PROBABILISTIC FORECASTS AND EVALUATIONS BASED ON A GLOBAL ENSEMBLE PREDICTION SYSTEM

YUEJIAN ZHU

Environmental Modeling Center
NCEP/NWS/NOAA, Washington, DC 20233, USA
E-mail: Yuejian.Zhu@noaa.gov

(Manuscript received 2 December 2002)

In the past decade, ensemble forecasting has developed into a major component of numerical weather prediction. With increases in computing resources, it is becoming more realistic to produce operational ensemble forecasts for compatible members with comparable resolutions in many numerical weather prediction centers around the world. Probabilistic forecasts based on a global ensemble prediction system, especially flow-dependent forecast probability distribution, which can be readily generated from an ensemble, allow for the identification of weather systems with high and low uncertainties. The potential economic benefit achieved by using ensemble probabilistic forecasts is significant when compared to that of a deterministic forecast. Among NCEP global ensemble-based applications, the relative measure of predictability is one of the excellent prediction tools used to estimate forecast uncertainty. Probabilistic quantitative precipitation forecasts can supplement short-/medium-range forecasts. However, the ensemble forecasts, like any numerical weather prediction system, are biased. The bias of ensemble forecasts is very similar to those of a deterministic forecast and comes from the imperfections of a numerical model such as initial conditions, physical parameterizations, numerical schemes, etc. The bias of ensemble forecasts can be removed, however, by applying a statistical calibration method. With the use of such a method, calibrated ensemble forecasts and ensemble-based, calibrated probabilistic forecasts can offer the possibility of bias-free products to the meteorological community and other users.

1. Introduction

One of the goals of the United States National Weather Service (US/NWS) for 2000-2005 is to provide weather, water and climate forecasts in probabilistic terms by the year 2005 (NWS 1999). In the past decade, global ensemble forecasting has been implemented into a major component of numerical weather prediction, such as the National Centers for Environmental Prediction (NCEP) ensemble prediction system (EPS), which is based on the breeding of growing mode (BGM) method (Toth and Kalnay 1993; Tracton and Kalnay 1993; Toth and Kalnay 1997), and the European Center for Medium-Range Weather Forecasts (ECMWF) EPS, which is based on a singular vector (SV) method (Palmer et al 1992; Molteni et al 1996). Both of the NCEP and ECMWF ensemble forecast systems add a small perturbation to the initial state of the model analysis. The Meteorological Service of Canada (MSC) also uses an EPS which is based on the system simulation experiment (SSE), but uses different methods to generate initial perturbations (Houtekamer and Derome 1996), including the use

of two different numerical models, different physical parameterization packages which are designed to simulate observation errors, model errors, imperfect boundary conditions, etc. The operational applications of EPS in major centers around the world have been offering dramatic information in addition to deterministic forecasts. There are many advantages of using ensemble forecasts, such as potentially providing case-dependent estimates of forecast uncertainty. The initial uncertainty and model uncertainty are two main sources that limit the skills of single/deterministic forecasts in a highly flow-dependent forecasting system. From a long term objective evaluation, there is a gain in skill of about 24 hours when comparing the ensemble mean (Fig. 1a and 1b, ENS, closed circle marks) to higher resolution deterministic forecast (MRF, cross marks) for root means square (RMS) errors and pattern anomaly correlation (PAC, not shown here) for an analysis based on Northern Hemisphere extra-tropical (20N-80N) 500hPa geopotential height 6-day forecasts (Zhu et al 1996; Fig. 1a and 1b). The season statistics of the ensemble spread (SPR, open square marks, where SPR is defined as standard deviation of the ensemble members from the ensemble mean). RMS errors between the NCEP analysis and climatology (CLM, closed square marks) are also shown in Fig. 1a and 1b. Note that the usage of computer resource for low resolution ensemble forecasts is less than or equal to the higher resolution deterministic forecast, but the ensemble probabilistic forecasts have more skill than the higher resolution deterministic forecast even for extreme events (Zhu and Toth 2001), based on Brier Skill Score (BSS) statistics. In fact, the potential economic value of an ensemble-based weather forecast (Zhu et al 2002) has indicated many advantages of using ensemble probabilistic forecast information. In the next section, I will briefly describe the current NCEP configuration, public data access and experiment setup. In section 3, selected current probabilistic forecast products would be introduced. The two probabilistic evaluation methods will be discussed in section 4. Finally, in section 5, conclusions and further discussions will be presented.

