained from ensembles, such as the average rain volume or areal coverage. Future work should examine probabilistic forecasts of object parameters, and expand the number of parameters analyzed. Sensitivity to user-defined parameters within the object-oriented verification techniques should also be explored further.

7. Acknowledgments

This research was funded in part by National Science Foundation grant ATM-0537043, and also by the WRF DTC. Special thanks are given to John Halley Gotway and Randy Bullock for assistance with MODE and methods of analyzing MODE output. Adam Clark kindly provided the precipitation output analyzed, and also provided useful feedback on the manuscript. Additional help was provided by Jon Hobbs, Daryl Herzmann, James Correia, Elizabeth Ebert,

Barbara Brown, Louisa Nance, and Pam Johnson. The author would like to thank Bob Gall and the entire WRF-DTC leadership for the opportunity to visit at the DTC and share in stimulating discussions there.

8. References

- Baldwin, M. E., and J. S. Kain, 2007: Sensitivity of several performance measures to displacement error, bias, and event frequency. *Wea. Forecasting*, **21**, 636-648.
- Casati, B., L.J. Wilson, D.B. Stephenson, P. Nurmi, A. Ghelli, M. Pocerich, U. Damrath, E.E. Ebert, B.G. Brown and S. Mason, 2008: Forecast verification: current status and future directions. *Meteorol. Appl.*, 15, 3-18.
- Chapman, M., R. Bullock, B.G. Brown, C.A. Davis, K.W. Manning, R. Morss, and A. Takacs, 2004: An object-oriented approach to quantitative precipitation forecasts: Part II – Examples. Preprints, *17th Conf. on Probability and Statistics in the Atmospheric Sciences*, Seattle, WA, Amer. Meteor. Soc., CD-ROM, J12.5.
- 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.* (in press).
- 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., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461-2480.
- Ebert E. E., and J. L. McBride, 2000: Verification of precipitation in weather systems: determination of systematic errors. *J. Hydrology*, **239**, 179-202.
- Ebert, E. E., and W. A. Gallus, Jr., 2009: Toward better understanding of the contiguous rain area (CRA) method for spatial forecast verification. Submitted to *Wea. Forecasting*.
- Ebert, E., L.J. Wilson, B.G. Brown, P. Nurmi, H.E. Brooks, J. Bally, and M. Jaeneke, 2004: Verification of nowcasts from the WWRP Sydney 2000 Forecast Demonstration Project. *Wea. Forecasting*, **19**, 73-96.
- Gilleland, E., D. Ahijevych, B.G. Brown, and E.E. Ebert, 2008: Spatial forecast verification methods inter-comparison. Submitted to *Wea. Forecasting*.
- Grams, J.S., W.A. Gallus, L.S. Wharton, S. Koch, A. Loughe, and E. E. Ebert, 2006: The use of a modified Ebert-McBride technique to evaluate mesoscale model QPF as a function of

convective system morphology during IHOP 2002. *Wea. Forecasting*, **21**, 288-306.

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

Skamarock, W. C., J. B. Klemp, and J. Dudhia, 2001: Prototypes for the WRF (Weather Research and Forecasting) model. Preprints, *Ninth Conf. on Mesoscale Processes*, Fort Lauderdale, FL, Amer. Meteor. Soc., J15. [Available online at <http://ams.confex.com/ams/pdfpapers/23297.pdf>].

Tartaglione, N., S. Mariani, C. Accadia, A. Speranza, and M. Casaioli, 2005: Comparison of rain gauge observations with modeled precipitation over Cyprus using Contiguous Rain Area Analysis. *Atmospheric Chemistry and Physics*, **5**, 2147-2154.

Xue, M., F. Kong, D. Weber, K. W. Thomas, Y. Wang, K. Brewster, K. K. Droegemeier, J. S. Kain, S. J. Weiss, D. R. Bright, M. S. Wandishin, M. C. Coniglio, and J. Du, 2007: CAPS realtime storm-scale ensemble and high-resolution forecasts as part of the NOAA Hazardous Weather Testbed 2007 spring experiment. *22nd Conf. Wea. Anal. Forecasting/18th Conf. Num. Wea. Pred.*, Park City, UT, Amer. Meteor. Soc., CDROM 3B.1.

Figure Legends

Figure 1: Standard deviation of 6-hr rainfall (mm) among the 8 ensemble members of Phys (CRA results dashed with diamonds, MODE solid with triangles) and IC/LBC (CRA results dash-dotted with squares, MODE dotted with circles) as a function of time. Differences in CRA results between Phys and IC/LBC statistically significant with p values less than .05 shown with asterisks. Slope of best-fit line for each set of data, expressed as percentage change relative to average rain rate SD, is shown in inset.

Figure 2: As in Figure 1 except for rain volume (km³).

Figure 3: As in Fig. 1 except for areal coverage (number of 10 x 10 km grid boxes) of rainfall.

Figure 4: As in Fig. 1 except for displacement error (km).

Figure 5: Rain area (in 20x20 km grid boxes) averaged for all objects in all ensemble members from MODE as a function of time for ENS4 (solid with squares) and ENS20 (solid with triangles), along with the observed value (dashed with asterisks) and the standard deviations for ENS4 (dotted with squares) and ENS20 (dotted with triangles). Slope of best-fit line for SD data, expressed as percentage change relative to average rain area SD, is shown in inset.

Figure 6: As in Fig. 5, except for rain rate (mm).

Figure 7: As in Fig. 5, except for rain volume in km³.

Figure 8: As in Fig. 5, except for displacement in km, with no observations plotted.

Figure 9: Difference between the forecasted rain rate and the observed (mm) as a function of time, based on CRA output, for a) Phys and b) IC/LBC. Curves are plotted to show the maximum and minimum values by any ensemble member (dotted lines with squares), the mean value (solid lines with squares), and the PM forecast (dashed lines with triangles).

Figure 10: Difference between the forecasted rain area and the observed (10 x 10 km grid boxes) as a function of time, based on CRA output for the mean value (solid lines with squares for Phys and triangles for IC/LBC), and the PM forecast (dotted lines with squares for Phys, triangles for IC/LBC).

Figure 11: Percentage of time that observed values of rain volume (dashed line with squares for Phys, triangles for IC/LBC), rate (solid lines with squares for Phys, triangles for IC/LBC), and areal coverage (dotted lines with squares for Phys, triangles for IC/LBC) fell within the minimum and maximum predicted by the two ensembles, based on CRA output. 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.

Figure 12: Mean absolute error for volume (km^3), rate (*2.54 mm) and areal coverage (*1000 grid boxes) of precipitation from CRA applied to Phys as a function of time for cases with standard deviations over 150% of the mean (squares with solid line for rate, dashed line for volume, and dotted line for area), and less than 50% of the mean (triangles with solid line for rate, dashed line for volume, and dotted line for area).

Figure 13: Displacement error (km) for the average of 4km ensemble members (solid line with squares) and the PM ensemble mean (dotted with squares), and for the average of the 20 km ensemble members (solid with triangles) and the 20 km PM ensemble mean (dotted with triangles), based on MODE output.

Figure 14: As in Fig. 13 except for rain area (20 x 20 km grid boxes) with observations also shown (dashed line).

Figure 15: As in Fig. 14 except for rain rate (mm).

Figure 16: As in Fig. 14 except for rain volume (in km^3).

