J11.3 On the Challenges of Identifying the "Best" Ensemble Member in Operational Forecasting

David R. Bright *
NOAA/NWS Storm Prediction Center, Norman, OK

Paul A. Nutter
CIMMS/University of Oklahoma, Norman, OK

1. INTRODUCTION

Operational weather forecasters have an abundance of numerical weather prediction (NWP) models that can be used in the forecast process. The Environmental Modeling Center (EMC), part of the National Weather Service's (NWS) National Centers for Environmental Prediction (NCEP), provides multiple runs daily of a variety of NWP models [e.g., Rapid Update Cycle (RUC), Nested Grid Model (NGM), Eta, Non-Hydrostatic Mesoscale Model (NMM), and the Global Forecast System (GFS)]. Short- and medium-range ensemble prediction systems are also supported by NCEP. Additionally, many universities and/or NWS field offices now produce their own real-time NWP forecast over limited area domains. The abundance of NWP output has led to a multimodel approach, such that forecasters implicitly create a "poor person's ensemble" by analyzing output from several models. Experienced forecasters use knowledge of model physics and biases to accept or dismiss certain aspects of a NWP solution, thereby weighting the NWP guidance and in effect creating a modified consensus forecast.

Ensembles continue to gain popularity in the operational environment and in many ways represent a formal extension of the familiar multimodel approach. But, time does not allow a forecaster to treat a large ensemble as individual, deterministic NWP forecasts. Considerable information is gleaned from the mean, spread, and probabilistic properties of the ensemble, but this approach may not be completely satisfying to meteorologists accustomed to viewing deterministic output. The reasons for some dissatisfaction in the ensemble approach is because the ensemble mean is dynamically inconsistent and generally smoothes details relevant to the mesoscale forecast, the ensemble spread is often under-dispersive, and probabilistic guidance may not translate directly to increased meteorological insight.

During the 2003 Storm Prediction Center/National Severe Storms Laboratory (SPC/NSSL) Spring Program (Levit et al. 2004), the NCEP Short-Range Ensemble Forecast (SREF) was used to determine if ensembles could aid in the prediction of severe convective weather. During the program, the question arose as to whether any SREF members could be eliminated from further consideration if they were not shadowing incoming observations during early portions of the run. This paper examines the concept of "thinning" SREF members from the ensemble if they perform poorly early in the run, and if early "best" members continue to verify best during the remainder of the forecast period.

* *Corresponding author address:* David R. Bright, Storm Prediction Center, 1313 Halley Circle, Norman, OK 73069; email: david.bright@noaa.gov

2. THEORETICAL RESULTS

The three variable Lorenz (1963) model may be used to illustrate simply why attempts to select the best member from an ensemble of *perfect model* forecasts do not yield productive guidance.

2.1: Flow Dependence and "Return to Skill"

An ensemble forecast system is comprised of a set of many individual deterministic forecasts. In each of those forecasts, errors generally increase with time up to a limit determined by the climate of the system. The error growth rate is different for each individual forecast and may even become negative so that errors decrease with time. Consider now the question of whether a single deterministic forecast can provide useful guidance once it has already experienced a loss of skill. We call this effect "return to skill" and demonstrate that it is equivalent to flow dependence.

Fig. 1 reveals the trajectories of two independent simulations having rather similar starting positions. The initial proximity of the simulations is shown by the small errors at time t=0 in the upper right panel. By following the three dimensional trajectories of the simulations, it is clear that the red trajectory makes two loops around each half of the attractor while the black trajectory has only one loop on the positive 'x' side of the attractor and three loops around the negative 'x' side.

The total error drops to nearly zero after about 3.75 units of time. Hence, there has been a return to skill, but it is not physically useful since the two simulations arrived at this point by following completely different routes through phase space. To use an adage, the simulations are, "right for the wrong reason." Hence, return to skill is simply an artifact of flow dependence, and could negatively impact forecast guidance in large dimensional weather prediction models.

Another feature seen in the error scores of Fig. 1 is the danger of comparing forecasts using only one variable. The difference between values of z is often rather small compared to differences between values of x and y. Thus, if the difference between the simulations is measured using only the z-component, the detrimental impact of the other variables on the overall quality of the forecast is ignored. This could be misleading since large differences between the x and y values form as the two simulations track through completely different parts of the phase space. Similar artifacts will be found in weather forecast models, especially when the number of variables we can evaluate is small compared to the enormous degrees of freedom.

2.2: Best Member Selection

The theoretical model is now used to illustrate how rapidly the "best" ensemble member changes. It may remain tempting to try and select a best member, as an ensemble is merely a collection of deterministic forecasts and the ensemble mean may not be attractive since it is dynamically inconsistent and often smoothes out mesoscale features that are of operational interest. Nonetheless, attempting to extract the single best member and apply it as a deterministic forecast is not productive. (Note that we are employing the fundamental assumption that each ensemble member is equally likely.)

Results from running 1000 independent ensemble simulations with the Lorenz model (Fig. 2a) show the ensemble mean square error (MSE) converges toward twice the climatological error variance. This result is expected (Leith 1974) since the model is unbiased. The deterministic forecast is usually considered to have lost skill when the ensemble MSE exceeds the climate error variance, in this case after about 6 time units. The results also show that the sum of ensemble mean error and ensemble dispersion (spread) is equal to the ensemble MSE. All three statistics converge to theoretically anticipated values after about 9 time units. The statistical curves have retained "steps" as the trajectories move from one side of the attractor to the other.