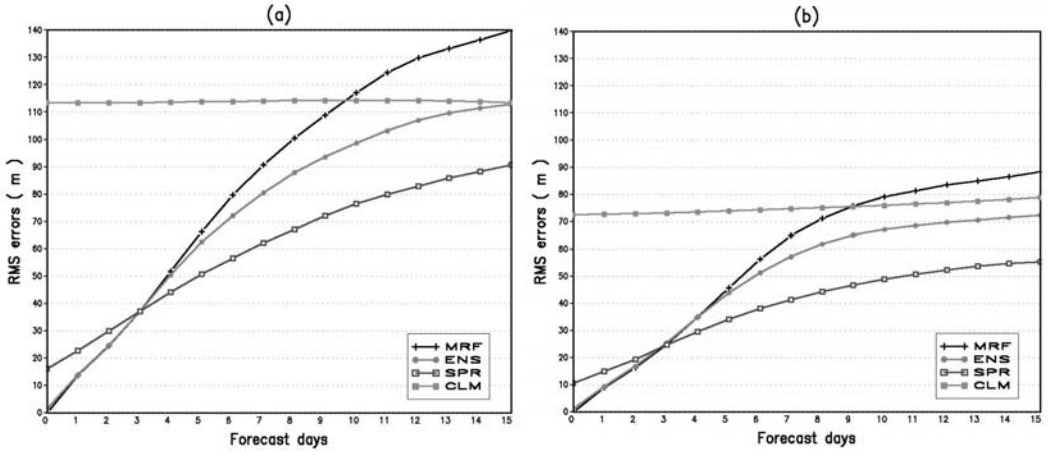

Figure 1. RMS errors for Medium-Range forecasts (MRF) and ensemble mean of northern hemisphere extra-tropical 500 hPa geopotential height: (a) for winter season (December 2001, January and February 2002), (b) for summer season (June, July and August 2002).

2. Ensemble Forecast

2.1. *NCEP Global Ensemble Configuration*

Currently, by adding small initial perturbations to an operational global analysis, the NCEP global ensemble forecasts are generated at 0000 UTC and 1200 UTC every day (Zoltan and Kalnay 1993; Tracton and Kalnay 1993; Zoltan and Kalnay 1997). The NCEP global ensemble forecasts consist each day of 25 individual independent forecasts run out to 16 days lead time, of which 5 members are control forecasts started from unperturbed analyses at 0000 UTC (global forecast system (GFS), which merges the former medium-range forecast (MRF) and aviation forecast (AVN) after October 2002 and ensemble control), 0600 UTC (GFS), 1200 UTC (GFS) and 1800 UTC (GFS), and 20 members are perturbed forecasts at 0000 UTC (10) and 1200 UTC (10) from initial conditions where bred perturbations of the size of estimated analysis uncertainty are added. Four GFS forecasts are integrated at T170/L42 resolution out to 180 hours, and then reduced to T62/L28 resolution out to 16 days. All perturbed members and an ensemble control forecast are run at T128/L28 resolution out to 84 hours, and then reduced to T62/L28 resolution out to 16 days. The breeding cycle is 24 hours in the current NCEP global EPS.

2.2. *NCEP Ensemble Data*

Two public access ftp sites are available to users for NCEP global ensemble forecast data:
1) ftp://tgftp.nws.noaa.gov
2) ftp://ftpprd.ncep.noaa.gov/pub/data/nccf/com/mrf/prod.
All available files are updated daily on these NCEP public access ftp sites.

2.3. *Experimental Data*

The data for this study are the NCEP global ensemble and deterministic forecasts (MRF) of 500hPa geopotential height (period from 01 DEC 2001 to 28 FEB 2002 for a winter season, period from 01 JUN 2002 to 31 AUG 2002 for a summer season), 850hPa temperature (period from 01 DEC 2001 to 28 FEB 2002) and gage 24-hour accumulated total precipitation (period from 01 DEC 2001 to 28 FEB 2002). The gage 24-hour accumulated precipitation data are used as observed precipitation. The climatology of 500hPa geopotential height and 850hPa temperature is from NCEP/NCAR re-analysis. All the model analysis and forecast data are calculated at 2.5 by 2.5 degree resolutions globally except of the precipitation analysis, which is based on 80km ETA grids (Baldwin and Mitchell 1996).