Figure 17: 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 (solid lines with squares for Phys and triangles for IC/LBC) and using MODE applied to probability forecasts with a 30% threshold (dotted lines with squares for Phys and triangles for IC/LBC).

Figure 18: As in Fig. 17 except for rain area (in 10 x 10 km grid boxes), and with the observed areas shown with dashed line and asterisks.

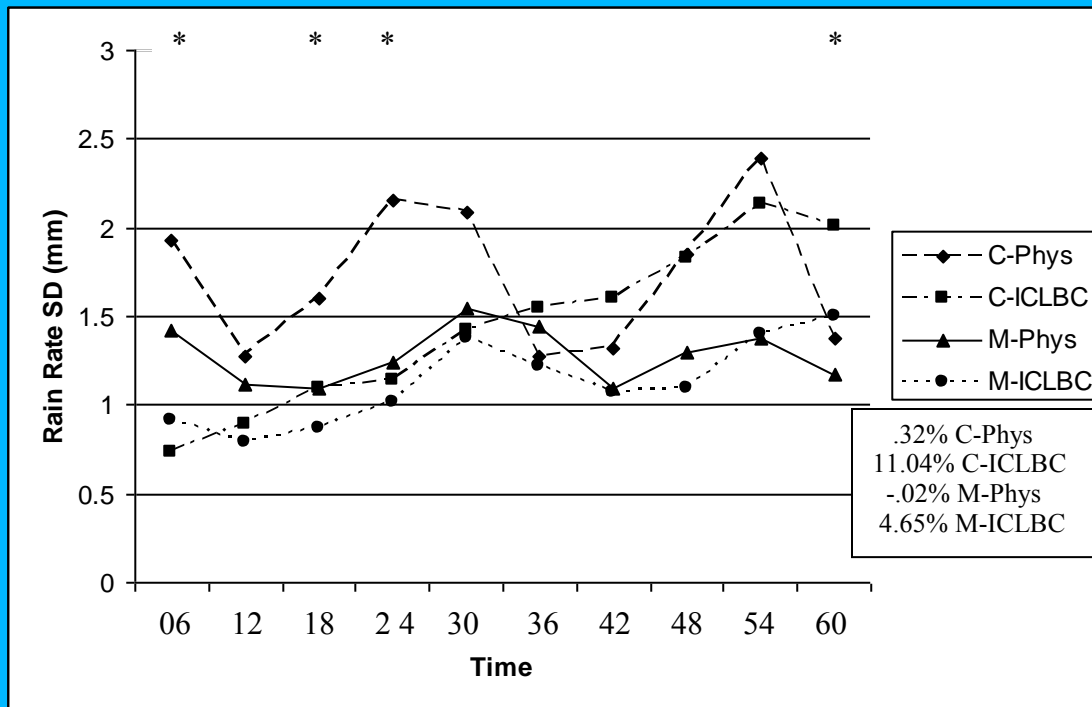


Figure 1: Standard deviation of 6-hr rainfall (mm) among the 8 ensemble members of Phys (CRA results dashed with diamonds, MODE solid with triangles) and IC/LBC (CRA results dash-dotted with squares, MODE dotted with circles) as a function of time. Differences in CRA results between Phys and IC/LBC statistically significant with p values less than .05 shown with asterisks. Slope of best-fit line for each set of data, expressed as percentage change relative to average rain rate SD, is shown in inset.

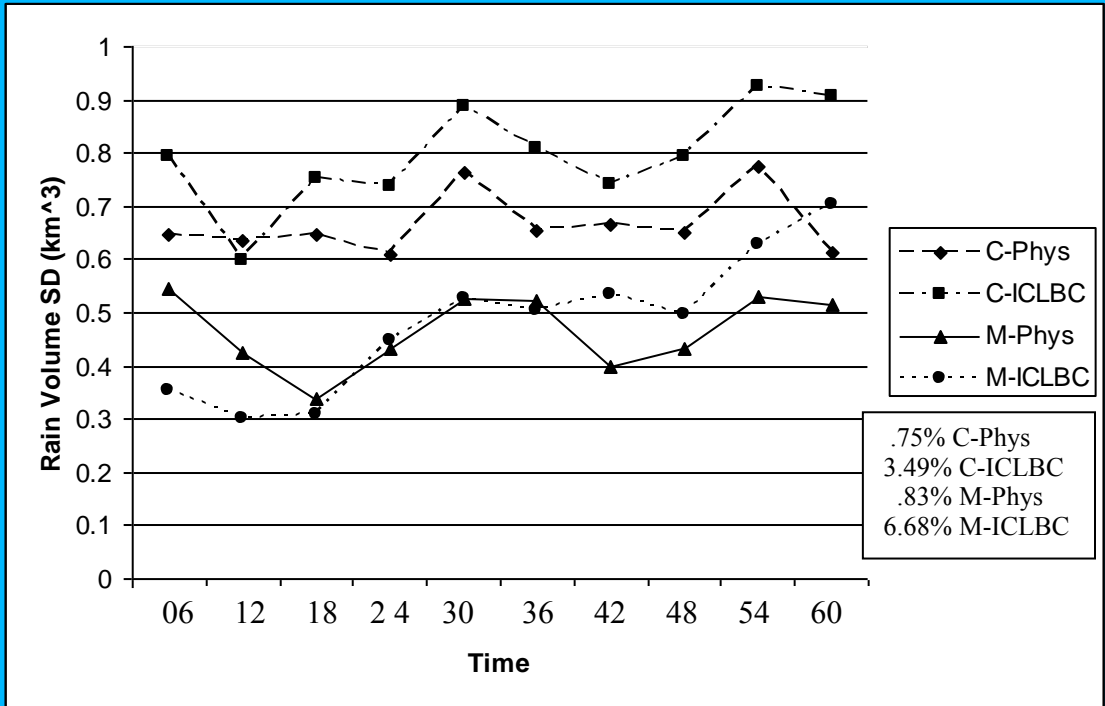


Figure 2: As in Figure 1 except for rain volume (km³).