The solid black curve in Fig. 2a is a count of how many ensemble members have uniquely had the smallest difference from the control simulation. To use a sports or racing analogy, it is a measure of the number of unique "leaders." The number of ensemble members limits the maximum value. We count only unique best members rather than the total number of "lead changes" because the latter would improperly result from flow dependence and "return to skill" as discussed above. Fig. 2a shows that in a perfect ensemble system, the number of unique best members grows linearly with time for about the first 0.5 time units, and then increases at an inverse exponential rate. About half the ensemble members could have been considered best within 1.5 time units, or about a quarter of the time in which the overall ensemble has lost skill.

Operational forecast models always contain biases and other model system errors. To emulate this effect, a small constant forcing term has been added to the x- and y-variables of the Lorenz system (Palmer 1995). Three different constants were used for 20-member subsets of the full 60-member ensemble to emulate the use of a multi-model ensemble. The additional forcing terms cause the solution trajectories to loop a few extra times around one side of the attractor before switching back to the other side. This behavior is seen in the error statistics as skill is lost within 1/3 of the time of the unforced model and ensemble mean error (and hence, the ensemble MSE) exceeds the climate error variance before converging to the saturation value.

While increasing the rate of error growth and spread, the additional model forcing actually *slows* the growth rate for the number of unique best ensemble members. The most probable reason for this can be explained by the enhanced residence time for solution trajectories around each attractor. Indeed, the control solution may switch from one side of the attractor and back in

the same time that a forced solution remains on just one side. Consequently, an individual ensemble member may have the best match to the control solution twice during that period, and as such does not contribute to the count of unique best members.

Nearly the same results were obtained for other model variables, and also when choosing the ensemble member that best matches the ensemble mean rather then the control simulation.

3. OPERATIONAL RESULTS: NCEP SREF

The NCEP SREF is comprised of 15 members and runs routinely at 09 UTC and 21 UTC through 63 forecast hours. At the time of this writing, the system employed three models: the Eta with Betts-Miller-Janjic parameterized convection; the Eta with Kain-Fritsch parameterized convection; and the Regional Spectral Model (RSM). Each model configuration contains five members including an unperturbed initial condition and four perturbed initial conditions created through breeding of growing modes. [More information on the NCEP SREF is available at the following URL: http://wwwt.emc.ncep.noaa.gov/mmb/SREF/SREF.html and from Du and Tracton (2001).] The NCEP SREF is examined herein to test whether an ensemble member exhibiting the best forecast early in the run is likely to retain that position at later times.

The 09 UTC and 21 UTC SREF forecasts were archived for 24 days during August 2003. Verification was based on the RUC analyses at 00 UTC and 12 UTC. The data were compared on the AWIPS regional grid 236 (see http://www.nco.ncep.noaa.gov/pmb/doc/on388/tableb.html for grid specifications) which has an approximate grid spacing of 40 km. Following the ideas of Roulston and Smith (2003), a root mean squared error (RMSE) based on 22 normalized variables was used to determine the best member (Table 1). Each variable was normalized by its standard deviation across the domain, and then all 22 variables were summed to arrive at the normalized RMSE (NRMSE). In order to avoid problems associated RUC lateral boundary conditions, calculations were restricted to a region over the central United States (Fig. 3). Comparisons to the RUC analyses were made at 00 UTC and 12 UTC corresponding to forecast hours of F03, F15, F27, F39, F51, and F63.

Table 1. A listing of the 22 normalized variables used to determine the best ensemble member.

LEVEL	VARIABLE
Surface:	Sea Level Pressure
Tropospheric:	Precipitable Water; CAPE
2 meter:	Temperature; Dew Point
10 meter:	U; V
700, 500, 300 hPa:	Temperature; Mixing Ratio; U; V; Geopotential Height

3.1 Time Averaged Results

Based on the NRSME for 24 days in August 2003, the ensemble mean is easily the best forecast for all times through 63 hours (figure not shown). Because the NRMSE continues to grow through the entire 63-hour forecast, and comparisons are made only at 00 UTC and 12 UTC, not all members have an "opportunity" to become the best member. Thus, a figure exactly like Fig. 2 is impossible to construct, and results are presented as a percentage of possible unique best members.

A simple mathematical calculation shows that the ensemble mean should always have the smallest error (Toth and Kalnay 1997), and therefore dominates the scoring of the best member (Fig. 4 – dashed line). However, if the ensemble mean is excluded, then the percentage of possible unique best members increases steadily with time (Fig. 4 – solid line), reaching about 45% by F63. In other words, with six 12-hourly verification steps indicated in Fig. 4, the NRMSE indicates that about three members are uniquely considered best out of the six that could have been identified during the forecast. These results do not consider whether or not the differences in NRMSE are statistically significant.

Results shown in Fig. 2b suggested that model biases may decrease the count of unique best members by slowing the forecast response to error growth. This may partly explain why only half of the available SREF ensemble members were considered best during the forecast period. Furthermore, repeat (non-unique) best members were not scored as a "lead change" because that would indicate a useless return to skill. On the other hand, we have not attempted to determine error growth rates or the expected time limit of SREF skill. If the forecasts remained skillful at the end of 63 hours, then we should not expect all of the available ensemble members to be uniquely counted as the best.

