Western Region Technical Attachment
No. 95-29
November 21, 1995

INTRODUCTION TO ENSEMBLE FORECASTING

Jon Mittelstadt - WRH - SSD - Salt Lake City, UT

Introduction

In a 1963 paper, E. N. Lorenz demonstrated that even very small errors in a numerical model's initial conditions will lead to large errors with time. Experience with numerical weather models has shown that the rate at which small errors grow varies greatly from day to day and for different geographic regions. Ensemble forecasting refers to the process of introducing small perturbations to the initial conditions and examining their growth in order to determine the predictability of model forecasts. This process is the first step in examining the chaos of a dynamical system. The National Center for Environmental Prediction (NCEP) has been testing ensemble forecasting using the Medium Range Forecast (MRF) model since December 1992. This Technical Attachment describes the medium and extended-range ensemble forecasting technique of the NCEP and some of the products available to NWS forecasters. Hints are given on how to best use the available products. Some preliminary verification statistics on the skill of the NCEP ensemble are also included.

How the Ensemble is Created

An "ensemble" is a set of model solutions such that each solution, or "member", is initiated with a slightly different set of initial conditions. The different members are created by introducing small errors, called "perturbations" to the initial conditions of a "control forecast". Statistically, the ensemble mean should, over time, result in better skill than the individual members. In general, increasing the number of ensemble members will increase the skill of the ensemble forecasts. However, as will be discussed in the next section, if perturbations are chosen carefully, a relatively small number of members may be capable of capturing most of the probabilistic information.

The Creation of Perturbations at the NCEP

The loss of model forecast skill at medium to extended ranges is caused by two factors: (1) the differences between the model's analysis and reality and (2) by the approximations of the forecast model. In midlatitudes, the loss of model forecast skill at extended ranges is usually dominated by the analysis errors. Random errors do not grow as fast as analysis errors because they are not organized dynamically. Thus, in an operational setting, the perturbations are chosen to be as close to the size and distribution of the analysis errors as possible.

Another consideration is that only a small subset of all analysis errors will grow rapidly with time. For example, errors located near a baroclinic zone may grow much more rapidly than those located within a broad ridge. These rapidly growing errors are responsible for the loss of a model's predictive skill.

Therefore, if perturbations are chosen to be similar to the size and distribution of these rapidly growing errors, then the effectiveness of an ensemble will increase.

At the NCEP, a technique called the Breeding of Growing Modes (BGM) was developed to create perturbations that are similar to the rapidly growing analysis errors. As mentioned above, within the analysis cycle, errors grow faster in certain unstable locations around the hemisphere, depending primarily on the synoptic situation. In locations where the analysis depends heavily on the first guess (the model's previous forecast), these errors will "survive" the analysis process and are then likely to grow again within the next forecast. In other words, any error that grew in the previous forecast will have a greater chance of remaining in the latest analysis and then to grow in the next short-range forecast. These errors are called "growing errors". The BGM process attempts to mimic the growing errors in the analysis cycle.

The BGM process is described schematically by Fig. 1. Initially (Day n), random perturbations are added and subtracted from an analysis to form two new initial states. The two new (slightly different) initial states are then run forward for 24 hours (to Day n+1). Some of the perturbations within these two forecasts will have grown to be large. The two solutions (at Day n+1) are then subtracted from each other, resulting in a set of differences. Large differences will exist at the locations of rapidly growing perturbations. The differences are then scaled to the size of the original perturbations. The scaling varies by hemispheric location based on the typical size of analysis errors at that location. Thus, over data-sparse regions, the perturbations are typically three times larger than over data-rich regions. The scaled differences are now treated as a new perturbation field and the cycle is repeated, i.e., the second cycle is started using the scaled perturbations rather than the random perturbations. The cycle is repeated infinitely, or until a computer problem requires the entire process to be restarted.

After a few cycles, the perturbations reach their maximum growth rates and will (in theory) resemble the actual analysis errors. The perturbations are then ready to be used for extended-range ensemble forecasting. (Note: Each "breeding cycle" creates two members of the ensemble.)

It is important to keep in mind that a small number of ensemble members can not capture all of the growing errors and, therefore, ensembles only estimate predictability. Further, in some cases, model error may dominate the analysis errors, in which case, ensemble data may mislead a forecaster.