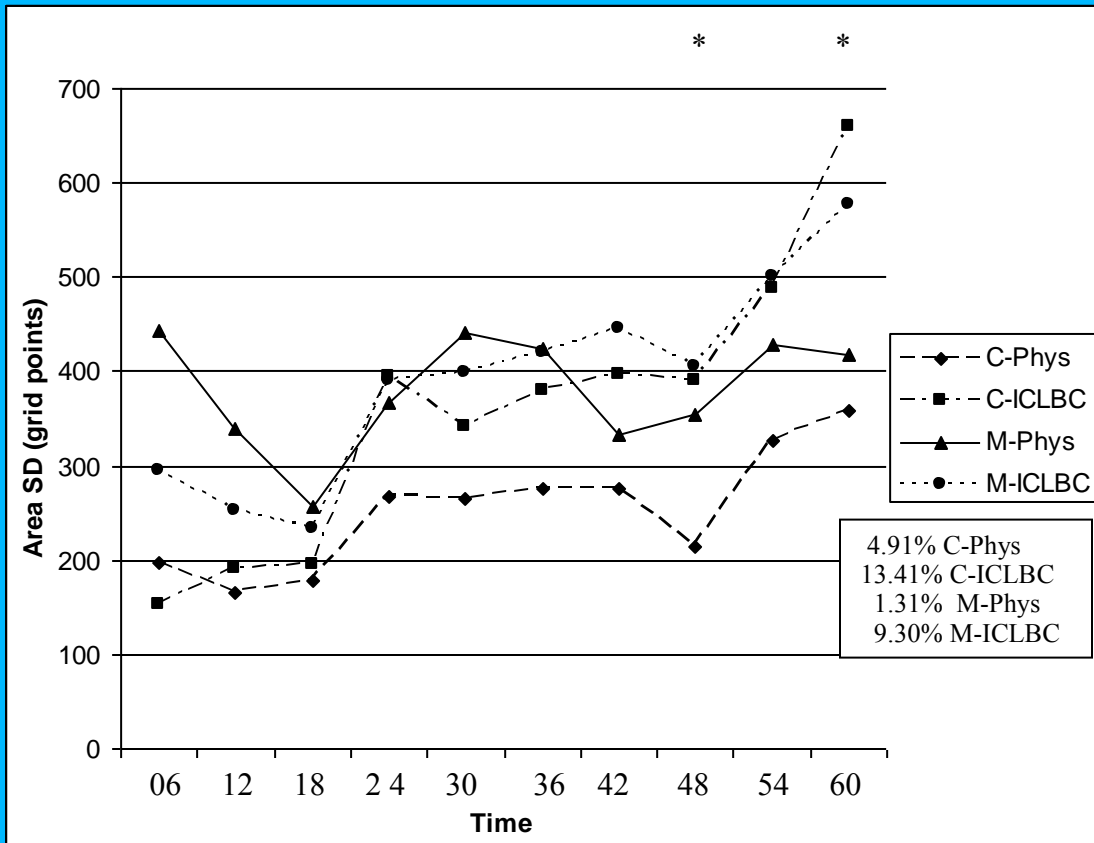


Figure 3: As in Fig. 1 except for areal coverage (number of 10 x 10 km grid boxes) of rainfall.

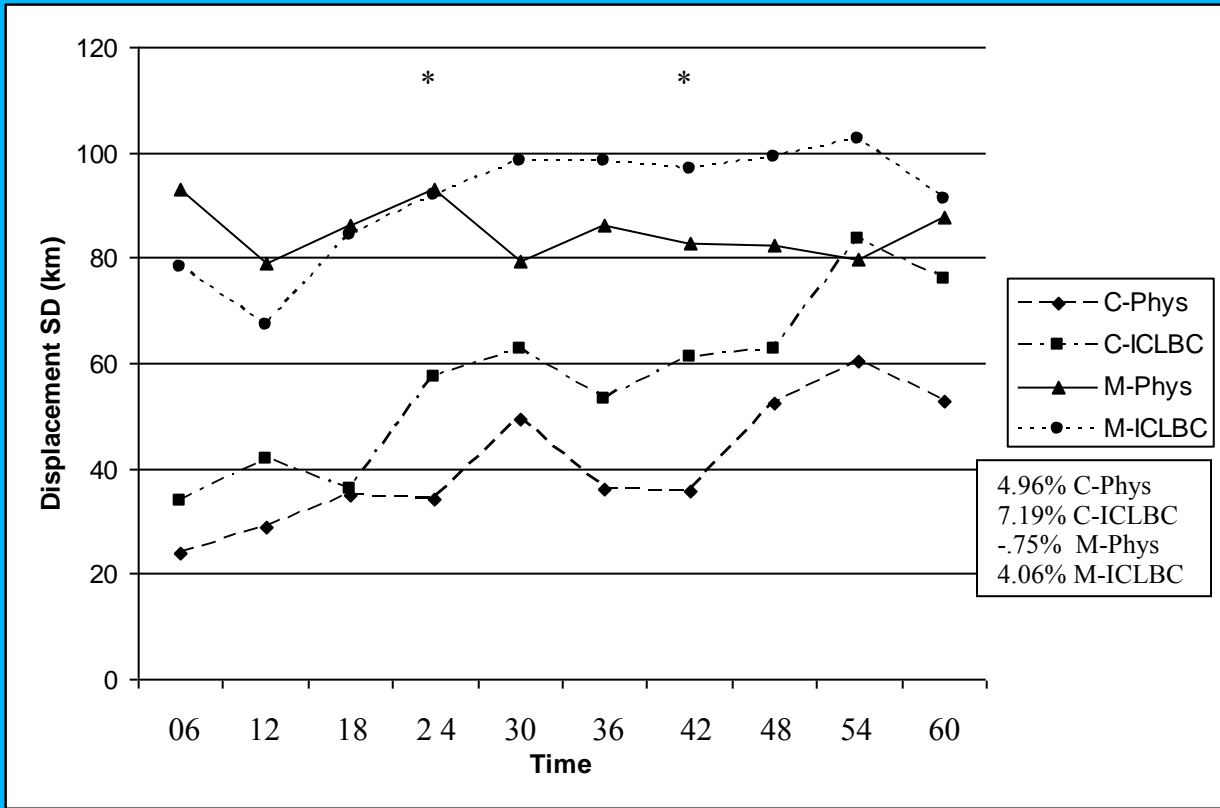


Figure 4: As in Fig. 1 except for displacement error (km).

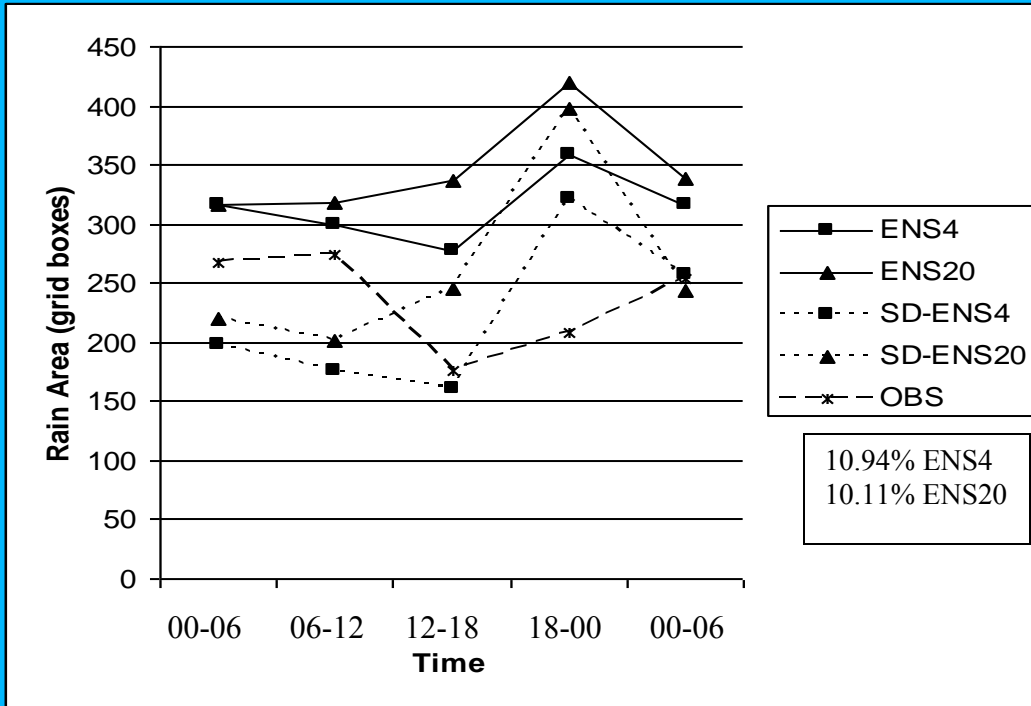


Figure 5: Rain area (in 20x20 km grid boxes) averaged for all objects in all ensemble members from MODE as a function of time for ENS4 (solid with squares) and ENS20 (solid with triangles), along with the observed value (dashed with asterisks) and the standard deviations for ENS4 (dotted with squares) and ENS20 (dotted with triangles). Slope of best-fit line for SD data, expressed as percentage change relative to average rain area SD, is shown in inset.