2.4. *Best Ensemble Forecast*

What is the best ensemble forecast? There are many methods to measure ensemble forecasts. One of them is evaluating the difference between RMS errors of ensemble mean and ensemble spread. A perfect ensemble model assumes that an initial perturbation represents growing errors from an analysis, and also assumes that the forecasting model is perfect (Toth

and Kalnay 1993; Toth and Kalnay 1997). Therefore, the spread of the ensemble forecasts should be fully equal to the size of the RMS errors of the ensemble mean when compared to observations. However, in a real application, the forecast model is not perfect and the initial perturbations do not fully represent the analysis growing errors, and so differences appear between RMS errors and the spread. In Figs. 1a and 1b, the differences of the ensemble mean RMS errors (ENS, closed circles) and spread (SPR, open squares) indicate the imperfection levels of NCEP ensemble forecasts relative to the analysis. The measurement would be more accurate if the observations, rather than analyses, could be used directly in the verification. In general, when a spread is smaller than RMS error, it means the ensemble forecast is under representing the model errors and forecast uncertainties; otherwise it is over representing them.

3. Probabilistic Forecast

It is well known that all environmental forecasts are associated with uncertainty and that the amount of uncertainty can be situation dependent. The use of probabilistic forecasts helps in estimating this uncertainty. By considering a wide range of forecasting information, forecasters could subjectively generate probabilistic forecasts by using different methods. For example, probabilistic forecasts could be generated from a set of deterministic forecasts valid at the same time, such as a lag forecast, which was very often used for climate prediction. Meanwhile, probabilistic forecasts could be made by using historical forecasts and observations, or by statistical methods such as the model output statistics (MOS) forecast, or by a multi-model super-ensemble which is based on a set of numerical models that use statistical regression to weight each ensemble member (Krishnamurti et al 1999). In this study, probabilistic forecasts that are based on initial perturbed NCEP global ensemble forecasts from only the same numerical model will be discussed.

There are many applications of probabilistic forecasts for the short/medium range. A "spaghetti diagram" is one such product. Two other major applications that will be described below include the probabilistic quantitative precipitation forecast (PQPF) and the relative measure of predictability (RMOP). PQPF is a product to help with a deterministic quantitative precipitation forecast (QPF). RMOP is a predictability measure for large and small uncertainty.

3.1. *PQPF Forecasts*

The PQPF forecasts have been produced operationally since 1997 in NCEP and are based on global ensemble forecasts using T62/L28 (T126/L28 out to 60 hours after 26 June 2000, T126/L28 out to 84 hours after 9 Jan. 2001) model resolution. The product includes nine threat amounts (0.254, 1.0, 2.54, 5.0, 6.35, 10.0, 12.7, 25.4, 50.8 mm) of 24-hour accumulation precipitation. To generate PQPF forecasts, the number of ensemble members that exceed a given precipitation threat level is divided by the total number of ensemble members at each grid point. For example, if the 24-hour forecasted precipitation amount exceeds 2.54 mm in 7 out of 10 total ensemble members at a particular grid point, then a 70%

probability of rainfall exceeding 2.54 mm (0.1 inches) for a 24-hour period (Zhu and Toth 1998) is assigned to that grid point. The synoptic PQPF maps shown in Fig. 2 show a strong fall/winter storm affecting most of the U. S. East Coast, with another system off the West Coast of Washington State. The four panels show 60-84 hour lead-time PQPF forecasts with four different 24-hour threat amounts of 2.54, 6.35, 12.7 and 25.4 mm. This is a highly predictable storm system (Predictability will be discussed in section 3.2). Additional sets of PQPF maps (not shown here) are very useful as well. For example, we could use a 3-parameter Gamma distribution or L-moment method to generate continued PQPF forecasts. By using continued PQPF forecasts, we may create a different precipitation forecast map for a specified probability (for example 30% probability, 75% probability, etc., Zhu and Toth 2003).