Current Configuration at the NCEP

The current 0000 UTC MRF ensemble has 17 members (see Table 1). All ensemble forecasts are run to 16 days. In order to increase the number of ensemble members within the constraint of available supercomputer resources, the NCEP runs several of the members using a lower resolution (T62 truncation) MRF model. T126 is roughly a 100-km resolution and T62 is roughly a 200-km resolution. Verification statistics indicate that on a hemispheric scale, the low resolution model is only slightly less skillful. However, some recent papers indicate that the loss of skill caused by coarser resolution may be important for situations when a large-scale trough is moving over the mountainous western states. There are five BGM cycles maintained at 0000 UTC and two maintained at 1200 UTC, creating 10 and 4 members of the ensemble, respectively.

Table 1 - THE NCEP ENSEMBLE

ANALYSIS TIME	MODEL	NUMBER OF MEMBERS	TYPE	RESOLUTION
0000 UTC	MRF	1	Operational	T126 to day 7 T62 days 7-16
0000 UTC	MRF	1	Low-resolution control	T62 days 1-16
0000 UTC	MRF	10	Perturbed by the BGM process	T62 days 1-16
1200 UTC	AVN	1	Operational	T126 to day 3 T62 day 3-16
1200 UTC	AVN	4	Perturbed by the BGM process	T62 days 1-16

A total of 10 members begin at 0000 UTC and a total of 5 members begin at 1200 UTC. The forecasts from 0000 UTC and 1200 UTC are combined, so that, for example, the 0000 UTC ensemble has 12 members that are current and 5 members that are 12 hours old (all verifying at the same time), for a total of 17 members.

Ensemble Products

Since the MRF ensemble has 17 members, it is not feasible to analyze each solution independently. For this reason, the NCEP creates several different products to display ensemble information as GEMPAK metafiles. (Alternatively, the CDC Internet home page contains ensemble products.) Both metafiles and the raw ensemble data should be available to NWS forecast offices in the near future. Not all of the NCEP ensemble products are discussed here. For a complete list, refer to the user's guide included in the reference list (below).

(a) Ensemble Mean

Statistically, the ensemble mean should, in general, have more predictive skill than any of its individual members. This, in part, is because an average filters out the unpredictable elements of a forecast. The NCEP constructs a weighted ensemble mean that weights the high-resolution operational run more than the low-resolution members and weights the current members more heavily than the 12-hour-old members. Note that the ensemble mean will be smoother than any of the individual members. However, if most of the members agree on the phase of a short wave (indicating a high predictability), then that feature would appear in the mean. Verifications indicate the NCEP ensemble mean has significant increased skill in medium and extended ranges during the winter but not during the summer.

(b) Ensemble Spread

The greatest contribution of ensemble forecasting (to a deterministic forecaster) is information about the uncertainty or confidence of a forecast. Ensemble spread is defined at the NCEP as the standard deviation of all the members about the ensemble mean. Plots of ensemble spread are available for 500 mb and 1000 mb height fields. If the members of the ensemble have widely different solutions at a certain location, than the spread is large and low confidence should be given to forecasts for that location.

(c) Spaghetti Diagrams

Perhaps the easiest of the products to analyze and apply is the spaghetti plot (see Fig. 2, Note: Fig. 2 is a black-and-white copy of a color image that can be viewed via the WR home page.) A single contour, e.g., the 5460 m (500 mb) height contour, is plotted on the same chart for each of the 17 members. High-spread (Low-confidence) areas are easy to spot, especially when the shortwave timing and/or amplitude differs among the 17 solutions. These diagrams are also useful in relating the spread to the synoptic situation and when viewing their time evolution. Each contour highlights the spread at different latitudes. It can sometimes be misleading to view a single contour, so several height-contour diagrams should always be examined. It is also useful to examine the position of the operational full-resolution MRF run relative to the other ensemble members. In situations when the resolution is important, the full-resolution contour may differ from all other contours and yet have the most skill.

Spaghetti diagrams can also be used to graphically estimate the probability of a certain event. If most of the members agree over a region, then a forecaster can assign a high probability to that solution. Probabilities can also be assigned objectively using "clusters".

(d) Clusters

Clustering refers to grouping together members of the ensemble that are similar in some respect. If 10 out of 17 members can be clustered than the mean of those 10 members can be given a 10 out of 17 chance of occurrence. (This assumes a perfect model and an ensemble that adequately samples the growing errors.) Currently, the NCEP chooses clusters for 1000 and 500 mb height fields based on the anomaly correlations between the members.

