ON THE RELATIVE MERITS OF USING AN ENSEMBLE VERSUS A SINGLE HIGH RESOLUTION CONTROL FORECAST

Zoltan Toth and Yuejian Zhu

GSC at Environmental Modeling Center
National Centers for Environmental Prediction
NOAA/National Weather Service
Washington DC, USA

Summary: We present results from a comprehensive objective verification study of the NCEP global ensemble. The results for a 10–member T62 ensemble are compared to similar statistics for the NCEP high resolution (T126) MRF control forecast. Note that creating the 10–member ensemble requires as much CPU as the T126 control forecast so the computational cost of the compared two forecast systems is equal. The results indicate that for the NH extratropics, beyond 3 days lead time probabilistic forecasts generated from the ensemble have potentially more utility than those formally identical forecasts generated from the high resolution control forecast. For the SH extratropics and the tropics probabilistic forecasts from the ensemble are superior at or beyond 12 hours lead time. Since the ensemble mode forecasts have an overall reliability similar to that of the control forecast, the advantage of the ensemble approach is due to the fact that it can reliably distinguish between highly and poorly predictable situations, as indicated by varying, flow dependent probability values, as opposed to fixed probability values for the control at each lead time, based on average forecast reliability.

1. INTRODUCTION

In this paper we present objective verification results based on the NCEP global ensemble system (Toth and Kalnay, 1997) for the spring of 1997. Performance in this transition season is characteristic of the ensemble's overall year–round performance; skill is typically somewhat higher in the winter and lower in the summer than that showed below. In an earlier paper (Zhu et al., 1996) we evaluated the NCEP ensemble's performance, and compared it to that of the ECMWF ensemble for the winter of 1995/96. Further verification statistics for the ECMWF ensemble can be found, for example, in Molteni et al. (1996). For brevity, we will not discuss the detailed definition and properties of most scores that we present. The interested reader is referred to the paper of Stanski et al. (1989).

All results below (except where noted) will relate to the performance of a 10–member subset of the operational NCEP ensemble, consisting of the 10 T62 forecasts started at 0000 UTC, perturbed by bred vectors. The CPU cost of creating this ensemble is equivalent to running the operational MRF control forecast that has a resolution of T126, double that of the other ensemble members (out to 7 days lead time). Beyond scores for accuracy, probabilistic scores that measure distribution characteristics for the 10–member ensemble will also be compared to those generated from both the high (T126) and low (T62) resolution control forecasts for 500 hPa geopotential height over three geographical regions: the NH and SH extratropics and the tropics (20N–20S). To save space, detailed results are shown below only for the NH extratropics.

2. MEASURES OF ACCURACY

2.1 Pattern anomaly correlation

Fig. 1 shows the pattern anomaly correlation (PAC) values for the two controls and the ensemble mean. For the first 4 days the three lines lie over each other. The high resolution control has a slight advantage over the other two curves but the difference is confined mainly to the third digit of PAC. After day 4, however, the advantage of using the ensemble mean as an estimator of the future flow becomes apparent; for example, the 0.6 PAC level is reached at day 6.5 by the controls while only at day 7 by the ensemble mean. At later lead times the ensemble mean develops a 3–day advantage against the controls. The median of the ensemble, which is a less smoothed field, performs similarly well against the control. This indicates that the advantage of the ensemble is primarily due to selective filtering of the unpredictable forecast components, rather than a general smoothing effect.

Note also that beyond day 5 there is no advantage of running a double resolution (T126) control as compared to the low resolution (T62) control at all – in fact, the score for the low resolution control happens to be higher. Statistically, the T126 model is still expected to perform slightly better. However the differences between the forecasts developing due to the chaotic nature of the atmosphere completely masks this very weak signal even when the scores are averaged over a whole season.