The ranking of the NRMSE scores were used to calculate the correlation coefficient

between the ensemble member rank at some forecast time and the rank at a later forecast time. Figure 5 shows the correlation of the ensemble membership rank, both with (dashed) and without (solid) the ensemble mean, to its rank 12 hours earlier. The correlation appears to increase gradually during the forecast, and inclusion of the ensemble mean always improves the result. Looking at longer lead times, the correlation of the ensemble rank at F15 to the ensemble rank 24, 36, and 48 hours later (F39, F51, and F63, respectively) decreases from about 0.5 to 0.3 (Fig. 6). Thus, it appears the ensemble rank early in the forecast has a rather low correlation to the ensemble rank one to two days later. *This indicates that an ensemble member should not be isolated as a preferred deterministic forecast for the remainder of the forecast, nor should remaining members be dismissed from further analysis.*

3.2 Individual Examples

Many members of an ensemble contribute useful information to the forecast, and that contribution appears to be a nonlinear function of location, time, and variable. For example, Fig. 7 shows the F15 NCEP SREF member (including the ensemble mean) that most closely matches the RUC analysis of 2-meter dew point (F15 SREF forecast valid at 00 UTC 28 May 2003). The number posted indicates the ensemble member closest to the RUC analysis at each grid point [Black 0=ensemble mean; Red=Eta-BMJ (members 1-5); Yellow=Eta-KF (members 6-10); Blue=RSM (members 11-15). Numbers 1, 6, and 11 represent the control members of the Eta-BMJ, Eta-KF, and RSM, respectively.] Only grid points with a dew point temperature greater than 50 F are plotted; it is clear there is little spatial correlation between ensemble members, and no one member dominates a large region of the domain.

In a second example, a NRSME multivariable calculation (as described in section 3.1) was made over a portion of the Northern Plains to determine if any SREF member had a better forecast of a short wave trough moving through the area on 2 June 2003 (Fig. 8). A scatterplot of F15 ensemble rank to F39 ensemble rank indicates a very low correlation (Fig. 9 – correlation coefficient only 0.28). The ensemble mean is ranked first at both F15 and F39, while an Eta member ranked 11th at F15 becomes the best individual member at F39.

Because the ensemble mean is usually the best member, one could attempt to locate the member closest to the mean. However, applying this approach to the 09 UTC 3 November 2003 NCEP SREF for 500 hPa geopotential height results in a collage of ensemble members and a dynamically inconsistent, noisy analysis (Fig. 10).

4. SUMMARY

While operational forecasters can extract useful information through detailed examination of individual NWP solutions, a comparison of observations with individual members in an attempt to extract a single best ensemble member will not yield the best forecast over time. Both the simple theoretical modeling approach and analysis of the NCEP SREF support this finding. Similarly, thinning an ensemble by eliminating the "worst" members early in the run degrades the future value of the ensemble. Both of these findings imply that a "model of the day" concept, where a single NWP solution is chosen upon which to base an entire forecast, is fundamentally flawed.

The theoretical results show that if models are perfect and unbiased, then forecasters could only evaluate the statistical aspects of the ensemble. However, our NWP models are far from perfect, and experienced forecasters can use knowledge of model physics and biases to add value to the ensemble by dismissing certain aspects of the solution due to a misrepresentation of simulated atmospheric processes (e.g., Baldwin et al. 2002).

Despite the model error contained within the NCEP SREF, there are some interesting similarities to the theoretical results. Most importantly, the number of unique best members derived from a multiparameter error calculation increases steadily with time through the 63-hr run. Thus, despite the existence of known model error, attempting to choose or eliminate members may degrade the future value of the SREF because "bad" members may appear as the best member at a later time. Experienced forecasters who understand the physics and biases of the component models can improve the ensemble forecast at a given forecast time by accounting for (either statistically or conceptually) known model error; nonetheless, all ensemble members should be retained for a complete analysis at other forecast hours.

Acknowledgements: Thanks to Steven J. Weiss, SPC Science and Operations Officer, for helpful

comments and discussion. The 2003 SPC/NSSL Spring Program provided the opportunity for many of these ideas to be discussed in the context of real-time severe weather forecasting. We appreciate the scientific dialogue fostered by the Spring Program, and remain grateful to COMET for helping to fund portions of the 2003 Spring Program through COMET Partner's Project S03-38671.

5. REFERENCES

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

Du, J., and S. Tracton, 2001: Implementation of a real-time short range ensemble forecasting system at NCEP: An update. Preprints, *Ninth Conf. On Mesoscale Processes*, Fort Lauderdale, FL, Amer. Meteor. Soc., 355-356.

Levit, J., D. Stensrud, D. Bright, and S. Weiss, 2004: Evaluation of short-range ensemble forecasts during the SPC/NSSL 2003 Spring Program. Preprints, *16th Conf. On Numerical Weather Prediction*, Seattle WA, Amer. Meteor. Soc.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409-418.

Lorenz, E. N., 1963: Deterministic Nonperiodic Flow. *J. Atmos. Sci.*, **20**, 130-141.

Palmer, T. N., 1995: Predictability of the atmosphere and oceans: From days to decades. *Seminar on Predictability, Vol. I.* Reading, United Kingdom, ECMWF, 83-141. [Available from European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading RG2 9AX, United Kingdom.]

Roulston, M., and L. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus*, **55A**, 16-30.

Toth, Z. and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev*, **125**, 3297-3319.