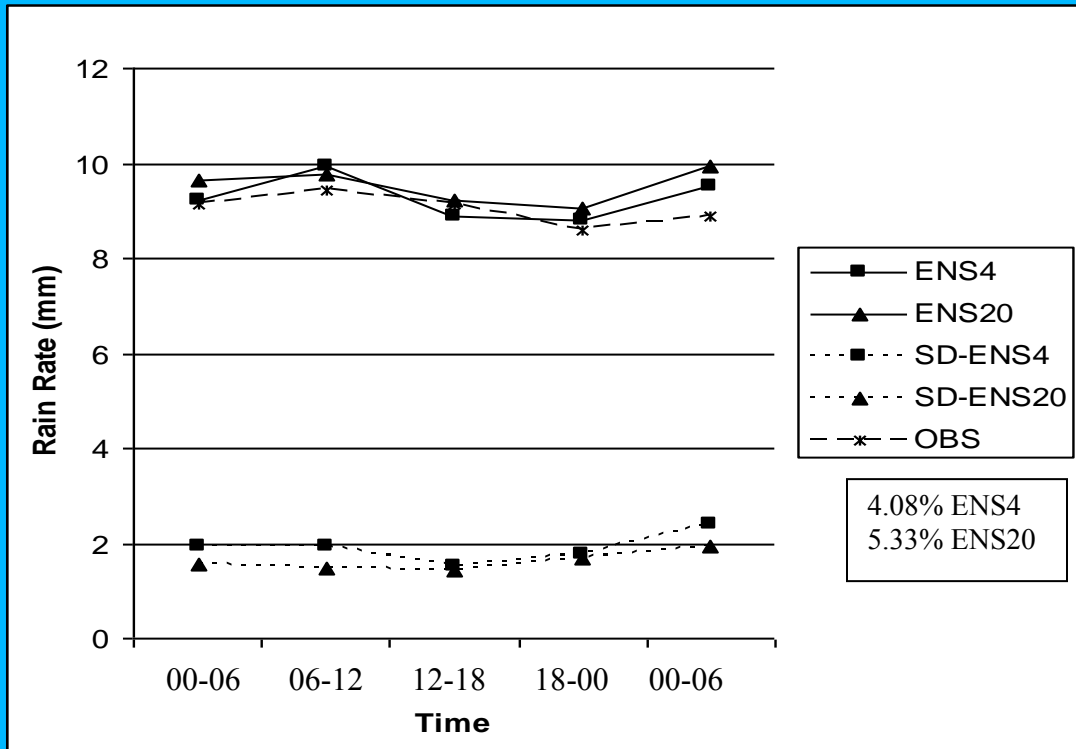


Figure 6: As in Fig. 5, except for rain rate (mm).

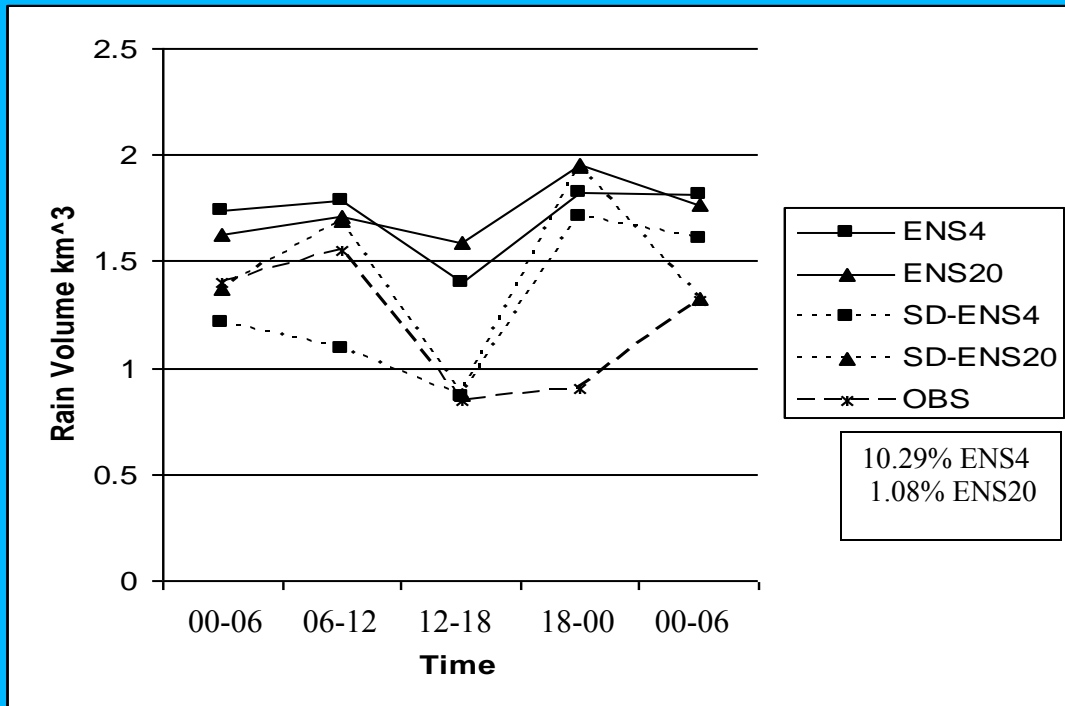


Figure 7: As in Fig. 5, except for rain volume in km^3 .

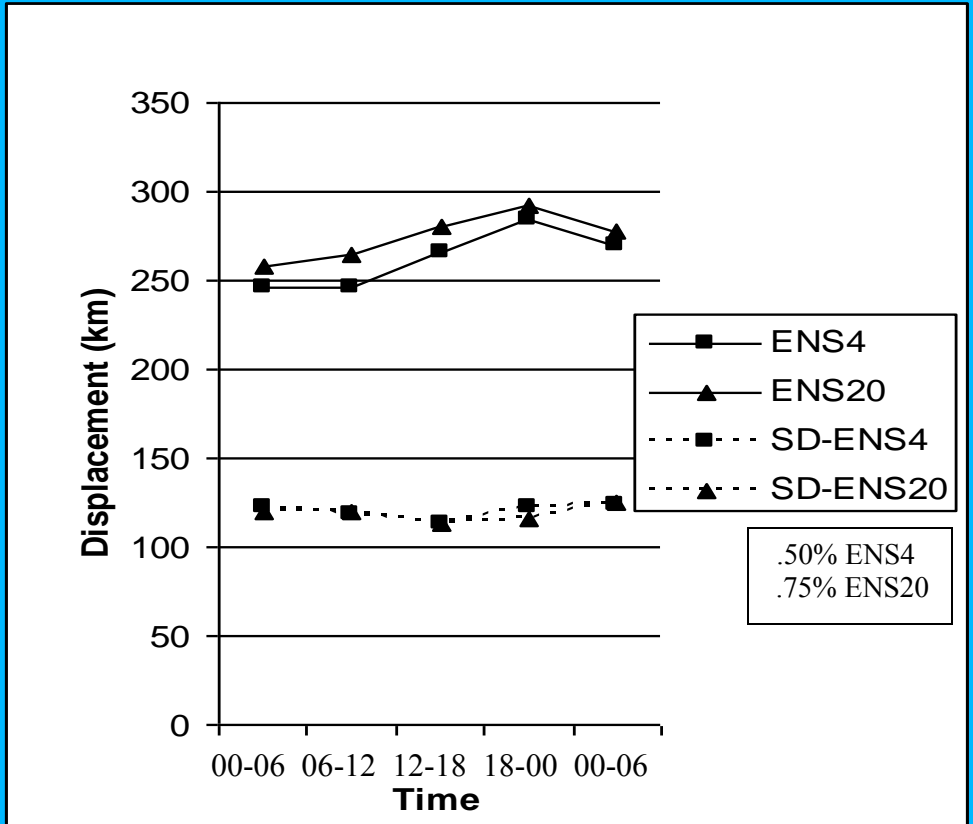


Figure 8: As in Fig. 5, except for displacement in km, with no observations plotted.