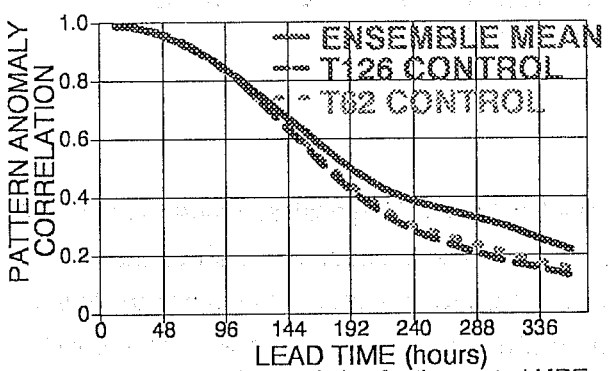

Fig. 1. Pattern anomaly correlation for the control MRF T126 forecast and the 10–member T62 ensemble mean for the NH extratropics, for March–May 1997.

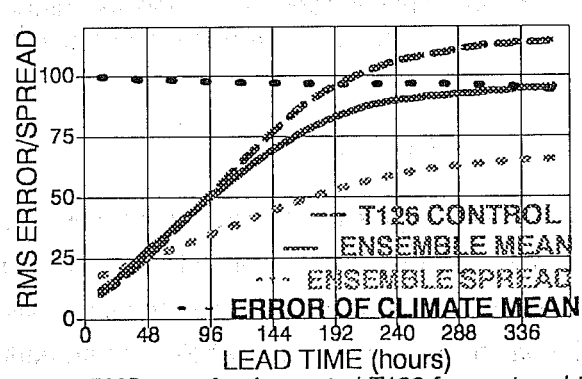

Fig. 2. RMS error for the control T126 forecast and the 10–member T62 ensemble mean for the NH extratropics, for March–May 1997. The ensemble spread is also shown.

2.2 Root mean square error and spread

As seen from Fig. 2, the ensemble mean also outperforms the high resolution control in terms of root mean square (RMS) error: at day 7 the ensemble has an error as low as the MRF control has at day 6; at day 14 as the MRF has at day 8. For an ideal ensemble the spread should equal the error in the ensemble mean. At later lead times the ensemble spread is roughly one third less than the rms error of the ensemble mean. Part of this deficiency can be attributed to case dependent bias in the MRF and a variance that decreases with increasing lead time in the T62 model (Mark Iredell, 1997, personal communication).

2.3 Average reliability

Let us consider the ensemble and control forecasts in terms of 10 climatologically equally likely bins at each grid point. In Fig. 3 we show the observed frequency of the verifying analysis falling into the same bin as the controls or the ensemble mode (bin with most ensemble members). The controls have up to 6–9 hours of advantage in reliability at short lead times while the ensemble has an advantage of a day or more beyond 9 days lead time.

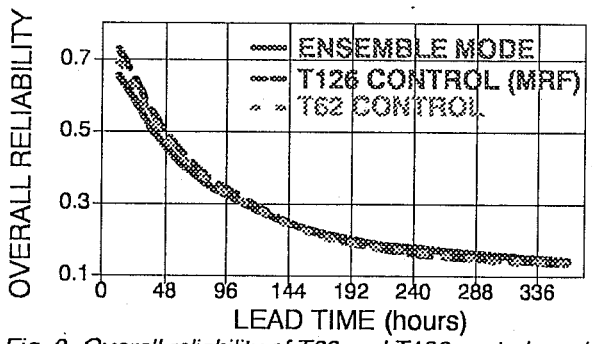

Fig. 3. Overall reliability of T62 and T126 controls and 10–member ensemble mode (most frequent value) forecasts for the NH extratropics, for March–May 1997.

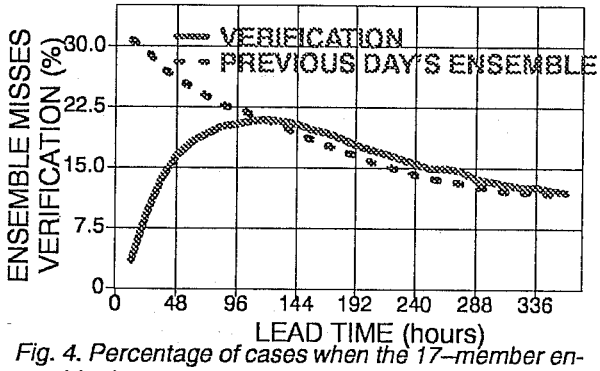

Fig. 4. Percentage of cases when the 17–member ensemble does not encompass the verifying analysis or next day's ensemble members (in excess of the 11.8% that is expected due to the limited size of the ensemble.) March–May 1997.