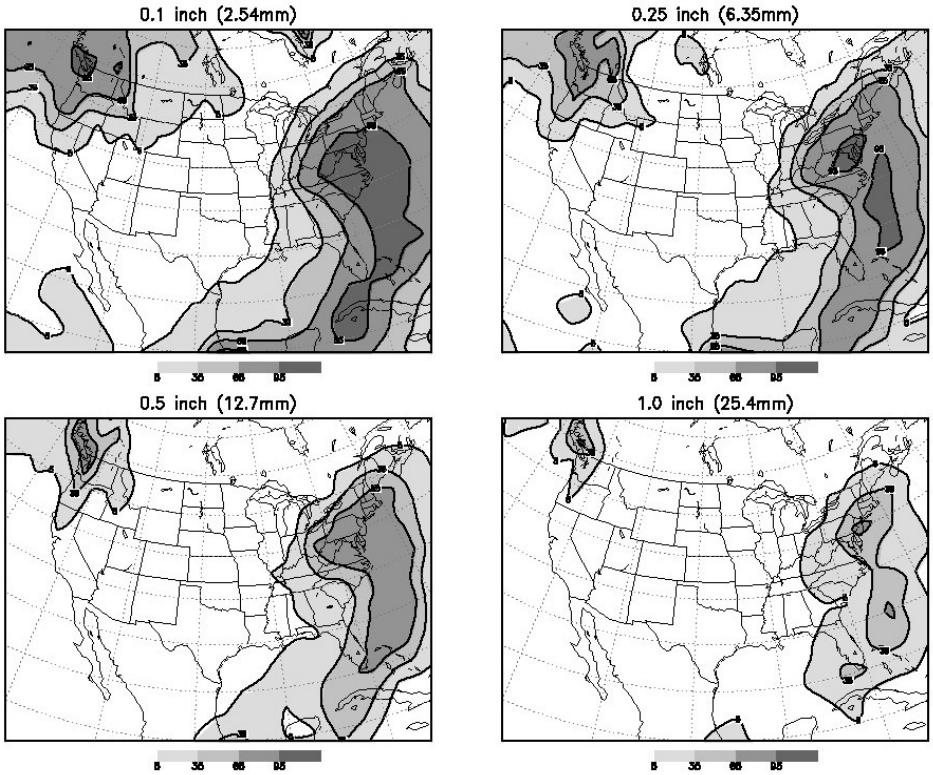

Figure 2. The 3-day lead-time probabilistic quantitative precipitation forecasts for 0.1, 0.25, 0.5 and 1.0-inch thresholds of the 24-hour period ending at 1200 UTC 17 Nov 2002, based on 23-member NCEP global ensemble forecasts. Contour lines are drawn at 5%, 35%, 65% and 95% probability levels.

As we know from long terms statistical objective evaluations, model forecasts are biased in most cases, and ensemble forecasts are biased as well. Especially for precipitation (Zhu and Toth, 1999, Zhu and Toth, 2003), most numerical model forecasts tend to over-forecast small precipitation amounts and under-forecast extreme amounts. In order to make reliable PQPF forecasts, it is necessary to remove the bias (or, first moments) and adjust the spread

(or, second moments) if possible. Real time ensemble based calibrated PQPF and GFS (or MRF) based calibrated QPF forecast maps can be found on the NCEP website at http://wwwt.emc.ncep.noaa.gov/gmb/ens/enshome.html.

3.2. *RMOP Forecasts*

RMOP is a probabilistic measure to assess the flow-dependent uncertainty in a single forecast. Statistics indicate that, for certain cases, 10-13 day lead time forecast skill of the low uncertainty (high predictability, top 10%-15%) cases could be as good as one day forecast skill of the high uncertainty (low predictability, top 10%-15%) cases (Toth et al. 2001). Continuing with the same synoptic case from section 2, Fig. 3 shows the 3-day RMOP forecast map for 500hPa geopotential height, valid at 0000 UTC 17 Nov. 2002. The contours indicate the ensemble mean state, and the colored shaded areas show forecast uncertainties, which are the measures of predictability numbered under the color bar. The reference probability values above the color bar are calibrated by using independent data, and they reflect forecast uncertainty due not only to initial error but also to model related errors. Based on the RMOP measure, Fig. 3 shows a very high predictability (90%) area north of the Gulf of Mexico at 72-hour lead-time, which is associated with the East Coast storm system (see Fig. 2).