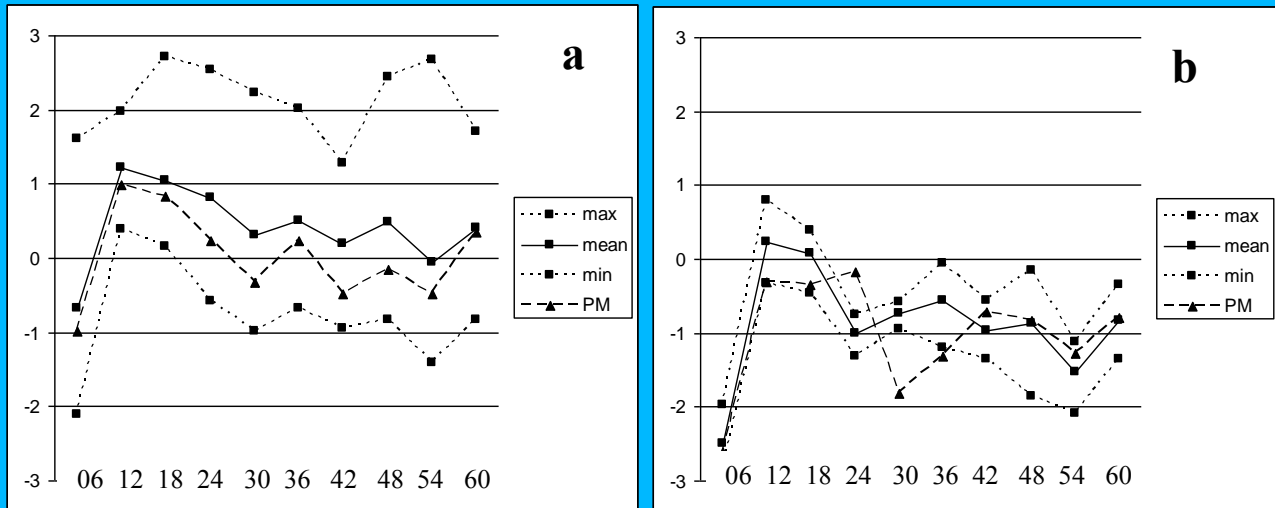


Figure 9: Difference between the forecasted rain rate and the observed (mm) as a function of time, based on CRA output, for a) Phys and b) IC/LBC. Curves are plotted to show the maximum and minimum values by any ensemble member (dotted lines with squares), the mean value (solid lines with squares), and the PM forecast (dashed lines with triangles).

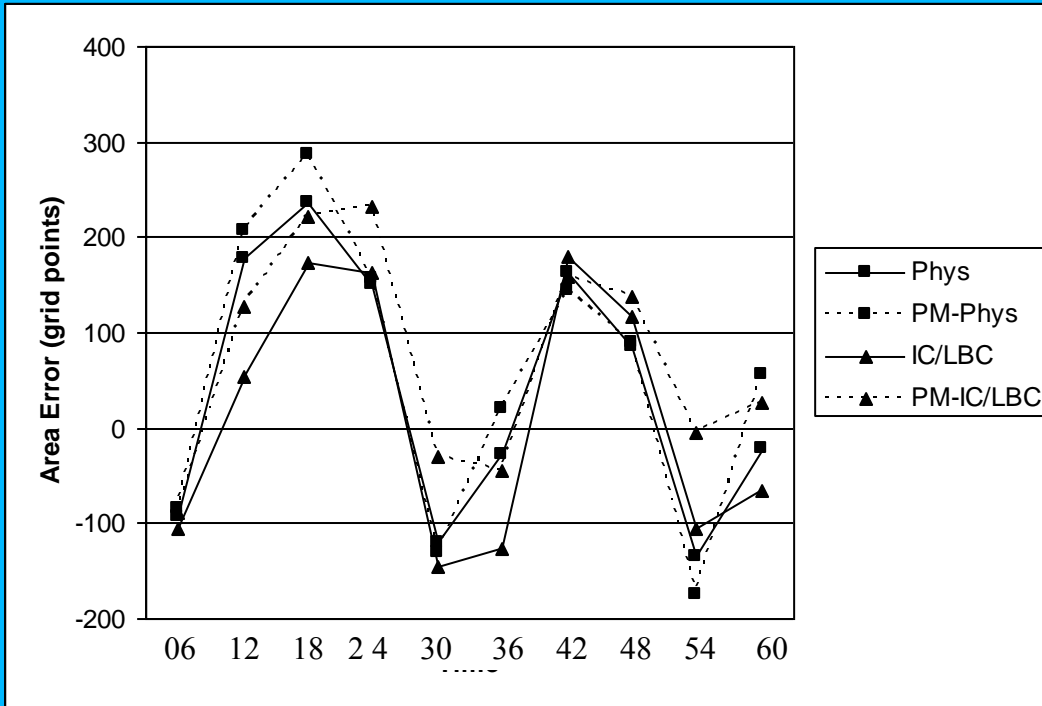


Figure 10: Difference between the forecasted rain area and the observed (10 x 10 km grid boxes) as a function of time, based on CRA output for the mean value (solid lines with squares for Phys and triangles for IC/LBC), and the PM forecast (dotted lines with squares for Phys, triangles for IC/LBC).