6. FIGURES

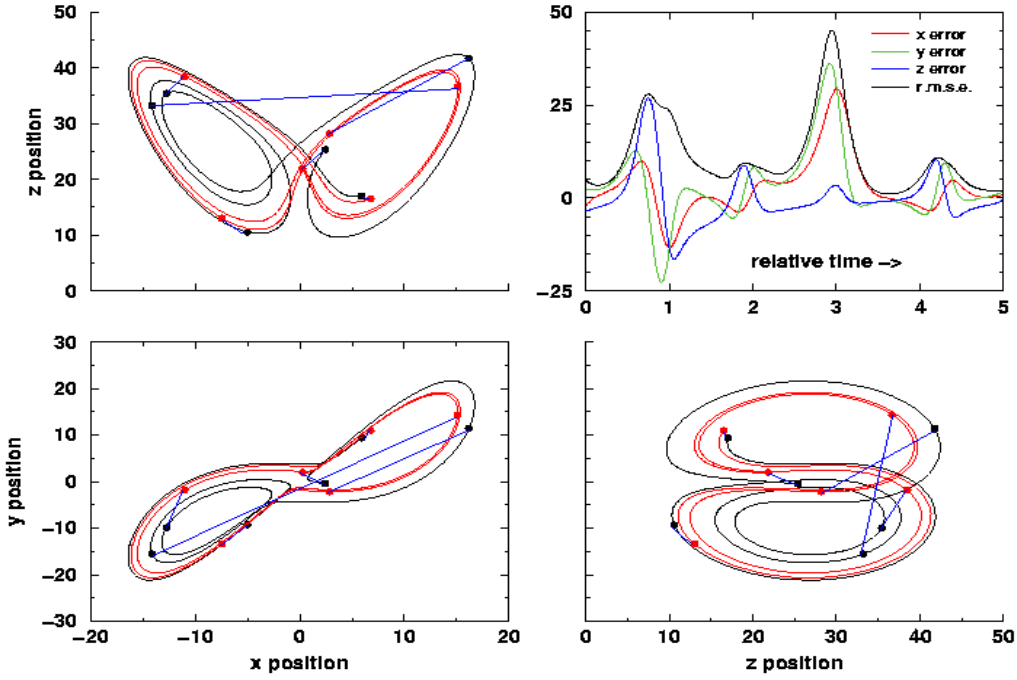

Fig. 1: Phase space trajectories of two independent Lorenz model simulations (red vs. black) originating at nearly the same position around (x=0,y=0,z=25). Dots connected by blue line segments identify solution values at regular time intervals. The upper right panel shows how the distance between the two trajectories changes with time for each model variable (x_2-x_1,y_2-y_1,z_2-z_1) and for the root mean square difference [$\sqrt{((x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2)}$].

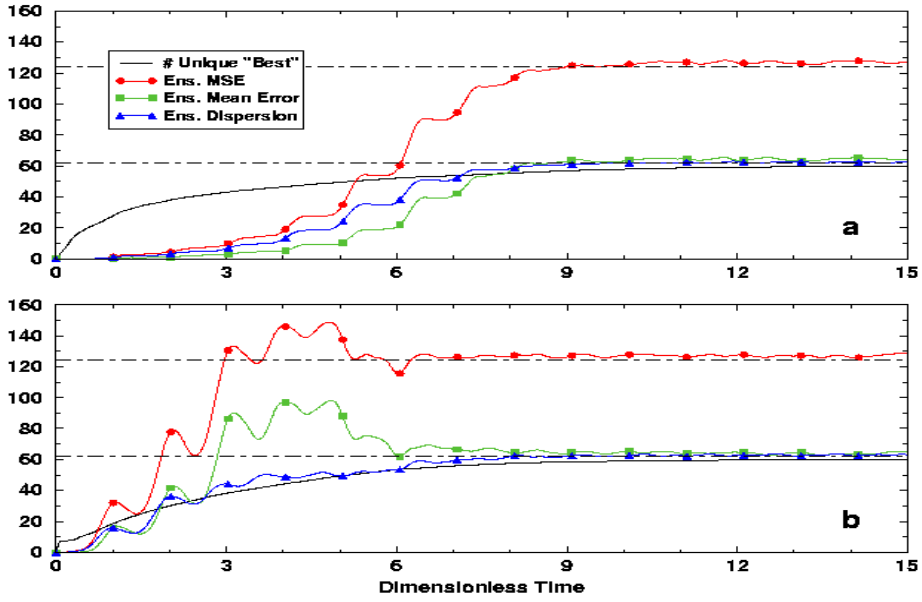

Fig. 2: Ensemble error statistics for the x-component of the Lorenz model obtained as averages over 1000 independent 60-member ensembles. Statistics from an unmodified (perfect) model are shown in (a) and those from a weakly forced version are shown in (b). The horizontal dashed lines show the values of 1 and 2 times the model's climate error variance calculated using random samples from the x-component of a long control simulation.