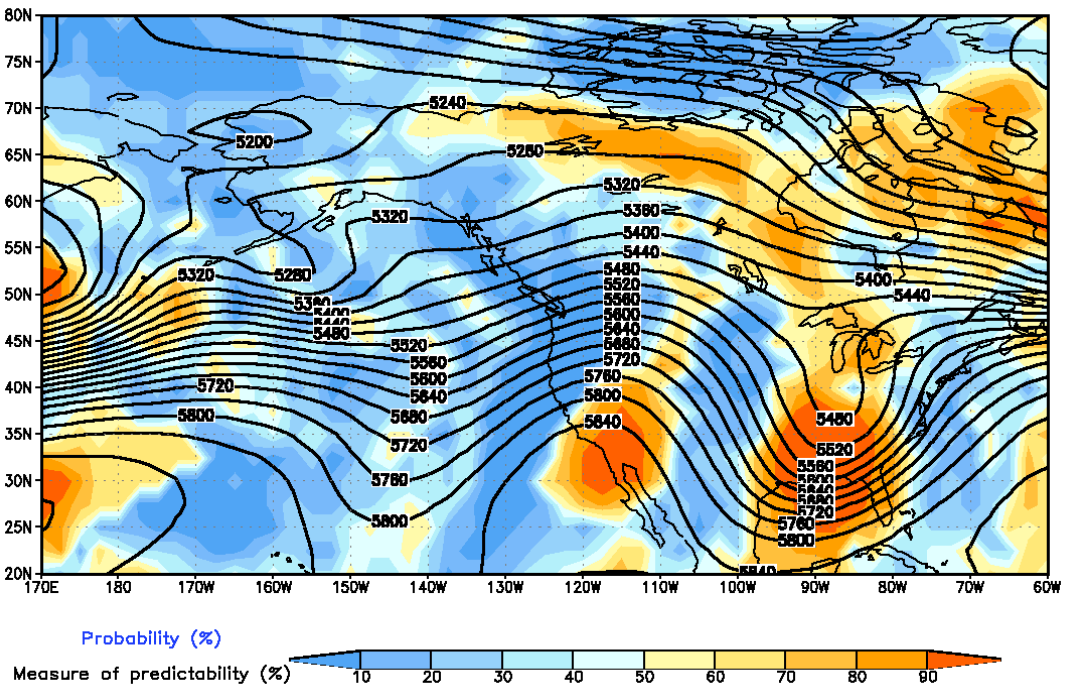

Figure 3 (color): The 3-day lead-time 500hPa geopotential height 10-member NCEP global ensemble mean forecast (contour lines) and associated relative measure of predictability (shaded), valid at 0000 UTC 17 Nov. 2002.

4. Probabilistic evaluation

An operational global ensemble forecast at NCEP and ECMWF is made by using an initial perturbation method, which is based on the assumption of a perfect forecast model. All errors are then considered to result from observations, first guess error and an imperfect data assimilation scheme. From this assumption, clearly the imperfections of the forecast model are not considered to be a source of errors. The forecast uncertainty is fully represented by model initial uncertainty. However, the forecast model is not perfect due to physical parameterization, boundary forces and other factors. Therefore, the forecast uncertainties are truly from both the initial condition (analysis) and the use of an imperfect numerical model. It is very difficult to separate initial errors and model errors quantitatively by using probabilistic evaluations of a model forecast. However, the probabilistic evaluations are the basic tools to measure model forecast uncertainties, which represent model resolution and reliability (Toth et al. 2002). There are many probabilistic evaluation methods that can be used, such as Talagrand distributions, rank probability scores (RPS), relative operating characteristics (ROC) area, information contents (IC), Brier scores (BS), outlier maps, etc. Two methods will be described below which can assess the statistical reliability and resolution of a numerical model (Wilks 1995).