The NCEP creates several types of cluster and probabilistic products. These products are in a state of flux and Steve Tracton (NCEP) is asking both for feedback on the current products and ideas for future products.

Verification Statistics

The primary verification statistics that are examined are the skill of the ensemble mean and the relationship between spread and forecast error, i.e., does the operational forecast really tend to have less skill in regions and on days when the spread is large?

Figure 3 compares anomaly correlation skill scores for the operational MRF, the low-resolution (T62) MRF, and the ensemble mean. These scores were computed for the 500 mb height field over the Northern Hemisphere for the period December though February, 1995. The operational MRF scores only marginally better than the nominal T62 MRF. The ensemble mean has a noticeable increase in skill after about day 6. Based on this one-season sample, the ensemble mean should be used for winter season forecasts beyond day 6. However, it should be noted that the ensemble mean crosses the 60 percent line at about day 6. (60 percent is often considered the cut-off level for useful predictive skill.) This loss of skill at extended ranges highlights the need to determine uncertainty at these forecast ranges. Please note that this verification is for the entire Northern Hemisphere, which tends to mask systematic errors for regions such as the western U.S.

Figure 4 is a "Reliability Diagram" and provides one indication that the ensemble information is useful beyond day 6. Figure 4 shows a good correlation between the probability of below normal 500 mb height (based on the ratio of the 17-members) and the corresponding observed frequency of below normal 500 mb height. (See the Fig. 4 figure caption for a description of how this reliability diagram was derived.) Toth and Kalnay (1995) provide additional verifications of ensemble spread that indicate that a useful correlation between spread and error exists for forecast times of day 6 through 10.

Another important verification is to count the number of times that the observed state of the atmosphere falls "outside" of the ensemble spread. A 17-member ensemble divides all possible solutions into 18 different bins, i.e., a bin between each ensemble member and 2 bins on either side, or "outside" the ensemble. Each of these bins is considered equally likely to contain the actual observed state of the atmosphere. Thus, using a perfect model, in 2 out of 18 cases (11 percent) reality is expected to be "outside" of the spread of ensemble solutions. Zoltan Toth (NCEP) reports that for the MRF ensemble, this occurs roughly 25 percent of the time (i.e., 14 percent above the perfect model score). Thus, about one out of 4 times the actual solution will not be within the ensemble spread. However, even in these cases, the ensemble mean may over time still have more skill than individual members and the ensemble spread may still provide useful confidence information.

Summary

The ensemble approach provides an objective measure of confidence, a feature important at extended-forecast ranges when model skill varies considerably from day to day and between different geographic regions. No ensemble is perfect and there will certainly be cases when the ensemble forecast provides misleading information. However, the use of ensemble forecasts should enable forecasters to more objectively assess their confidence in any particular long-range model forecast.

Acknowledgements

The author thanks Steve Tracton, Zoltan Toth, and Dick Wobus for freely providing and discussing the information contained in this Technical Attachment. Thanks also, to Steve Tracton, for providing the figures.

References

The NCEP Ensemble User's Guide is available via anonymous ftp. Due to security concerns the IP address is not printed here. Contact jmittelstadt@smtpgate.ssmc.noaa.gov for further information.

H. Brooks et al., 1995, Short-Range Ensemble Forecasting: Report from a Workshop (25-27 1994), *Sept. 1995 Bulletin of the AMS.*

Z. Toth and E. Kalnay, 1995, Ensemble Forecasting at NMC and the Breeding Method, *NMC Office Note 407.*

Z. Toth and E. Kalnay, 1993, Ensemble Forecasting at NMC: The Generation of Perturbations, *Dec. 1993 Bulletin of the AMS.*

S. Tracton and E. Kalnay, 1993, Operational Ensemble Prediction at the National Meteorological Center: Practical Aspects, *Weather and Forecasting,* Vol. **8,** pp. 379-398.