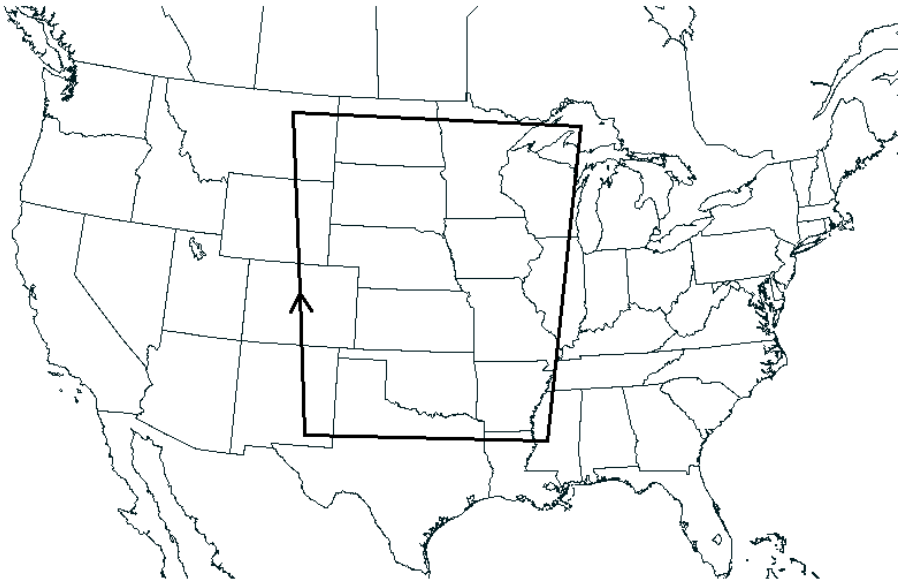

Fig. 3: The sub-domain over central U.S. considered in determining the best NCEP SREF member (inside black box).

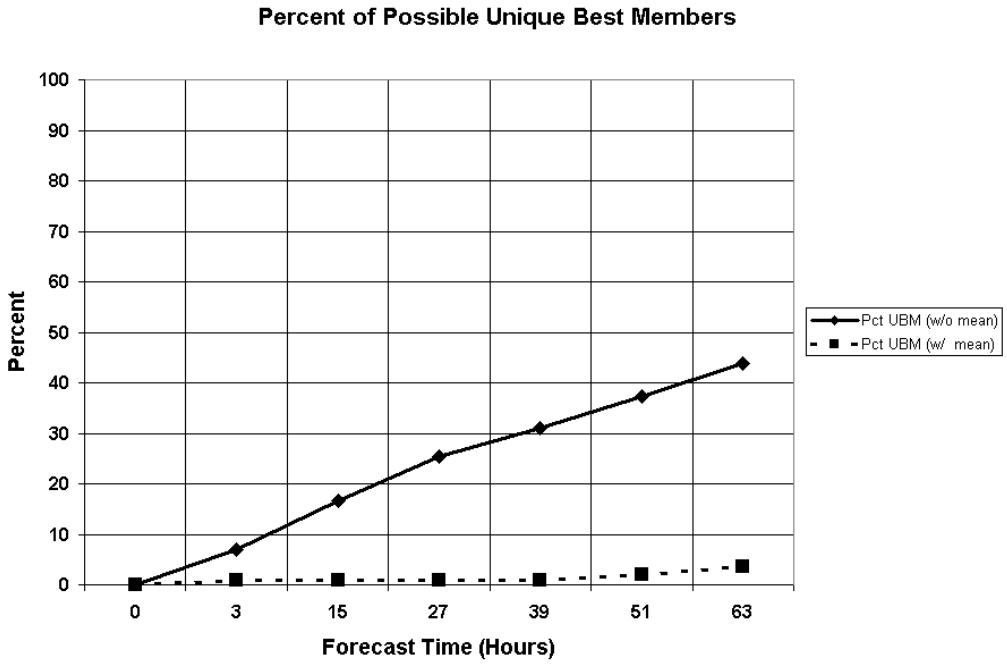

Fig. 4: The percent of possible unique NCEP SREF best members, both excluding the mean (solid) and including the mean (dashed).