4.1. *Reliability Diagram*

For a given set of forecast probabilities of an event, one can compare with observations and 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 frequency, which would indicate perfect reliability (Fig. 4, diagonal line), in fact, the result is a curve indicating a model that is less than perfectly reliable (Fig. 4, solid line with closed circles). However, the reliability curve could be adjusted to near diagonal (dash line with open circles) by simple calibration by using past forecast and analysis information that is independent of current forecast data. This reliability diagram is based on Northern Hemisphere extratropical 500hPa geopotential height forecasts at day 5 (120-hour). The forecasts and verifying analysis data are separated at each grid point into 10 climatological equally likely intervals (bins). These intervals are defined uniquely for each grid point and each month of the year using NCEP/NCAR re-analysis (Kalnay et al, 1996). From this study (Fig. 4), the forecasts are still not perfect after a simple calibration (dash line, open circles), but they are more reliable when compared to the uncalibrated (raw) forecasts.

4.2. *Economic Value*

A decision maker becomes a user of a weather forecast if he/she alters his/her actions based on forecast information. In 2002, Zhu et al introduced the concept of economic value (EV) of a weather forecast by using cost-loss analysis method and considering user reaction. Different from the reliability measurement that was discussed in section 4.1, the EV is an estimation of

forecast resolution, which is the ability of a forecast system to discern sub-sample forecast periods with different relative frequencies of an event. In this study, we evaluate the potential EV associated with the use of ensemble forecasts (T126/L28 out to 84 hours, then reduced to T62/L28), in terms of equivalent costs, as compared to a higher resolution deterministic forecast (MRF, T170 /L42 out to 7 days, then reduced to T62/L28). For this objective comparison, 850hPa temperatures in the Northern Hemisphere extratropics (20N-80N) have been used. By comparing lower resolution ensemble forecasts and higher resolution MRF forecasts at day 5 (Fig. 5), ensemble forecasts (solid line, closed circles) have better EVs than MRF forecasts (dash line, open circles) over all reasonable and selected cost-loss ratios from 0.01 to 1.0. Based on 3-month statistics, the highest EV is apparently around 1:10 cost-loss range for both the ensemble and deterministic forecasts, which means that the probabilistic forecasts from numerical models are the best fit for user groups who are most comfortable with around a 1:10 cost-loss ratio. Therefore, when considering 1:10 cost-loss ratio only, Fig. 6 shows the EVs for lead time out to 15-day. The ensemble forecasts (solid line, closed circles) have 20% more economic value than the MRF forecasts (dash line, open circles) at early lead times. The EVs of ensemble forecasts are almost double those of the MRF forecasts at longer lead time, but the EV's are already very small and close to zero.

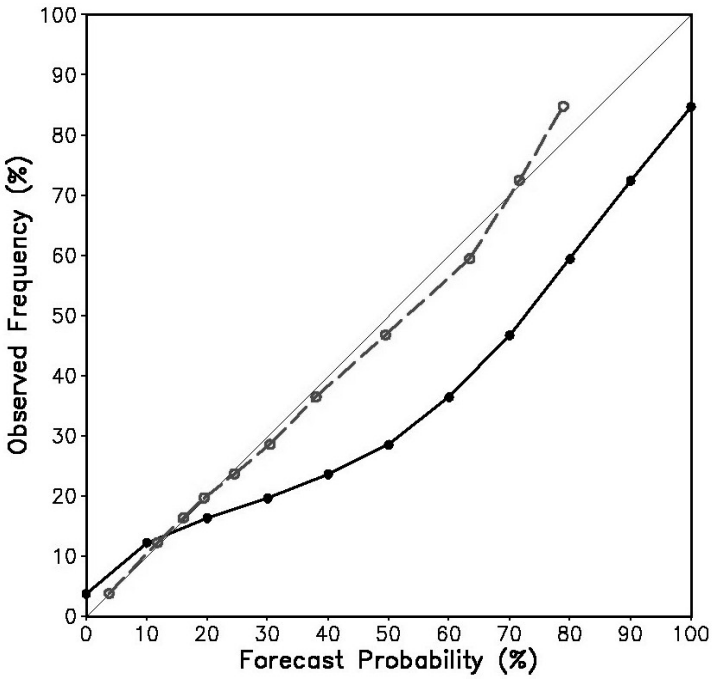

Figure 4. The 5-day lead-time ensemble-based reliability diagram for January 2002 for NH extra-tropical 500hPa geopotential height. Calibrated forecast probabilities (dash line, open circles) are based on observed frequencies associated with the same number of ensemble members falling in a climatologically equally likely bin during 1-20 December 2001. The uncalibrated (raw) forecast is shown on the solid line (closed circles). The vertical axis shows observed relative frequencies.