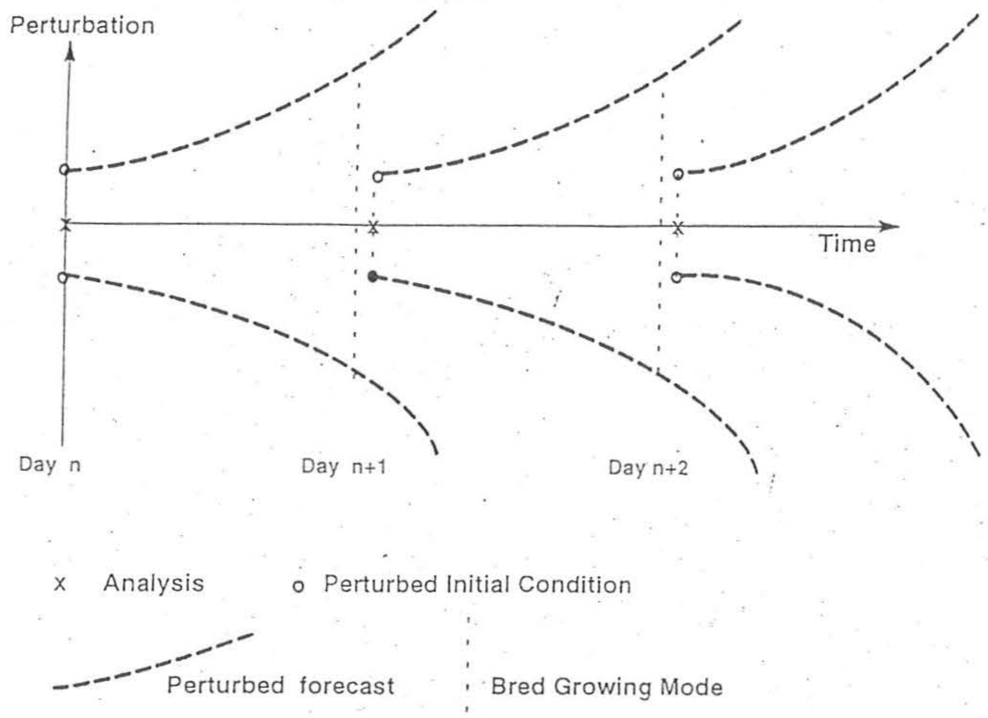


Fig. 1 Schematic Diagram of the Breeding of Growing Modes (BGM) process. See text for details.

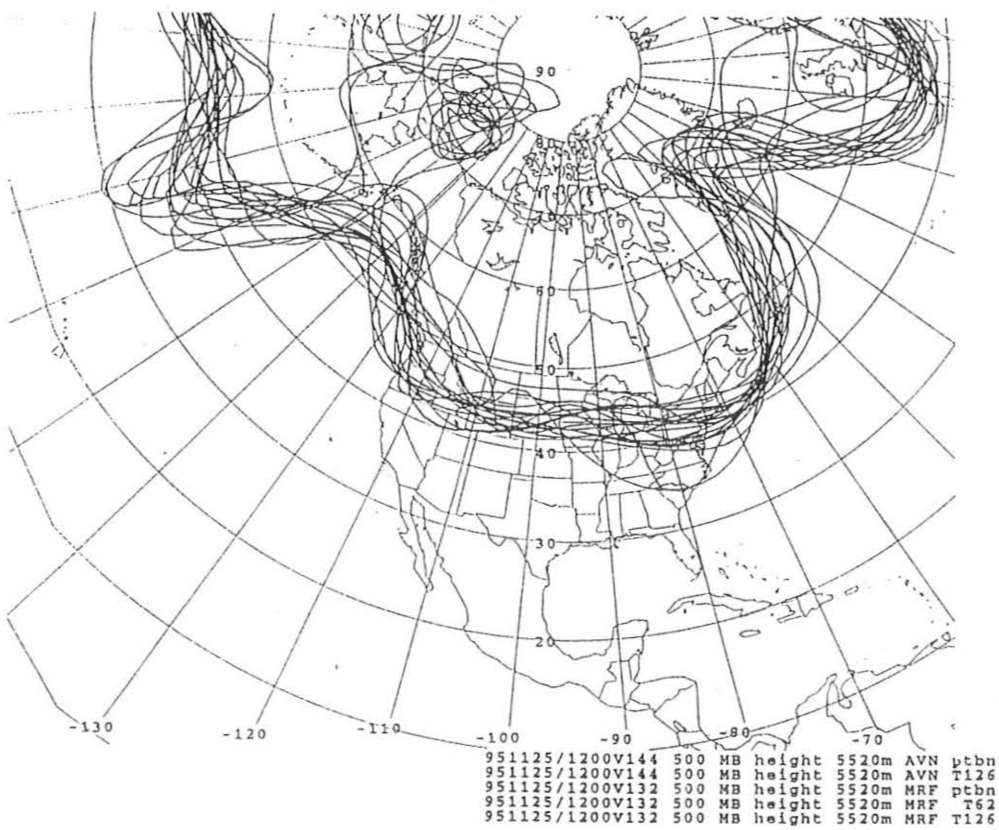


Fig. 2 Spaghetti Diagram of (500 mb) 5520 m height contours for the NCEP 17 member ensemble. The valid time is 951125/1200. Note that this is a black-and-white copy of a color product and color is used to identify the various contours.