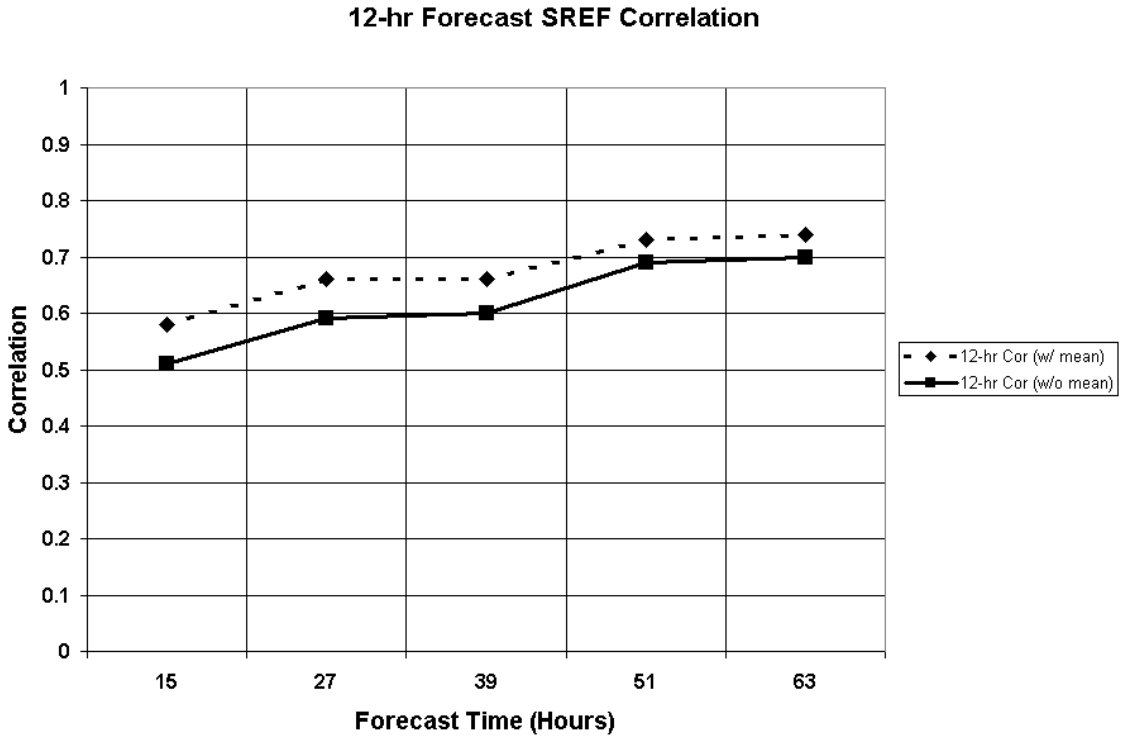

Fig. 5: The correlation coefficient of NCEP SREF member rank to the rank 12-hours earlier, both excluding (solid) and including (dashed) the ensemble mean.

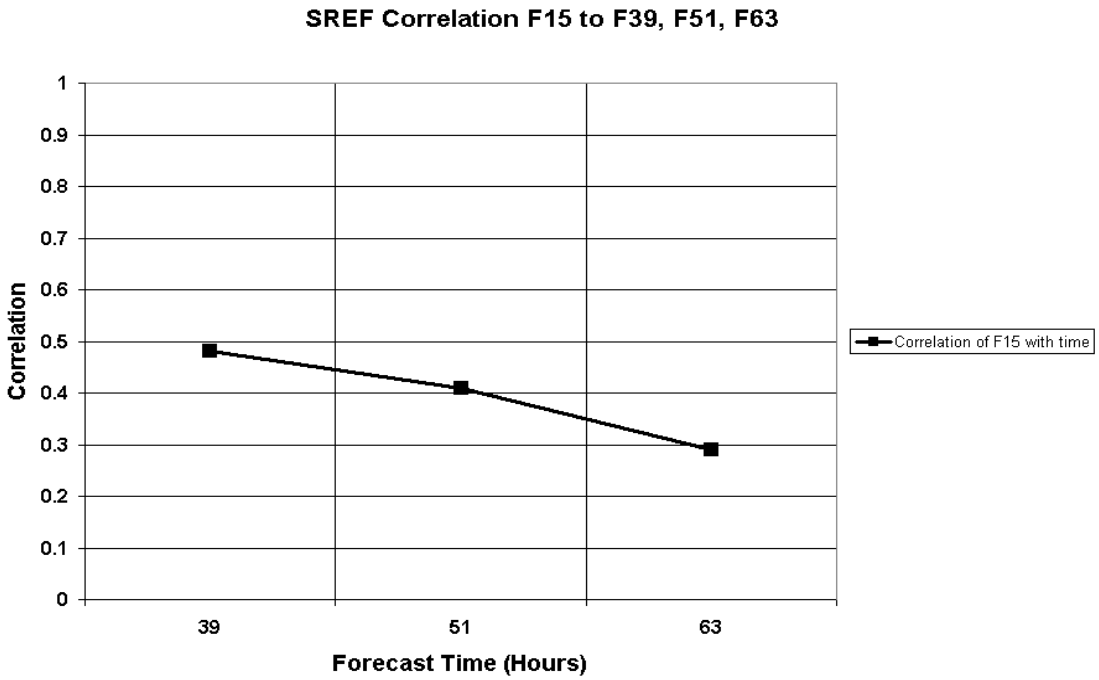

Fig. 6: The correlation coefficient of the NCEP SREF (excluding the ensemble mean) member rank at F15 to the rank at F39, F51, and F63.