5. Summary and Conclusions

Ensemble-based probabilistic forecast products can provide dramatic information to users. With the use of such probabilistic forecasts, users have access to significantly more information than is available through a single, deterministic forecast, and this information can be used to make more accurate forecasts and allow users to alter their decisions, especially for extreme events that can be associated with higher and lower uncertainty weather systems. PQPF forecasts and RMOP forecasts are two examples of probabilistic products that were discussed in this paper. The PQPF predicts the future possibilities for many precipitation amount thresholds at all lead times and physical locations, while the RMOP describes the degree of uncertainty associated with the forecasting of future weather systems.

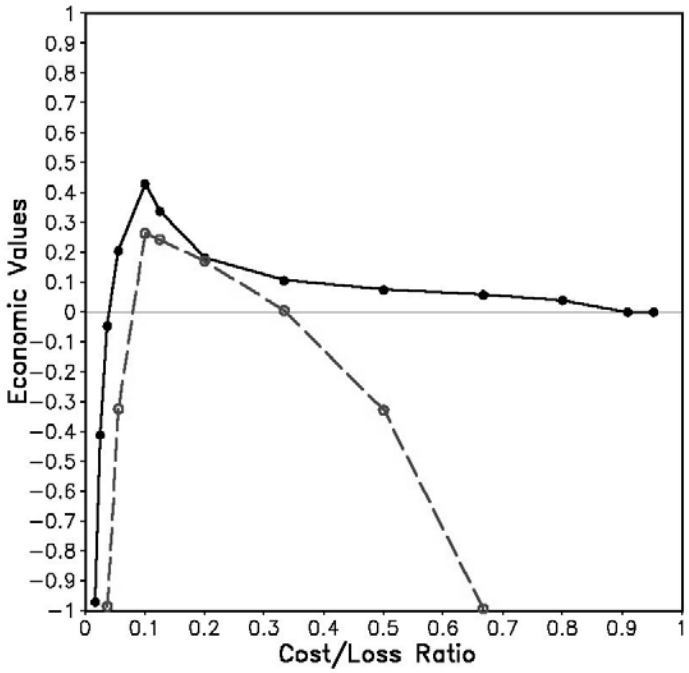

Figure 5. *The economic values for NH extratropical 850hPa temperatures, 5-day lead 10-member NCEP ensemble forecasts (solid line, closed circles) and NCEP deterministic forecast MRF (dash line, open circles) for a winter season (Dec. 2001, Jan. and Feb. 2002). The vertical axis is economic values.*

Reliability and resolution are two major attributes that are considered in the evaluation of these ensemble-based probabilistic forecasts. The reliability indicates how statistically consistent these probabilistic forecasts are with observations, such as through the use of Talagrand distributions, reliability diagrams, outlier maps, etc. The resolution summarizes how much more information the forecasts have with respect to climatology, such as through the use of EVs, ROC areas, ICs etc.

Finally, recent studies indicate that statistical calibration can improve model forecast reliability, although it does not have much effect on the resolution. Since forecast reliability is mostly due to system bias, that bias can be removed by applying historical information from previous forecasts with the same model, if there has been no significant change in the forecasting model in the recent past.

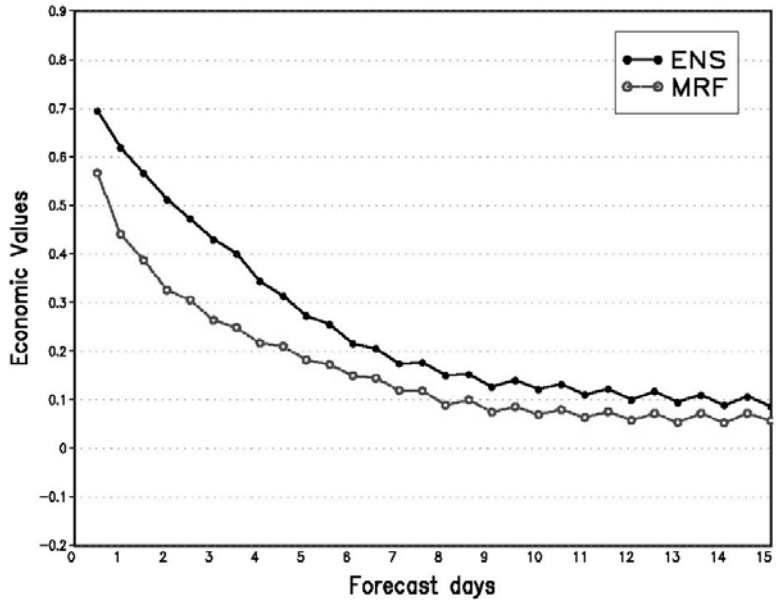

Figure 6: The same as Fig. 5 except for all lead-time forecasts and 1:10 cost-loss ratio only.