3. DISTRIBUTION CHARACTERISTICS

In this section, we evaluate whether the full 17–member NCEP ensemble cloud typically encompasses the verifying analysis. Following Talagrand (1994, personal communication) and Anderson (1996), we checked where the verifying analysis falls with respect to the ensemble forecast data (arranged in increasing order at each grid point; "Talagrand" distribution.) Since all perturbations are intended to represent equally likely scenarios, this distribution should be flat. In reality, because of insufficient spread and potential case dependent model bias, the verification falls more often into the two open ended categories (extremes) than expected by chance (2/18 bins=11% combined for the two extremes). Fig. 4 shows the percentage of cases when the verification falls into the two extreme categories *in excess* of the expectations. As we see, on average, the ensemble does not encompass the verification one out of 7 cases. We can also see from Fig. 4 that the ensemble changes relatively little from one day to the next – a feature that practicing forecasters appreciate a lot.

4. PROBABILISTIC MEASURES

The most important application of the ensemble forecasts is their use for the generation of probabilistic forecasts. In this section we will evaluate the performance of such forecasts, created by simply determining the percentage of the ensemble members at each grid point that fall into any of 10 climatologically equally likely categories and then using that value as the forecast probability of the event. All verification scores below are averaged over all 10 climate bins.

4.1 Reliability diagrams

Given a particular forecast probability of an event(which is a climate bin at a gridpoint), one can determine the relative frequency at which an event with that forecast probability is observed. Because of the insufficient spread in the ensemble, the ensemble forecasts are generally overconfident, i. e., they produce sharper probabilistic forecasts (probability values closer to 0 and 1) than the reliability of the forecasts allows.

It was shown by Zhu et al. (1996) that since the behavior of the ensemble is stable in time, the probabilistic ensemble forecasts can be well calibrated. The calibrated forecast probabilities are given as the observed frequencies corresponding to the forecast probabilities from a *previous time period.* For example, when 8 out of a total of 14 members fall in a climate bin, the forecast probability is not simply 8/14 but rather the observed frequency at which verifying observations fell into bins with 8 ensemble members during a preceding verification time period. All forecasts in this study have been calibrated using reliability statistics from a period 45–15 days *prior* to the actual forecast period so only verification data that are operationally available were used.

In Fig. 5, reliability diagrams are shown for different lead times. Generally, the calibrated probabilities match very well the observed frequencies, indicating close to perfect reliability. Discrepancies from perfect reliability occur at longer lead times and at higher forecast probability values due to (1) small sample sizes and (2) possible flow dependent bias in the MRF model.