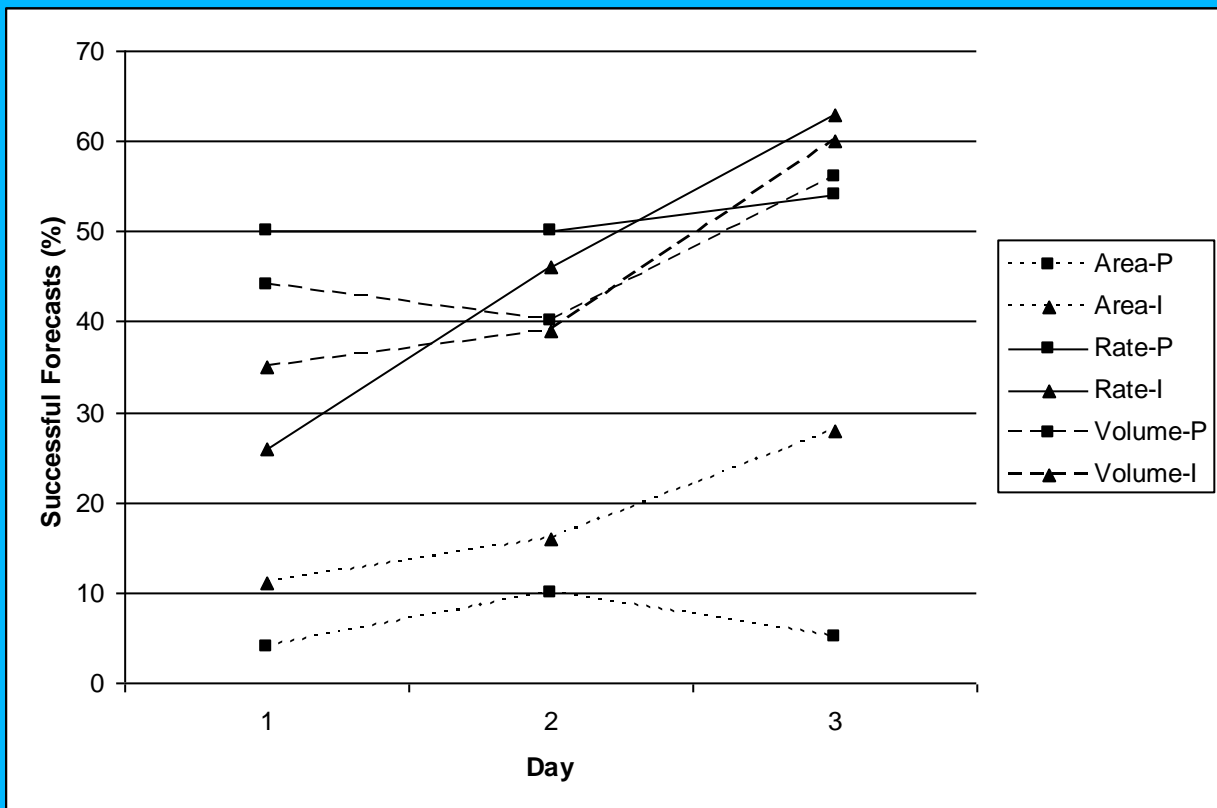


Figure 11: Percentage of time that observed values of rain volume (dashed line with squares for Phys, triangles for IC/LBC), rate (solid lines with squares for Phys, triangles for IC/LBC), and areal coverage (dotted lines with squares for Phys, triangles for IC/LBC) fell within the minimum and maximum predicted by the two ensembles, based on CRA output. 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.

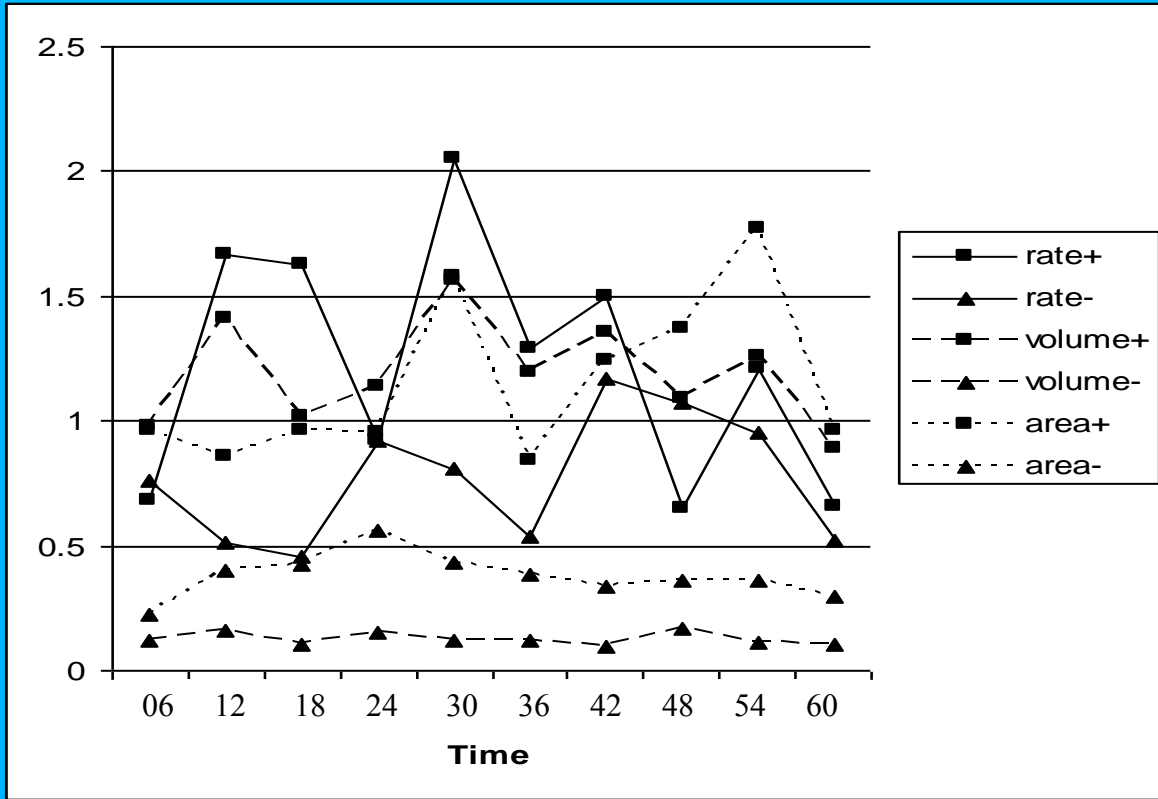


Figure 12: Mean absolute error for volume (km^3), rate ($*2.54 \text{ mm}$) and areal coverage ($*1000$ grid boxes) of precipitation from CRA applied to Phys as a function of time for cases with standard deviations over 150% of the mean (squares with solid line for rate, dashed line for volume, and dotted line for area), and less than 50% of the mean (triangles with solid line for rate, dashed line for volume, and dotted line for area).

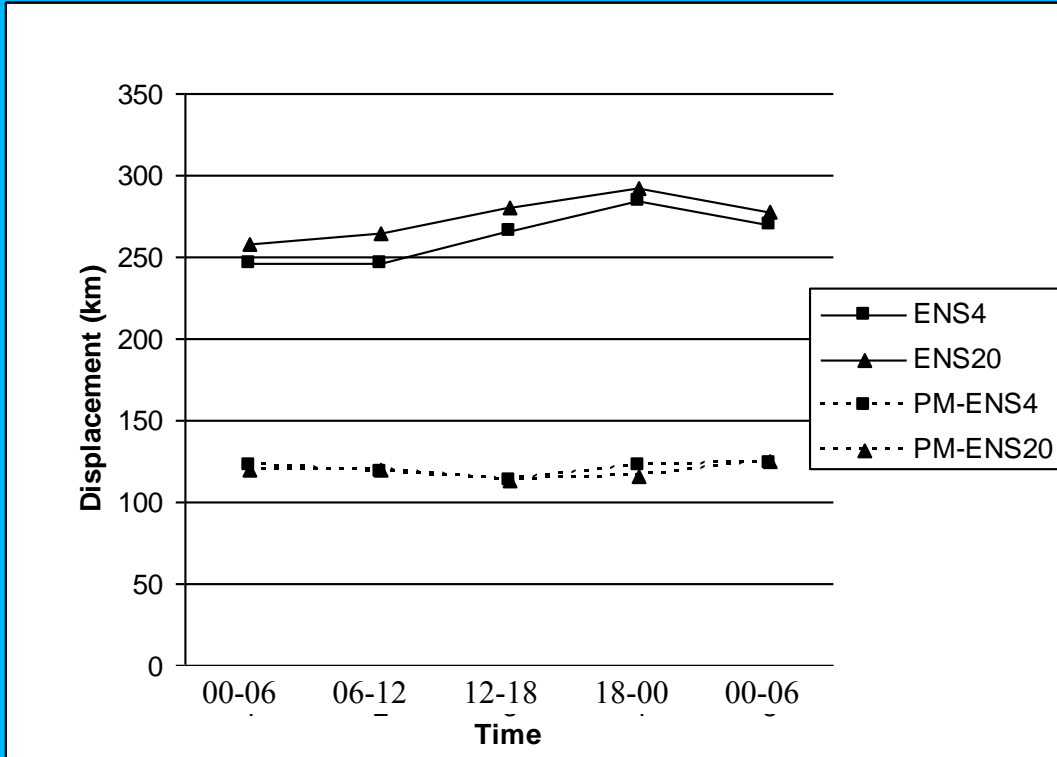


Figure 13: Displacement error (km) for the average of 4km ensemble members (solid line with squares) and the PM ensemble mean (dotted with squares), and for the average of the 20 km ensemble members (solid with triangles) and the 20 km PM ensemble mean (dotted with triangles), based on MODE output.