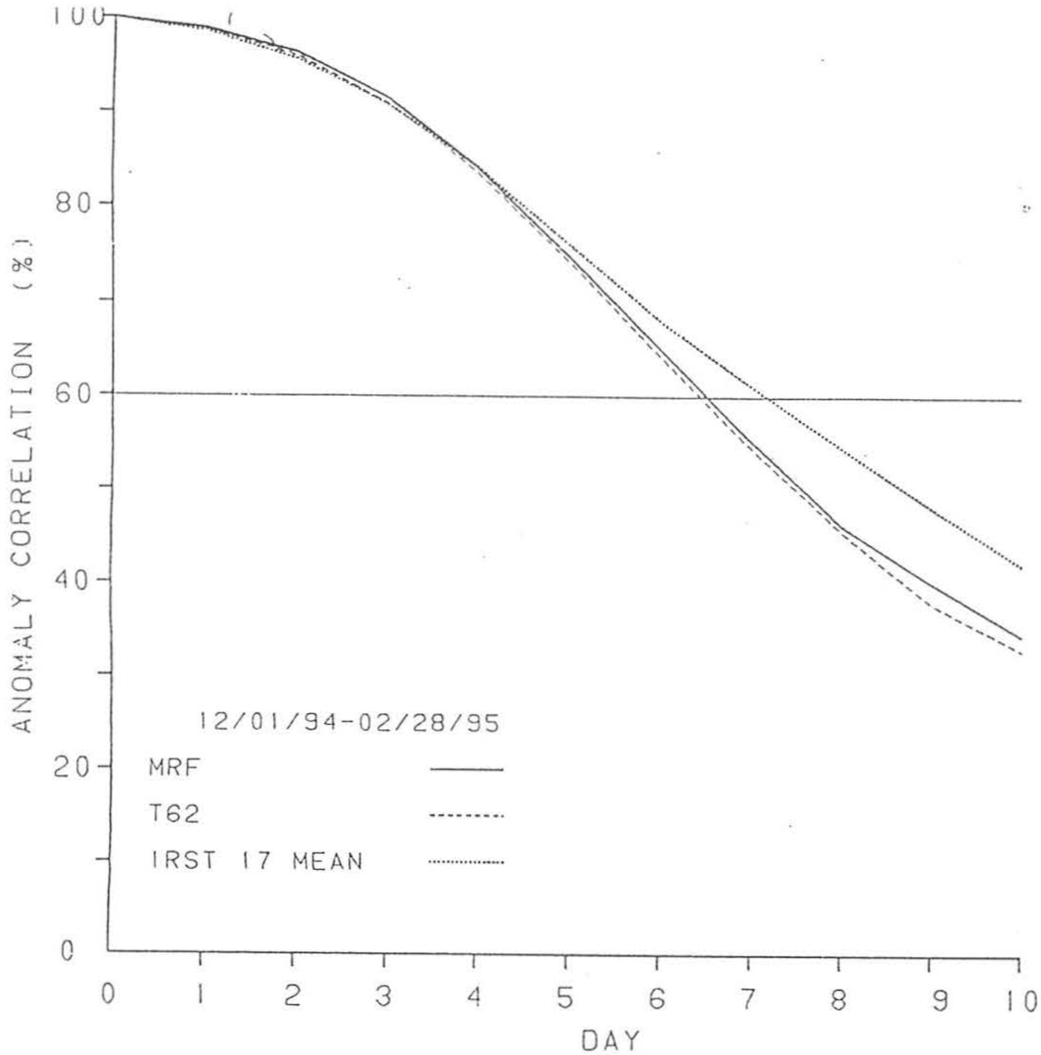


Fig. 3 Northern Hemisphere, 500 mb, anomaly correlations for days 1 through 10 for the 1995 winter season. The anomaly correlations are verifications of the operational MRF (solid line), T62 MRF (dashed line) and of the ensemble mean (dotted line).