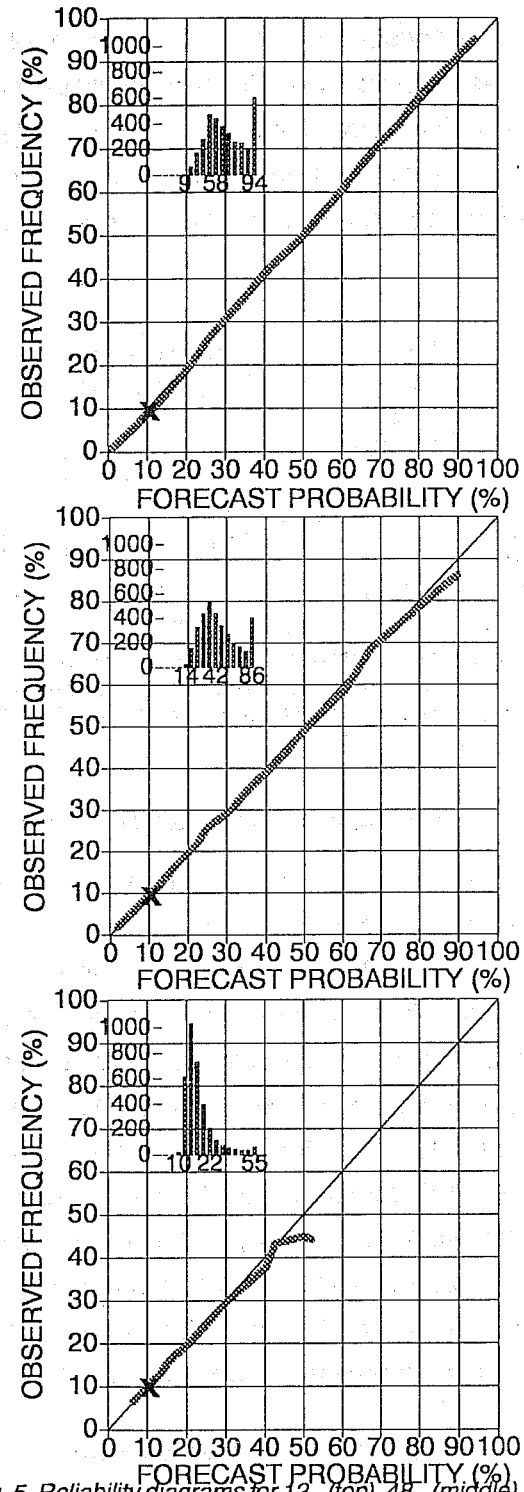

Fig. 5. Reliability diagrams for 12– (top), 48– (middle) and 240–hour (bottom) lead time 500 hPa height NH extra-tropics forecasts between March and May 1997. Forecast probabilities are based on how many ensemble members fell in any of 10 climatologically equally likely bins at each gridpoint, and are calibrated using verification statistics from the winter of 1995–96. Inserts in upper left corners show in how many events a particular forecast probability was used for the most likely bin (ensemble mode). A cross marks the performance of climatological forecasts.

4.2 Probabilistic forecasts based on the control forecast

Following Akesson (1995), categorical MRF forecasts have been converted to probabilistic forecasts, for the sake of a comparison with the ensemble–based probabilities. Initially, a probability of 1 is assigned to the climate bin where the MRF falls at each grid point, and 0 probability to the other 9 climate bins. These probabilities are then calibrated in the same fashion as the ensemble forecast probabilities, resulting in "binary" probabilistic forecasts that are practically perfectly reliable.

In order to have a fair comparison between the control and ensemble forecasts in terms of probabilistic measures, the ensemble probabilistic forecasts were degraded by retaining the probability value for the ensemble mode (P_m, most likely climate bin) and distributing the remaining probability ($1-P_m$) over the other 9 bins. Thus the probabilistic forecasts from the MRF control and ensemble mode are formally identical: one (the most likely climate bin) has a higher probability value whereas the other 9 bins have all the same and lower value of probability. The only difference is that for the ensemble, P_m varies depending on how predictable the flow is (i. e., how many ensemble members fall into the most likely bin) whereas for the MRF this value is fixed for each lead time.

4.3 Ranked probability skill score (RPSS)

RPSS is a generalization of the Brier skill score for multi–categorical forecasts where the categories can be ordered. It rewards probabilistic forecasts that are both reliable and sharp, as compared to the background climatological probabilities. In Fig. 6 we compare RPSS values for the controls and the ensemble mode forecast. For the first 2–3 days both the high and low resolution control forecasts (which perform rather similarly) have a slight advantage over the ensemble mode, perhaps due to better average reliability (see Fig. 3). Beyond 3 days, however, it is clear that the ensemble mode forecast, which is able to distinguish highly predictable cases (reliable sharp forecasts) from poorly predictable situations, has a large advantage. Note also that the full ensemble (which is not directly comparable to the performance of the controls since it contains a full probability distribution that the control does not) substantially improves the forecasts, especially at short lead times.