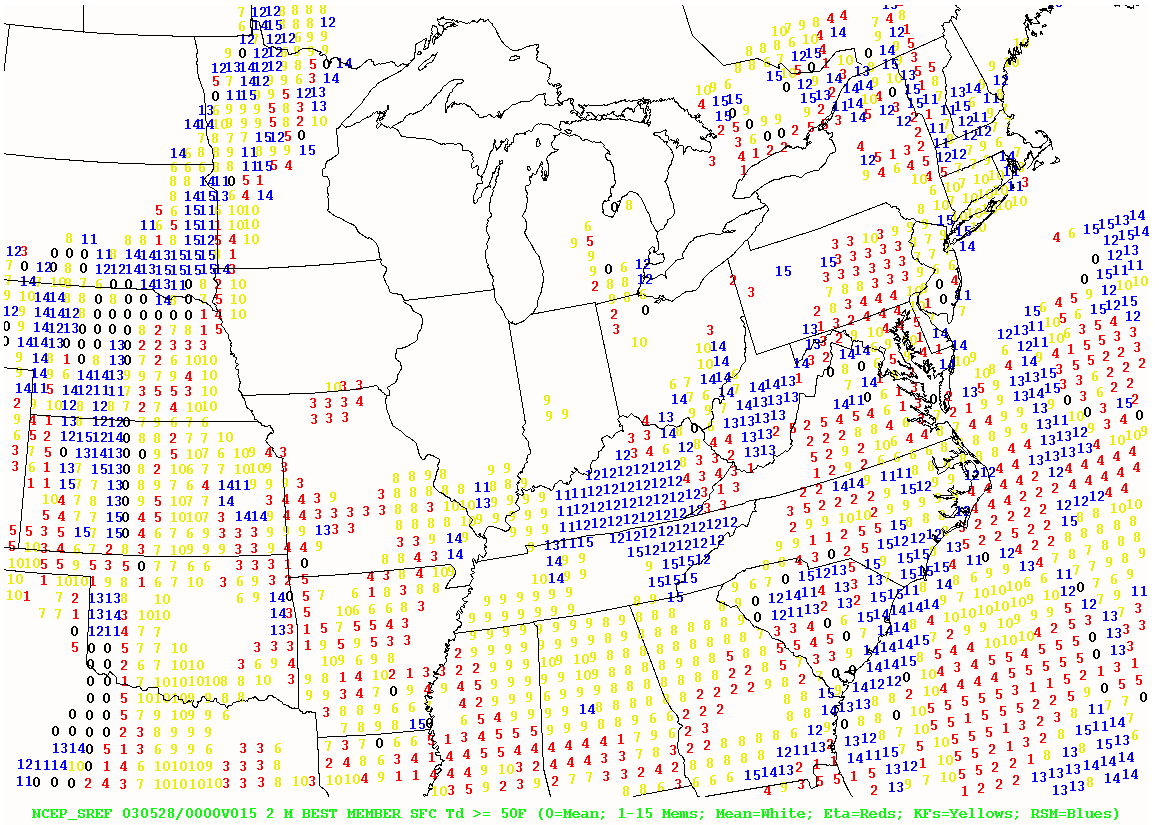

Fig. 7: The NCEP SREF ensemble member closest to the RUC analysis valid at 00 UTC 28 May 2003. (Black 0 indicates the ensemble mean; Yellow=Eta-BMJ members; Red=Eta-KF members; and Blue=RSM members)

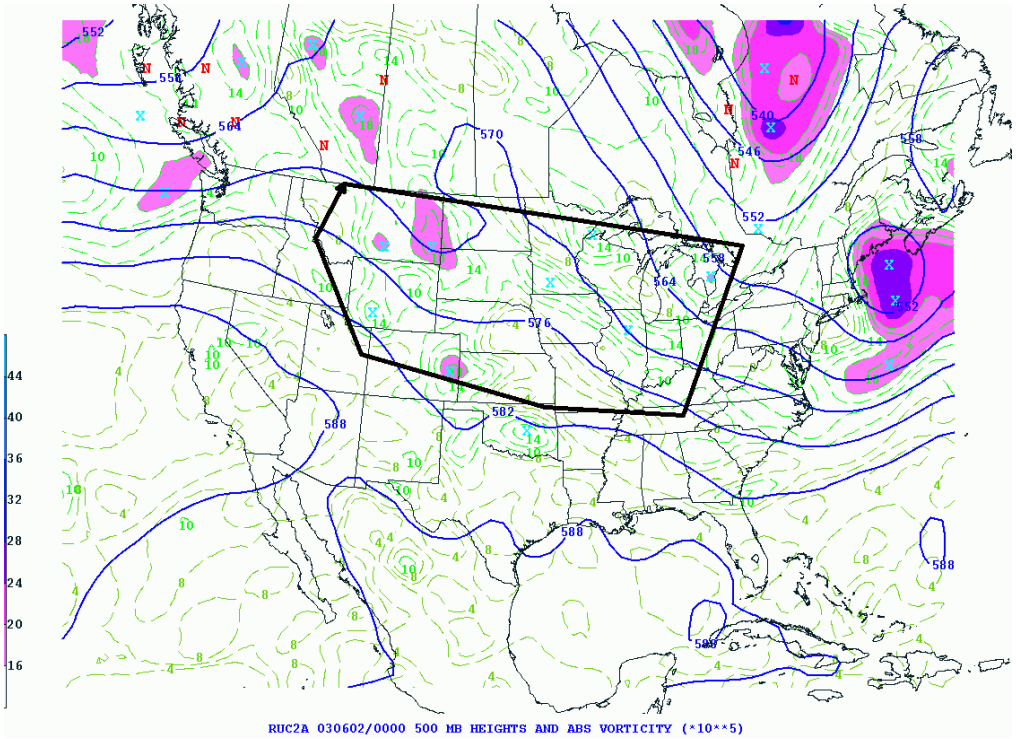

Fig. 8: Domain considered (black, solid line) for determining the best NCEP SREF ensemble member from the 09 UTC 01 June 2003 case. The 500 hPa geopotential height (solid) and absolute vorticity (dashed) are from the RUC analysis valid at 00 UTC 02 June 2003.

