OBJECTIVE EVALUATION OF THE NCEP GLOBAL ENSEMBLE FORECASTING SYSTEM

JP1.12

Yuejian Zhu[1], Gopal Iyengar[2], Zoltan Toth[1], Steve M. Tracton and Tim Marchok[1]

Environmental Modeling Center, NCEP, NWS/NOAA
Washington DC 20233

1. INTRODUCTION

In this paper we present objective verification results based on the NCEP global ensemble system (Kalnay and Toth, 1996). In accompanying papers, we discuss the utility of ensemble forecasting in general (Tracton et al., 1996), and give a synoptic overview of the performance of the NCEP ensemble (Wobus et al., 1996).

Ensemble prediction is a relatively new component in operational numerical weather forecasting. Not surprisingly, the verification of ensemble forecasts is also an emerging new area. For brevity, we will not discuss the definition and detailed properties of most scores that we present. The interested reader is referred to the paper of Stanski et al. (1989).

Our goal is the evaluation of the NCEP ensemble during the winter of 1995/96 but we will also show results for the ECMWF ensemble prediction system (see Molteni et al., 1996). For a clear comparison, we include only those days for which both ensembles were available in our archive (a total of 78 days during December–February 95/96.) At NCEP, there are 10 and 4 perturbed forecasts started at 00Z and 12Z, respectively, so in all comparisons the first 10 (or 14) members of the ECMWF ensemble is used so that the two ensembles we compare have always the same membership. Results shown are for the 500 hPa geopotential height for the Northern Hemisphere extratropical belt, for 24, 48,..., 240 hours lead time. For each ensemble, its own control analysis is used as the verifying field.

2. PERFORMANCE OF THE CONTROL FORECASTS

Most members of the ensemble at NCEP are run with the T62, 28-level version of the MRF model, while a similarly configured, T63 model is used for the perturbed forecasts at ECMWF. The two models have similar RMS errors, though the ECMWF model, especially after 4–5 days, has a slightly better performance (Fig. 1.) Any differences found in terms of the performance of the ensemble, at least for the first part of the time period, must be related to the initial perturbations (and not model performance.)

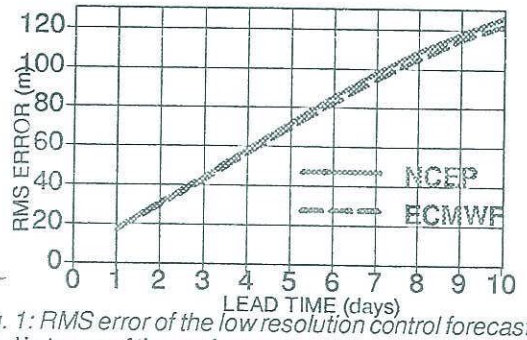

Fig. 1: RMS error of the low resolution control forecasts.

3. ENSEMBLE MEAN AND SPREAD

In this section, RMS results are shown for 10-member ensembles. The use of pattern anomaly correlation yields very similar results.

The most basic measure of the performance of an ensemble is a comparison of the errors associated with the ensemble mean and the control forecast. Fig. 2 shows the improvement in RMS scores for both 10-member ensembles due to ensemble averaging. In the medium (3–7 days) range the NCEP ensemble mean has a one day advantage over the ECMWF ensemble in a relative sense. When considering all 34 members of the ECMWF ensemble (including the more skillful high resolution control), this advantage is reduced to about 10–12 hours. Because of the ECMWF model itself performs somewhat better (see Fig. 1), the ECMWF ensemble mean outperforms the NCEP ensemble mean in absolute terms over all lead times.

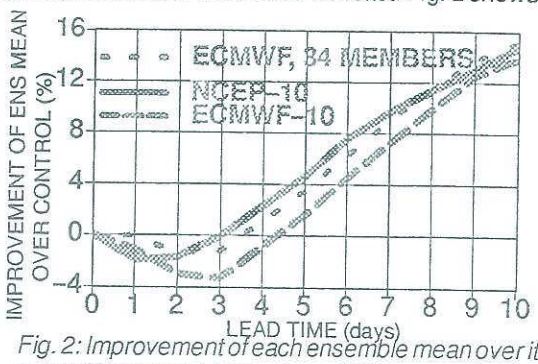

Fig. 2: Improvement of each ensemble mean over its control in terms of percentage RMS error reduction.