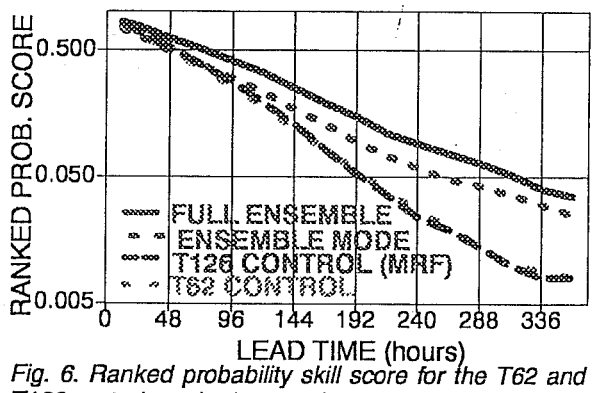

Fig. 6. Ranked probability skill score for the T62 and T126 controls and a 10–member ensemble forecast for the 500 hPa height, NH extratropics, March–May 1997.

4.4 Relative operating characteristics (ROC)

ROC, a measure from signal detection theory, offers another way of comparing the performance of the control forecast with that of the ensemble. Its main advantage is that categorical forecasts (like

the control forecasts in our case) can be directly compared with probabilistic forecasts (like those from the full ensemble, see Stanski et al., 1989). Cases are classified according to observations, so reliability is not considered at all. A forecast system (i. e., control falling in a bin or ensemble exceeding a certain probability threshold in a bin) is better than another if its hit rate is higher and false alarm rate is lower than the other's.

The ROC–area measures the quality of forecasts in these terms. As can be seen from Fig. 7, the ensemble has a large advantage over the control forecasts, indicating a better separation between forecast probabilities used when an event occured vs. when it did not occur.

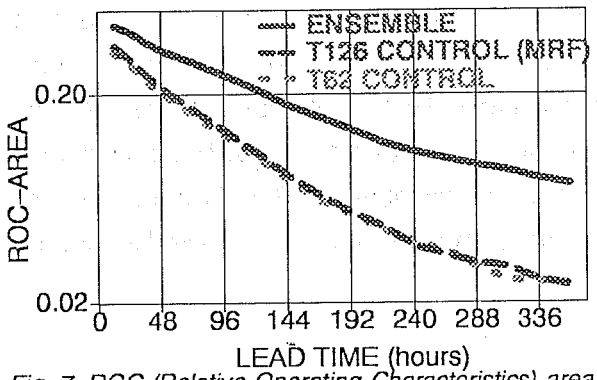

Fig. 7. ROC (Relative Operating Characteristics) area for T126 and T62 control and 10–member T62 ensemble forecasts for the 500 hPa height, NH extratropics, March–May 1997.

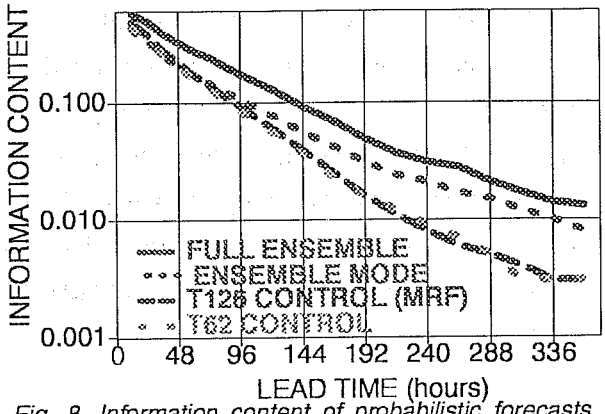

Fig. 8. Information content of probabilistic forecasts based on the T62 and T126 control forecasts and the mode (most frequent value) of a 10–member ensemble, for the NH extratropics, for March–May 1997.

4.5 Information content

Measuring the information content of the forecasts offers perhaps the most direct way of evaluating how useful the forecasts can be for the end users. We define the information content of a single forecast (expressed in terms of probabilities over the 10 equally likely climate bins) as: $I = 1 - \sum_{i=1}^{10} P_i \log_{10} P_i$, where P_i is the observed frequency associated with a forecast probability (assuming that perfectly reliable forecasts can be made). In Fig. 8 we can compare the information content of the control forecasts with that of the ensemble mode. The curves run rather similarly to those for RPSS (Fig. 6). Again, for the first 2 days or so there is an up to 6 hour advantage for the high resolution control forecast. However, after day 3 the ensemble mode forecast is a clear winner. In fact, the information content in the ensemble mode forecasts beyond 8 days is double that of the controls; the information content of a 14–day ensemble mode forecast is as high as that of a 9–day control forecast. Average reliability of these two forecasts is rather similar (see Fig. 3) so the advantage of the ensemble is clearly due to its ability to correctly predict cases with high vs. low predictability.