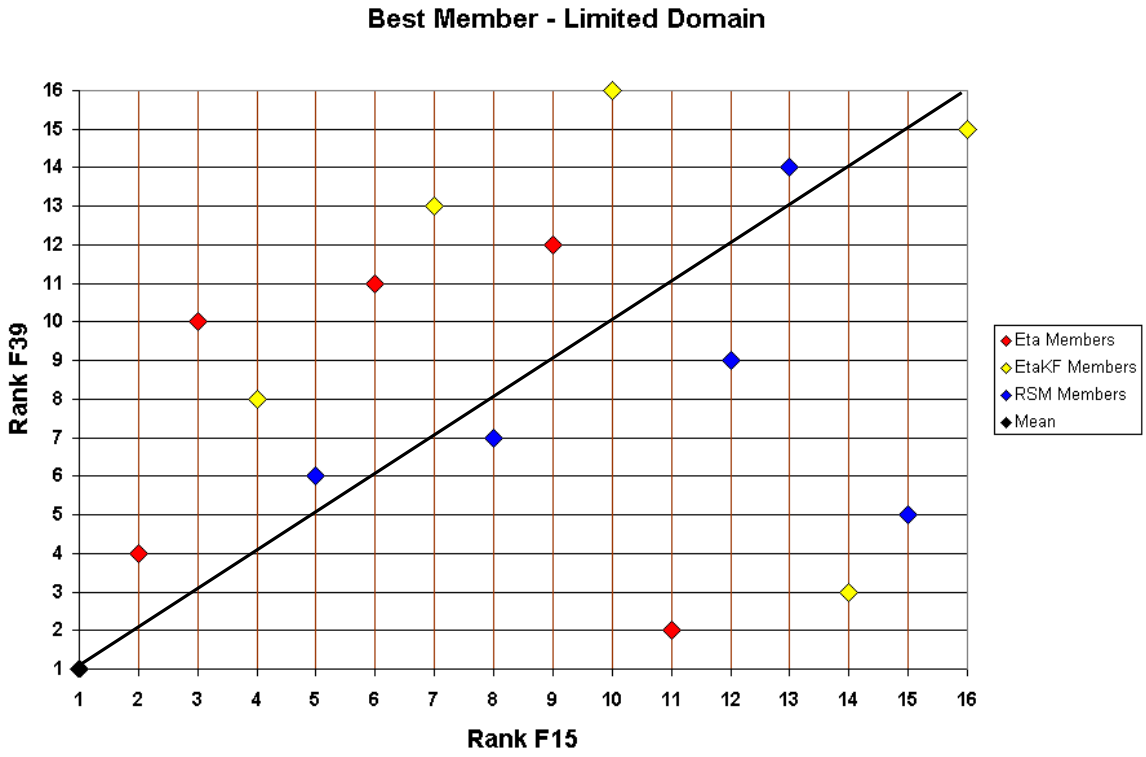

Fig. 9: Scatterplot of the NCEP SREF ensemble rank, including the ensemble mean (black), at F15 to F39 (Red=Eta-BMJ; Yellow=Eta-KF; Blue=RSM) from the 09 UTC 01 June 2003 NCEP SREF. (Correlation coefficient = 0.28.)

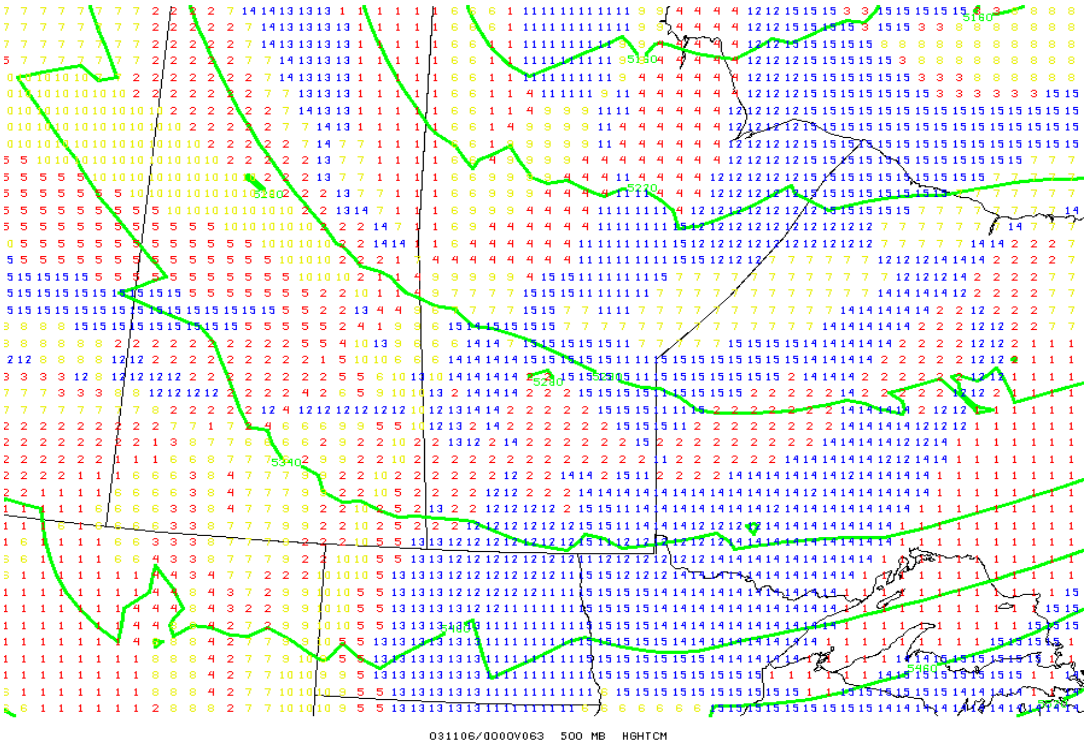

Fig. 10: As in Fig. 7, except the NCEP SREF member closest to the mean at 500 hPa (F63 valid 00 UTC 06 November 2003). The solid line is the 500 hPa geopotential height based on the member closest to the mean.