Ideally, the ensemble spread around the mean should be equal to the error of the ensemble mean. From comparing Figs 1, 2 and 3, we can see that this is not the case: both ensembles have a deficiency in rms spread of 25–30 % in the medium and extended range (and even in the short range for the ECMWF ensemble.) Due to the special formulation of the initial perturbations, the ECMWF ensemble spread increases more rapidly than the NCEP ensemble spread during the first two days. At days 3–4, the spread is equal in the two ensembles, while it increases slightly less afterwards in the ECMWF ensemble. Note that the initial spread in the ensemble could be easily increased. However, other characteristics of the ensemble may be negatively impacted.

4. DISTRIBUTION CHARACTERISTICS

1 GSC (Laurel, MD) at NCEP. Corresponding author address: Y. Zhu, NCEP/EMC, 5200 Auth Rd., Room 207, Camp Springs, MD 20746. wd20yz@sun1.wwb.noaa.gov

2 National Center for Medium Range Weather Forecasting, New Delhi, India

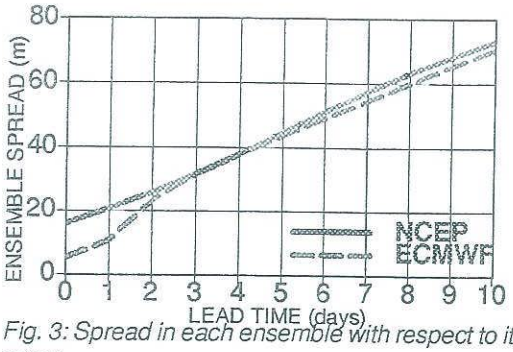

Fig. 3: Spread in each ensemble with respect to its mean.

In this and the following section, results are shown for 14-member ensembles (though results are very similar for 10 members.)

4.1 *Verification rank distribution*

Following the suggestions of Talagrand (1994, personal communication) and Anderson (1996), we checked where the verifying analysis usually 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. The results for the NCEP ensemble are not very far from this ideal situation (Fig. 4). However, both ensembles display an excessive

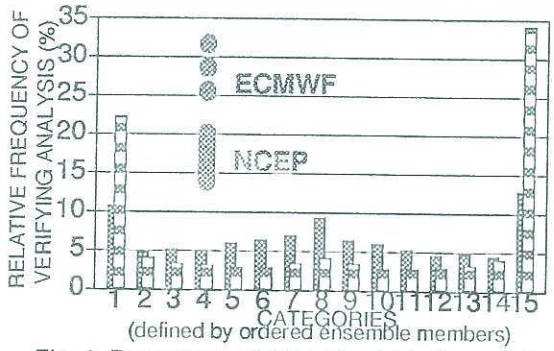

Fig. 4: Percentage of cases in which the verifying analysis falls in each of the 15 bins defined by the 14 ordered ensemble members at each grid point at 24 hours lead time. The expected value is 6.6 %.

number of cases in which the verification falls outside the range of the ensemble (Fig. 5): 18–43 % of all cases for the ECMWF and 10–21 % for the NCEP ensemble.

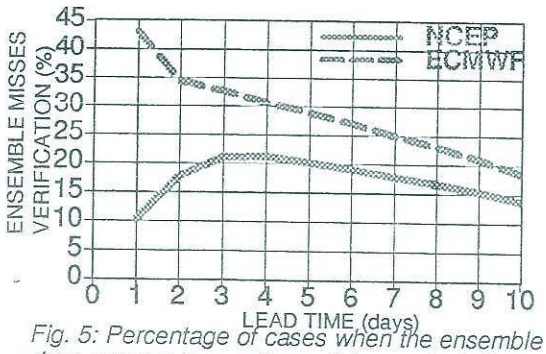

Fig. 5: Percentage of cases when the ensemble does not encompass the verifying analysis (in excess of the 13.3% that is expected due to the limited size of the ensemble.)

4.2 *Time consistency in the ensemble forecasts*

The "Talagrand" distribution can also be used to test how much the ensemble valid on a particular day differs from that valid on the same day but issued a day earlier. The overlap of today's ensemble with that of yesterday's is gratifying, with only 6–20 % of the two sets of time lagged ensembles differing at or beyond day 4 lead time. However, the NCEP ensemble displays considerably more time consistency at short lead times (Fig. 6).