Acknowledgments. The author would like to thank Dr. Zoltan Toth of NCEP/NOAA for his encouragement. The comments of Tim Marchok help to improve the presentation of the manuscript. I acknowledge the support of Hua-Lu Pan, Chief of Global Climate and Weather Modeling Branch, EMC and Stephen Lord, Director of EMC.

References:

Baldwin, M.E., and K. E. Mitchell, 1996: The NCEP hourly multi-sensor U. S. precipitation analysis. Preprints, *11th AMS Conf. on Numerical Weather Prediction*, Norfolk, VA. Amer. meteor. Soc. J95-96.

Houtekamer, P. L., and J. Derome, 1995: Methods for ensemble prediction. *Mon. Wea. Rev*., **123**, 2181-2196.

Kanamitsu, M., and Coauthors, 1991: Recent changes in the global forecasting system at NMC. *Wea. Forecasting*, **6**, 425-435.

Krishnamurti, T. N., C. M. Kishtawal, T.LaRow, D. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, **285**, 1548-1550.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc*., **122**, 73-119.

NWS, 1999: *Vision 2005: National Weather Service Strategic Plan for Weather, Water, and Climate Services 2000-2005*. 24 pp. [Available from NWS, 1315 East-West Highway, Silver Spring, MD 20910]

Palmer, T. N., F. Molteni, R. Mureau, R. Buizza, P. Chapelet, and J. Tribbia, 1992: Ensemble prediction. *ECMWF Research Department Tech. Memo*. **188**, 45pp.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc*., **74**, 2317-2330

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

Toth. Z., Y. Zhu, and T. Marchok, 2001: On the ability of ensembles to distinguish between forecasts with small and large uncertainty. Weather and Forecasting, 16, 436-477

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2002: Probability and ensemble forecasts. In: Environmental Forecast Verification: A practitioner's guide in atmospheric science. Edits.: I. T. Jolliffe and D. B. Stephenson. Wiley, in press.

Tracton, M. S., and E. Kalnay, 1993: Ensemble forecasting at NMC: Operational implementation. *Wea. Forecasting*, **8**, 379-398.

Wilks, D. S., 1995: *Statistical Methods in the Apmospheric Sciences*. Academic Press, New York, 467pp.

Zhu, Y., G. Iyengar, Z. Toth, S. Tracton, and T. Marchok, 1996: Objective evaluation of the NCEP global ensemble forecasting system. *15th AMS Conf. on Weather Analysis and Forecasting*. Norfolk, VA. Amer. Meteor. Soc. J79-82.

Zhu, Y., Z. Toth, E. Kalnay, and S. Tracton 1998: Probabilistic Quantitative Precipitation Forecasts based on the NCEP global ensemble. *Special Symposium on Hydrology*, Phoenix, AZ. Amer. Meteor. Soc. J8-11.

Zhu, Y., and Z. Toth, 1999: Objective Evaluation of QPF and PQPF Forecasts Based on NCEP Ensemble, Preprints, *Third International Scientific Conference on the Global Energy and Water Cycle*, 16-19 June 1999 Beijing China, 47-48.

Zhu, Y., and Z. Toth, 2001: Extreme weather events and their probabilistic prediction by the NCEP ensemble forecast system. Preprints, *Symposium on Precipitation Extremes: Prediction, Impact, and Responses*, Albuquerque, NM. Amer. Meteor. Soc., 82-85.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne 2002: On the economic value of ensemble based weather forecasts. *Bull. of Amer. Meteor. Soc*., **83**, 73-83.

Zhu, Y., and Z. Toth, 2003: A synoptic evaluation of ensemble based probabilistic quantitative precipitation forecasts. Submit to *J. Hydrometeorology*.