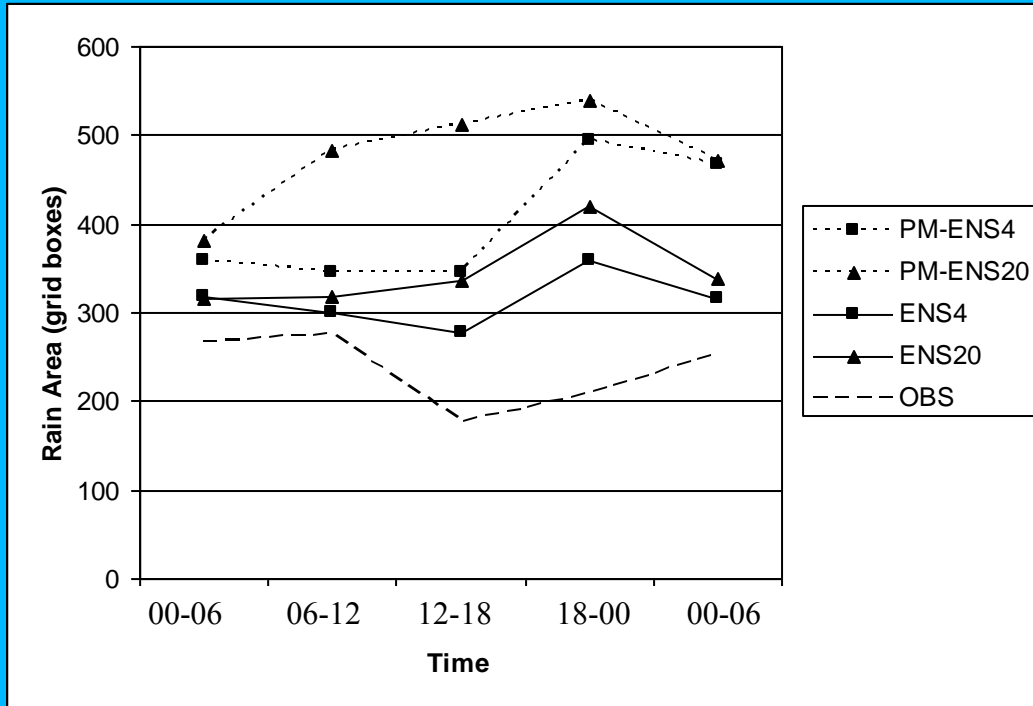


Figure 14: As in Fig. 13 except for rain area (20 x 20 km grid boxes) with observations also shown (dashed line).

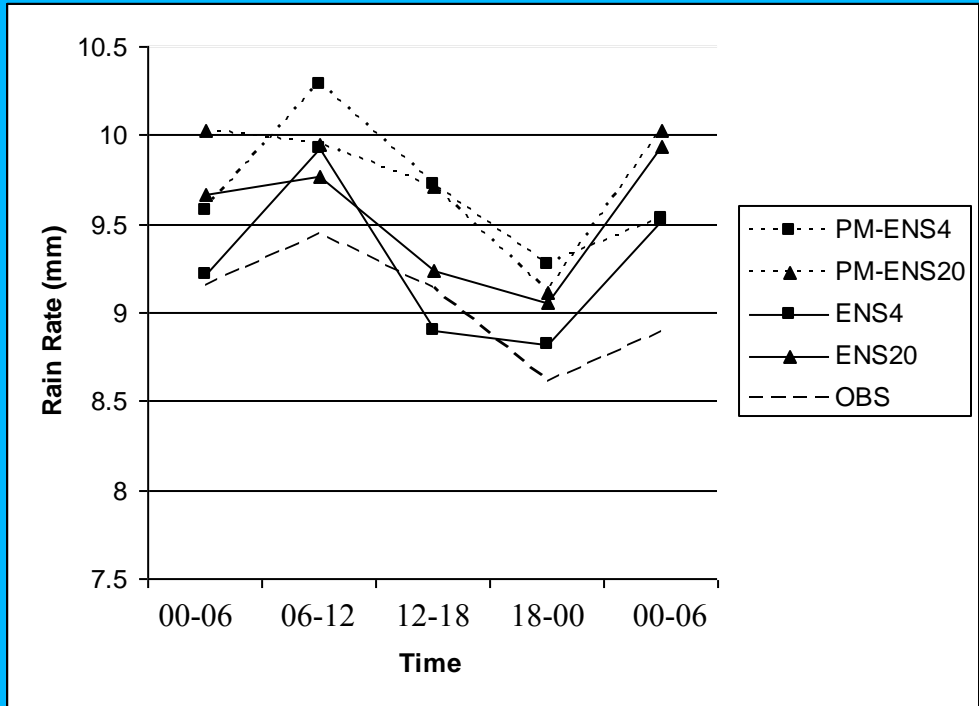


Figure 15: As in Fig. 14 except for rain rate (mm).