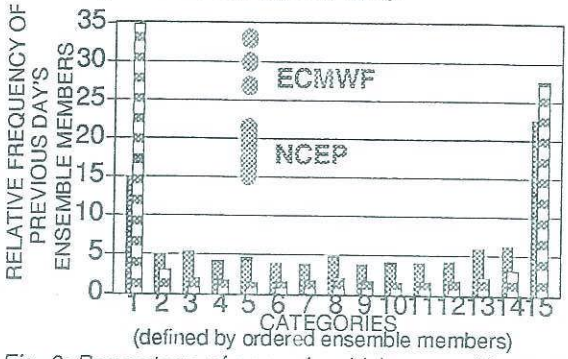

Fig. 6: Percentage of cases in which ensemble members from previous day fall in each of the 15 bins defined by the 14 ordered ensemble members at each gridpoint at 24 hours lead time. The expected value is 6.6 %.

5. PROBABILITY MEASURES

Probably 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 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.

5.1 *Reliability, resolution and sharpness*

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. Ideally, one would like to see that the observed frequency is close to the forecast probability. As seen from Fig. 7, both ensembles work well, though the NCEP ensemble offers more reliable probabilistic forecasts (i.e., verification curve is closer to the 45 degree line).

It is well known that the reliability of probabilistic forecasts can be improved if the forecast process is stable, i. e., the conditional observed frequencies do not change in time. As seen in Fig. 8, this is the case with the ensembles: after a simple calibration, in which the forecast probabilities are given as the observed frequencies from a previous time period, both the NCEP and ECMWF ensembles provide probabilistic forecasts with excellent reliability even at extended lead times.

One can also see from Figs. 7 and 8 that the NCEP ensemble can make a somewhat better distinction between likely and unlikely events. For example, the observed frequencies of different events at day 3 (Fig. 8.a) are in the range of 3–86% (instead of 4–65% for the ECMWF ensemble). Resolution, as this quality of probabilistic forecasts is called, is good even at day 8, where in some cases the forecast probability (and corresponding

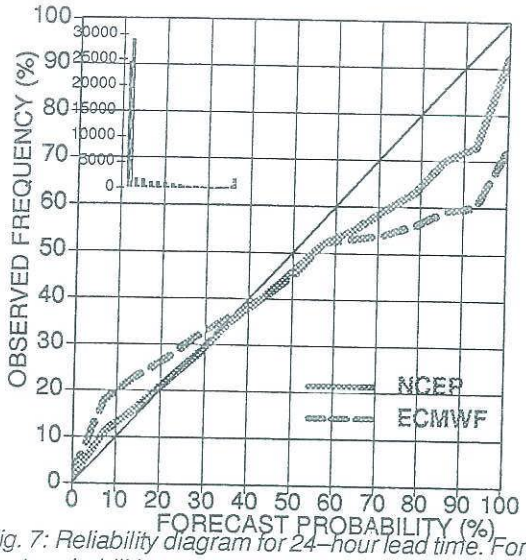

Fig. 7: Reliability diagram for 24–hour lead time. Forecast probabilities are based on how many ensemble members fell in any particular climate bin at each grid-point. Insert in upper left corner shows in how many events a particular forecast probability was used.

observed frequency) is still almost 7 times the 10% climatological likelihood.

As seen from the inserts in Figs. 7 and 8, the ECMWF ensemble, especially at short lead times, offers "sharper" probabilistic forecasts, i. e., there are more cases with extreme (0 or 1) forecast probabilities. However, this extra sharpness does not seem to be justified by the generally poorer performance in terms of reliability and resolution.

5.2 *Ranked probability skill score*

RPSS is used as another measure of the performance of the probabilistic forecasts based on the ensemble. 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 have high resolution, as compared to the background climatological probabilities.

Confirming results from the reliability diagrams, Fig. 9 indicates that probability forecasts from both ensembles have a useful skill over the entire 10–day lead time period. At short and intermediate lead times, where the RPSS has high values, the NCEP ensemble has an advantage (of a half day or so) over the ECMWF ensemble while beyond day 7 the situation is reversed.

5.3 *Relative operating characteristics*

ROC, a measure from signal detection theory, is especially useful in ensemble verification because it offers another way of comparing the performance of the control forecast with that of the ensemble. Cases are classified according to observations. A forecast system (i. e., control falling in a bin or ensemble exceeding a certain probability in a bin) is better than another if its hit rate is higher and false alarm rate is lower than the other's.