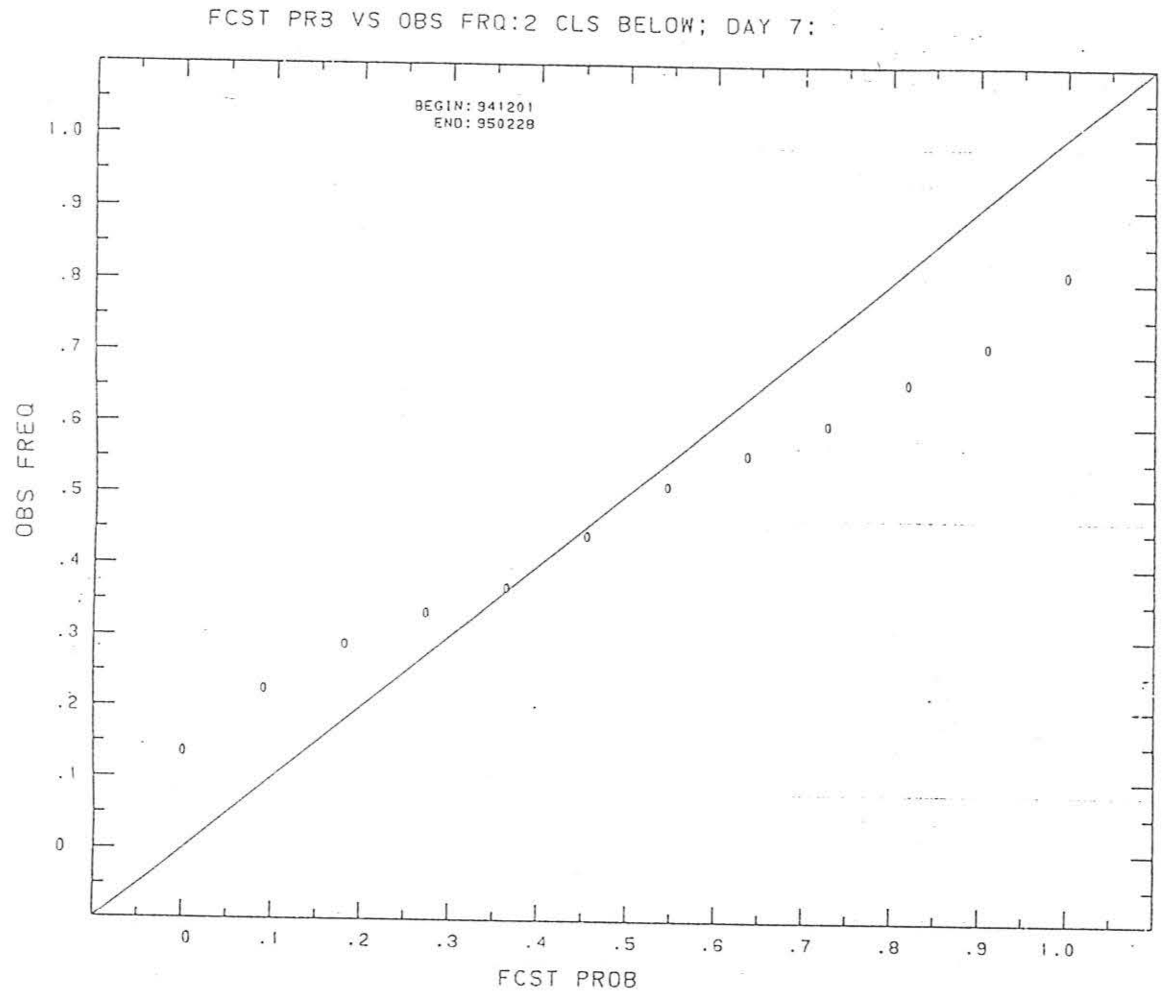
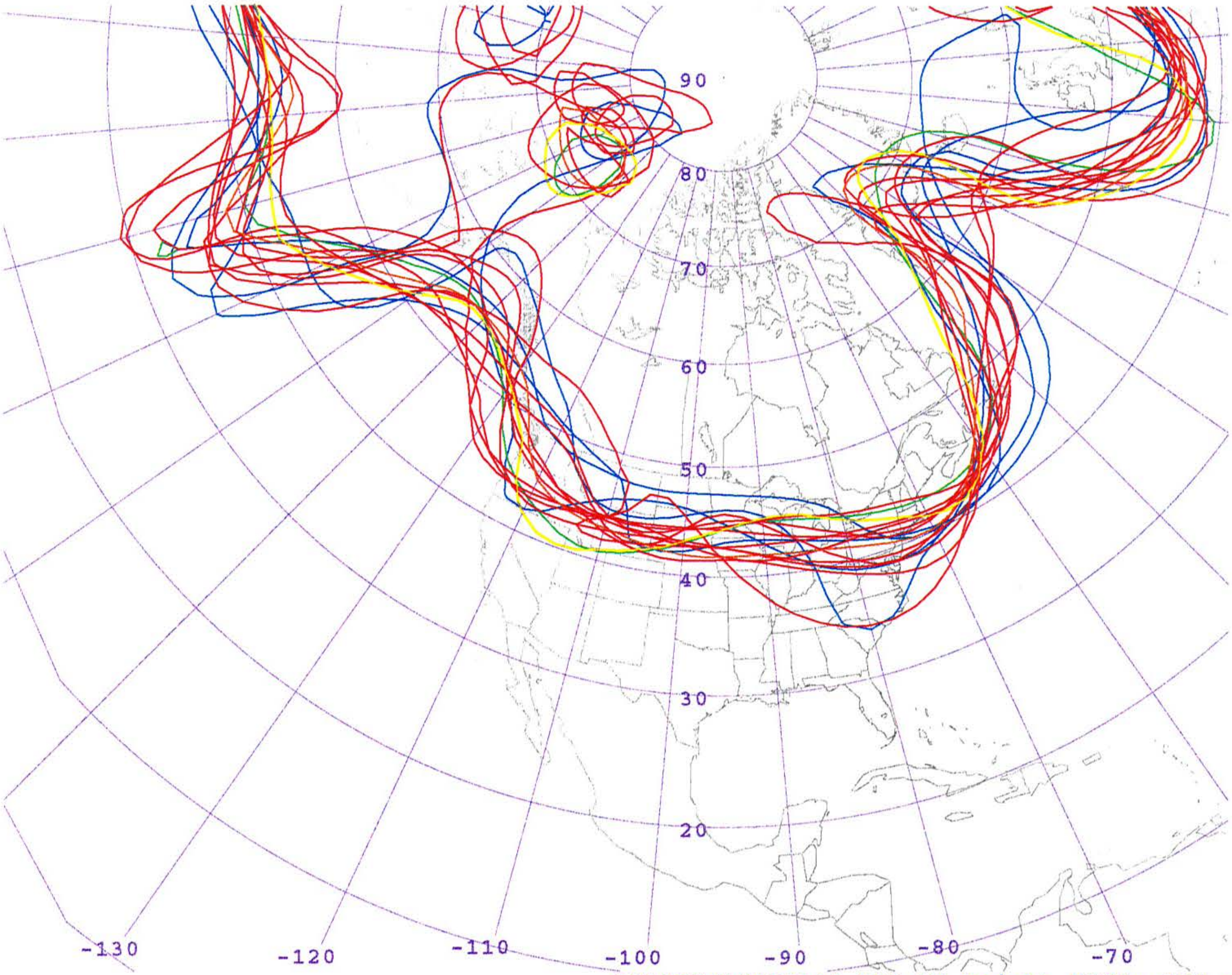


Fig. 4 Reliability Diagram showing the correlation between the probability of below normal 500 mb height (based on the ratio of the 17 members) and the corresponding observed frequency of below normal 500 mb height. This reliability diagram was derived by first categorizing every Northern Hemisphere gridpoint over the 1995 winter season according to the ratio of ensemble 7-day forecasts with below normal 500 mb heights. For example, a .1 on the horizontal axis of Fig. 4 marks the category containing all (gridpoint) forecasts such that 10 percent of the 7-day forecasts had below normal 500 mb height. The vertical axis indicates the observed frequency of below normal 500 mb heights among all the cases in a given category. For example, for all gridpoints (over the season, over the Northern Hemisphere) such that 10 percent of the members were forecasting below normal 500 mb heights, about 22 percent of the cases verified with below normal heights. A perfect score would match the diagonal.



951125/1200V144 500 MB height 5520m AVN ptbn
951125/1200V144 500 MB height 5520m AVN T126
951125/1200V132 500 MB height 5520m MRF ptbn
951125/1200V132 500 MB height 5520m MRF T62
951125/1200V132 500 MB height 5520m MRF T126