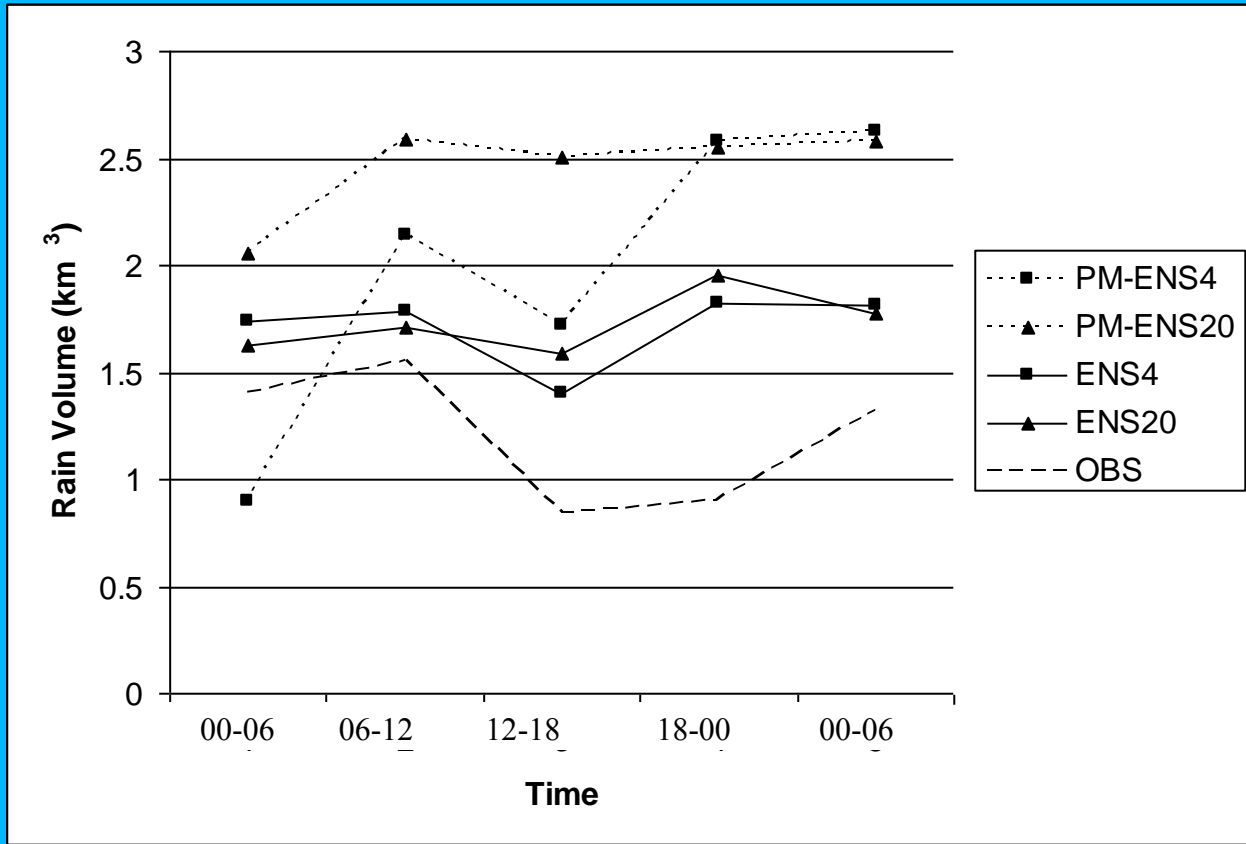


Figure 16: As in Fig. 14 except for rain volume (in km³).

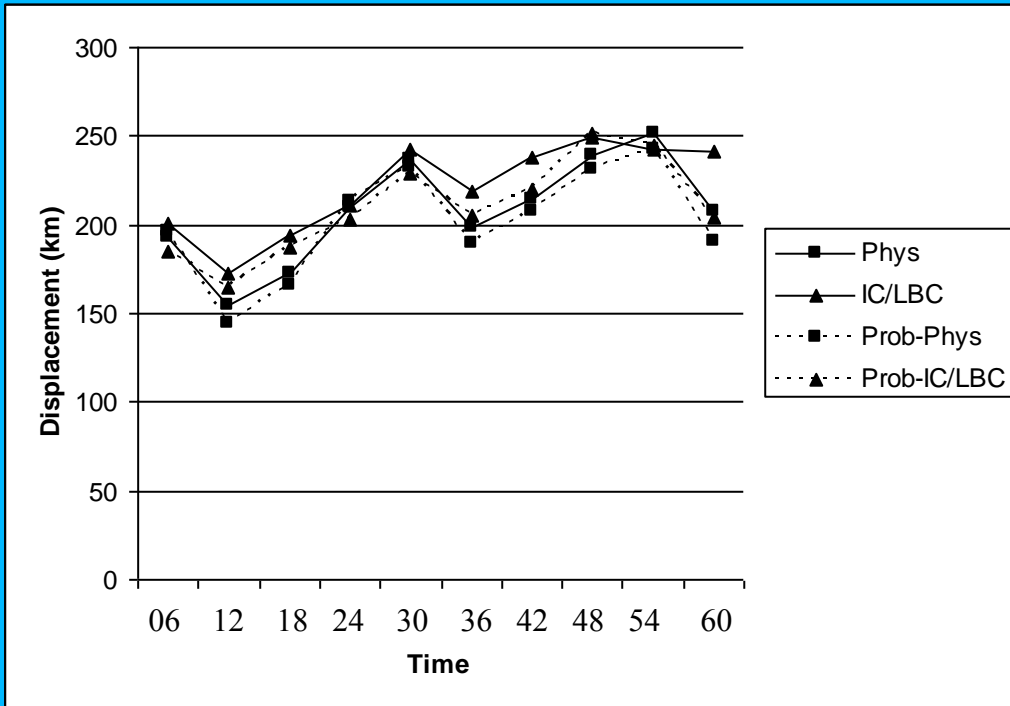


Figure 17: Comparison of displacement errors (km) by time between the 2 ensembles using the average of the MODE output applied to QPF amount for all individual members (solid lines with squares for Phys and triangles for IC/LBC) and using MODE applied to probability forecasts with a 30% threshold (dotted lines with squares for Phys and triangles for IC/LBC).

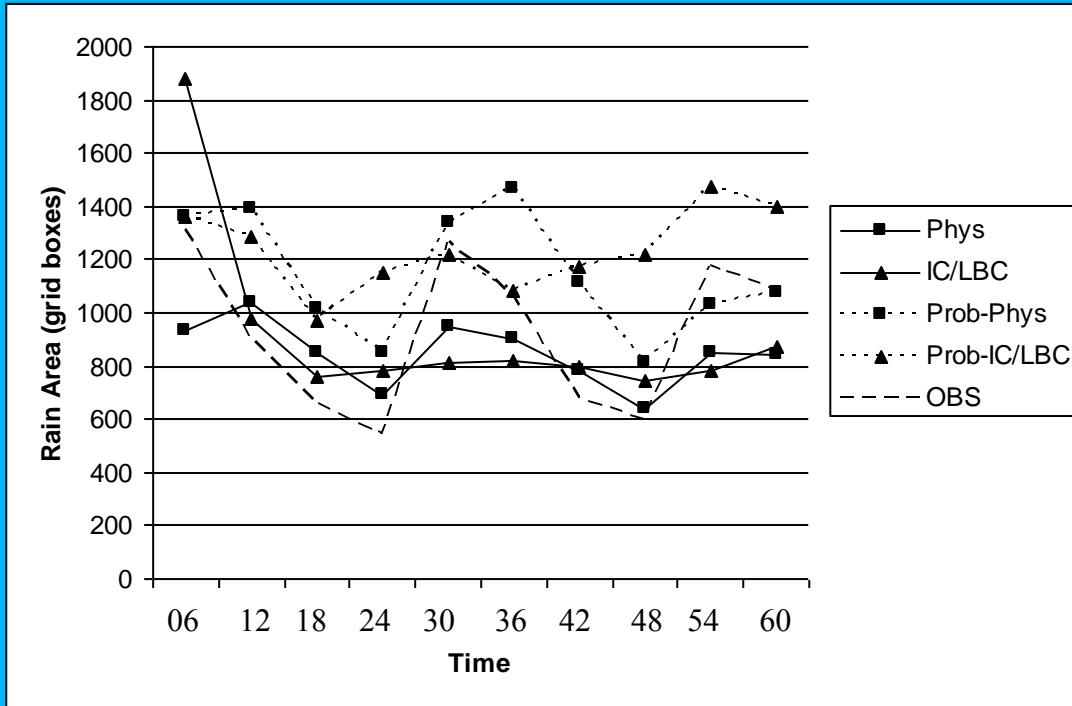


Figure 18: As in Fig. 17 except for rain area (in 10 x 10 km grid boxes), and with the observed areas shown with dashed line.