It is worth noting that in this measure the ECMWF control has an advantage over the NCEP control at all lead times. Despite this fact, the NCEP ensemble has somewhat higher ROC scores at short and intermediate lead times (Fig. 10). It is encouraging that in terms of ROC, the

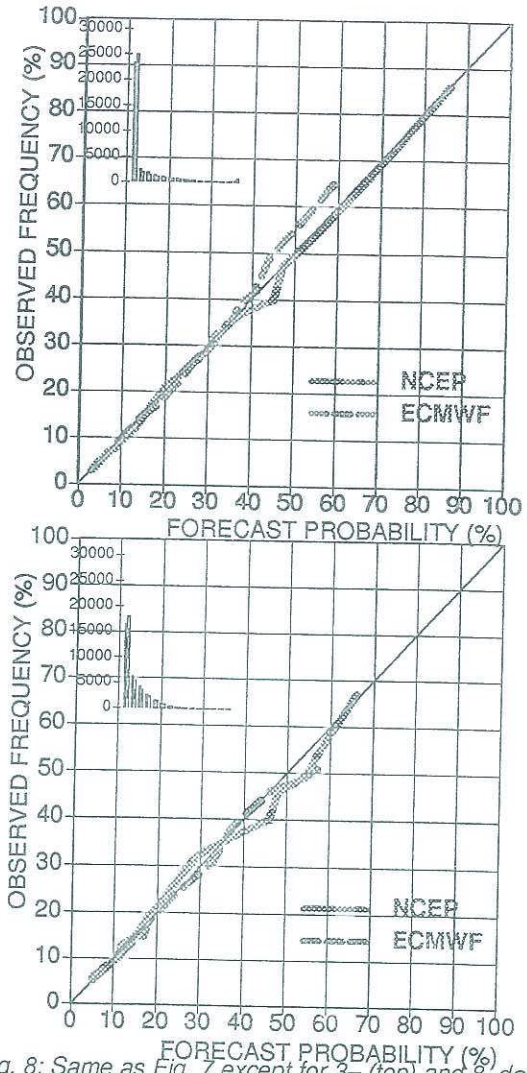

Fig. 8: Same as Fig. 7 except for 3– (top) and 8–day (bottom) lead time for January 1996. Forecast probabilities are based on observed frequencies associated with the same number of ensemble members falling in a particular bin during December 1–20, 1995.

NCEP (ECMWF) ensemble is better than the corresponding low resolution control forecast right from the beginning (from day 3 on), and better than the high resolution control from day 4 (day 6) on.

5.4 *Likelihood of individual forecast members*

In an operational forecast environment it is important to know the likelihood that the verifying analysis would be closest to any one member of the ensemble. In this context, we considered all the 17 and 34 members of the NCEP and ECMWF ensembles respectively.

In Fig. 11 we compare the likelihood of the verifying analysis falling next to the high resolution control (HRC) to that expected for any other member. The results are strikingly different, indicating that it is several times more likely that the analysis will fall next to the HRC in the ECMWF ensemble while this is only 1.5 times more likely to happen in the NCEP ensemble. We also note that in this sense, the

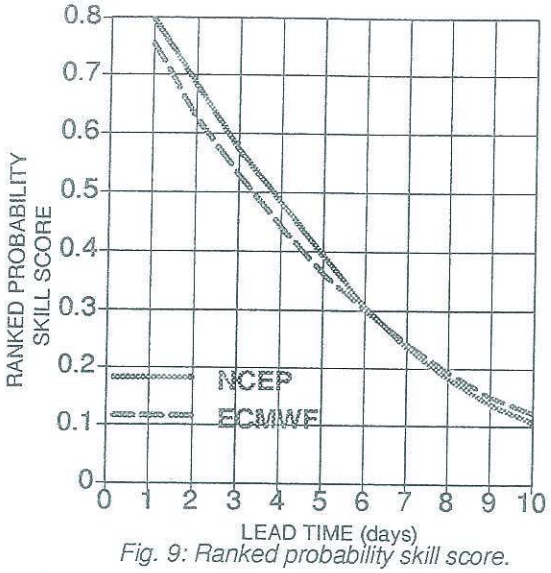

Fig. 9: Ranked probability skill score.