Note that beyond 3 days lead time the full ensemble distribution has an information content that is more than double that of the control. In other words, a full probability distribution forecast at 6–day

lead time has as much information as a categorical forecast based on a high resolution control run at day 4. Although the full ensemble forecasts are formally not directly comparable to the control probabilistic forecasts, they are operationally available. These statistics point to the great value that probabilistic forecasts based on the ensemble hold for the user community.

5. DISCUSSION

In this paper we compared the performance of a 10–member T62 ensemble forecast to that of a T126 single control forecast in different terms, including pattern anomaly correlation and root mean square error. In these terms the ensemble mean forecast becomes more skilful than the high resolution control forecast beyond 84 hours lead time for the NH extratropics, and beyond 48 and 108 hours lead time for the SH extratropics and for the Tropics, respectively. Accuracy alone, however, measures only one aspect of a forecast system. Distribution characteristics that address the quality of probabilistic forecasts, such as ranked probability skill score, relative operating characteristics and information content, measure the performance of forecast systems in a more complex way (including information on accuracy as well). In these measures probabilistic forecasts based on the ensemble mode, which are formally comparable with probabilistic forecasts generated based on the control forecasts, show higher quality than the control beyond 3 days for the NH extratropics, and right from the beginning or after 12 hours lead time over the tropics and SH extratropics. This is due to the fact that the ensemble can distinguish between highly and poorly predictable flow configurations.

These results indicate that even if the accuracy of the high resolution control forecast is considerably above that of the lower resolution ensemble (as is certainly the case in the tropics), the overall utility of the ensemble mode forecast, as indicated by the distribution measures, can be considerably higher than that of the MRF control. The results for the SH extratropics, and especially for the tropics indicate that there may be no need to run a control at a resolution higher than that used for generating the global ensemble. Results for the NH extratropics vary, with some measures suggesting a possible gain for running a higher resolution control up to 2–3 days, while others (ROC) indicating no need for a higher resolution control at all.

Comparing the information content of binary probabilistic forecasts constructed from the MRF control to that of full probability distributions from the ensemble, the importance of providing and using full probabilistic forecasts, instead of categorical forecasts, becomes evident. This way the information content of the forecasts can be more than doubled beyond three days lead time, and the predictability limit can also be extended by several days. This is due to the combined impact of (1) providing full forecast probability distributions (which could possibly also be accomplished with using only one control forecast) and more importantly, (2) distinguishing between more and less predictable situations (which cannot be done without an ensemble).

6. ACKNOWLEDGEMENTS

We are grateful to our colleagues, Eugenia Kalnay, Tim Marchok, Richard Wobus, Steven Tracton, and Stephen Lord for their help and encouragement during the coarse of this work.

7. REFERENCES

Akesson, O., 1995: Comparative verification of precipitation from EPS and from the operational T213 forecast. Proceedings of the Fifth Workshop on Meteorological Operational Systems, 13–17 November 1995, Reading, England, 297–299. [Available from: ECMWF, Shinfield Park, Reading, RG2 9AX, UK.]

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations, *J. Climate*, **9**, 1518–1530.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble system: Methodology and validation. *Q. J. R. Meteorol. Soc.*, **122**, 73–119.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in meteorology. WMO World Weather Watch Technical Reprot No. 8, WMO/TD. No. 358.

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

Zhu, Y., G. Iyengar, Z. Toth, M. S. Tracton, and T. Marchok, 1996: Objective evaluation ofthe NCEP global ensemble forecasting system. Preprints of the 15th AMS Conference on Weather Analysis and Forecasting, 19–23 August 1996, Norfolk, Virginia, p. J79–J82.