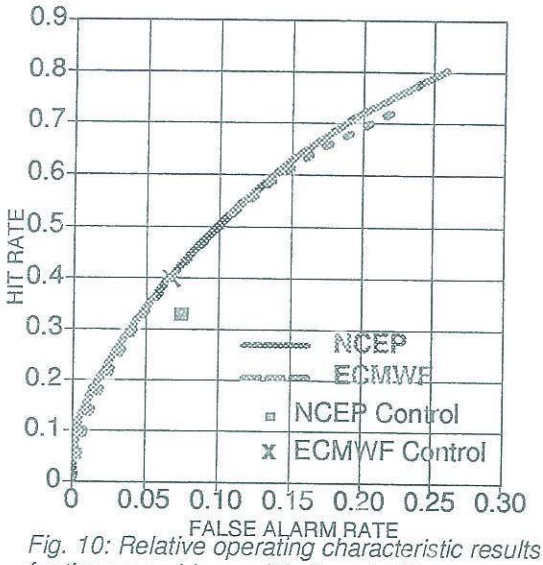

Fig. 10: Relative operating characteristic results for the ensembles and their control forecasts.

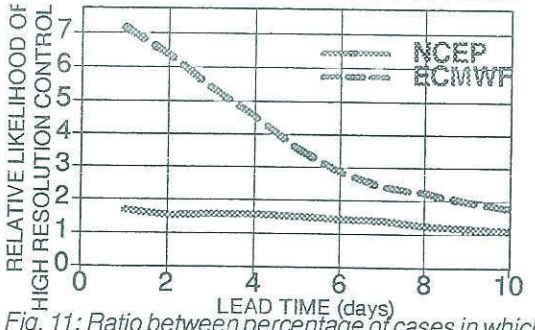

Fig. 11: Ratio between percentage of cases in which the verifying analysis falls next to the high resolution control forecast and that expected by chance (assuming all members are equally likely).

perturbed forecasts are as likely as their low resolution control in both ensembles.

6. SUMMARY AND DISCUSSION

In this paper we have computed objective verification statistics for the NCEP and ECMWF operational ensemble forecasts. The results indicate that the ensembles provide important extra information that is not available by using the control forecasts only. In particular we found that the ensemble mean provides an improvement over the control forecasts in the medium and extended ranges. In addition, probabilistic forecasts based directly on the ensemble forecasts provide useful guidance with respect to the likelihood of alternate forecast scenarios.

We find it especially encouraging that with minimal postprocessing, very reliable probabilistic forecasts can be formulated using the ensembles. In the present paper, we considered only 500 hPa geopotential height forecasts. However, in the near future we plan to evaluate, in a similar manner, forecasts for other variables that are more closely related to surface weather. We hope that with a relatively simple postprocessing, reliable probabilistic forecasts can be made for those variables as well.

At both centers work is underway to further improve the ensemble forecasts which will undoubtedly improve their usefulness as well. The advent of NWP changed weather forecasting forever. The advance of the ensembles now promises to bring another new era in which a weather forecast would not be complete unless it is expressed in terms of a probability distribution for different, more or less likely events.

As for the comparison of the NCEP and ECMWF ensembles, most of our results indicate that for the first 6 days or so the NCEP ensemble can offer a somewhat better forecast guidance. The difference in performance roughly equals to a half day's or a day's skill. This is despite the fact that the ECMWF control forecast is slightly better. This indicates that the difference in the performance of the ensembles must be related to the difference in the initial perturbations that the two centers use (bred vectors vs. singular vectors, see e. g., Toth et al., 1996). The results suggest that the NCEP perturbations may be more representative of the actual uncertainty in the control analysis.

7. ACKNOWLEDGEMENTS. We would like to express our gratitude to the directors of ECMWF and NCEP, David Burridge and Ronald McPherson, who gave their approval for the experimental exchange of the two centers' ensembles. Thanks are also due to Horst Bottger and John Hennessy of ECMWF and John Ward of NCEP who, along with their colleagues, made the data exchange possible.

8. REFERENCES

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

Kalnay, E., and Z. Toth, 1996: Ensemble prediction at NCEP. Preprints of the 11th AMS Conference on Numerical Weather Prediction, 19–23 August 1996, Norfolk, Virginia.

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., I. Szunyogh, and E. Kalnay: Singular, Lyapunov and bred vectors in ensemble forecasting. Preprints of the 11th AMS Conference on Numerical Weather Prediction, 19–23 August 1996, Norfolk, Virginia.

Tracton, M. S., Z. Toth, and R. Wobus, 1996: On the principles and operational utility of ensemble forecasts. Preprints of the 15th AMS Conference on Weather Analysis and Forecasting, 19–23 August 1996, Norfolk, VA.

Wobus, R., Z. Toth, S. M. Tracton, and E. Kalnay, 1996: How the NCEP ensemble works: Synoptic examples. Preprints of the 15th AMS Conference on Weather Analysis and Forecasting, 19–23 August 1996, Norfolk, Virginia.