ADVANCES IN ATMOSPHERIC SCIENCES, VOL. 22, NO. 6, 2005, 781–788

Ensemble Forecast: A New Approach to Uncertainty and Predictability

Yuejian ZHU*

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

(Received 31 January 2005; revised 29 June 2005)

ABSTRACT

Ensemble techniques have been used to generate daily numerical weather forecasts since the 1990s in numerical centers around the world due to the increase in computation ability. One of the main purposes of numerical ensemble forecasts is to try to assimilate the initial uncertainty (initial error) and the forecast uncertainty (forecast error) by applying either the initial perturbation method or the multi-model/multi-physics method. In fact, the mean of an ensemble forecast offers a better forecast than a deterministic (or control) forecast after a short lead time (3–5 days) for global modelling applications. There is about a 1–2-day improvement in the forecast skill when using an ensemble mean instead of a single forecast for longer lead-time. The skillful forecast (65% and above of an anomaly correlation) could be extended to 8 days (or longer) by present-day ensemble forecast systems. Furthermore, ensemble forecasts can deliver a probabilistic forecast to the users, which is based on the probability density function (PDF) instead of a single-value forecast from a traditional deterministic system. It has long been recognized that the ensemble forecast not only improves our weather forecast predictability but also offers a remarkable forecast for the future uncertainty, such as the relative measure of predictability (RMOP) and probabilistic quantitative precipitation forecast (PQPF). Not surprisingly, the success of the ensemble forecast and its wide application greatly increase the confidence of model developers and research communities.

Key words: ensemble forecast, predictability, uncertainty

1. Introduction

In the past decade, the methodologies followed at the National Centers for Environmental Prediction (NCEP) of the National Weather Service of the United States, the European Centre for Medium-Range Weather Forecasts (ECMWF) and the Canadian Meteorological Centre (CMC) of the Meteorological Service of Canada have been developed to simulate the effect of initial and model uncertainties onto the forecast errors. In early studies, the characteristics of these three global ensemble prediction systems (EPS) have been discussed, and objective evaluations have been made by using the three ensemble forecasts for a 3-month period; May, June, and July 2002 (Buizza et al., 2005). The probabilistic applications, the probabilistic evaluations and the differences between deterministic and ensemble forecasts from the NCEP EPS system have been presented in past years (Zhu et al., 1996, 2002; Zhu and Toth, 1999; and Zhu, 2004). In the present study, experiments were performed based on the most recent global ensemble forecasts (June, July, and August 2004) from the world numerical centers to capture the improvements made in the numerical models and ensemble techniques in recent years. Furthermore, synoptic examples of probabilistic quantitative precipitation forecasts (PQPFs) from three numerical prediction centers have been exhibited side by side to allow us to compare each one. The multi-center ensembles, which are the combined NCEP EPS and ECMWF EPS, are studied to demonstrate the new approach of the ensemble method, which is from different initial condition generation methods, different assimilation systems (initial conditions), different forecast models (dynamics and physical parameterizations) and different model resolutions (vertical and horizontal resolutions). The importance of all these studies is not to rank the performance of the ensemble systems, but to identify possible reasons for superior/inferior performance, thus drawing a guideline

*E-mail: Yuejian.Zhu@noaa.gov

for future ensemble development and improving the ensemble forecast system and predictability.

This paper will discuss the importance of the ensemble forecast in the next section. After that, the methods of the ensemble forecast will be briefly reviewed, and the objective evaluations of the forecasts from the improved, state-of-the-art ensemble systems will be presented in terms of deterministic (and/or ensemble mean) and probabilistic (distribution) concepts in section 4. In section 5, the multiple applications of the ensemble forecast will be introduced. The experimental multi-center ensemble forecast will be discussed in section 6 through selected combinations from comparable ensemble systems. In additional to the discussion of the forecast skill in section 4, the effect of model initial conditions and model resolutions will be investigated from one-year statistics.

2. Why do we need ensemble the forecast?

There are two main reasons that emphasize the importance of the ensemble model forecast. One is the forecast error (uncertainty), which comes from each process of a numerical weather prediction system, such as observation and data collection (observation system), data assimilation (analysis system) and forecast model (dynamical process, computation, physical parameterization, etc.). Early studies (Lorenz, 1969, 1982) suggested that the initial error could grow very fast into the different scales no matter how small the initial error. In fact, the forecast error will increase continually with the model integration before it is saturated. The optimum solution to capture and reduce this forecast error (uncertainty) is to use an ensemble forecast instead of a single (deterministic) forecast, because the ensemble forecast produces a set of randomly-equally-likely (independent) solutions for the future. The diversity of these solutions, which is called the forecast spread, mostly represents the forecast uncertainty. The relationship between ensemble spread and ensemble mean error (uncertainty) has been discussed in an earlier study (Zhu, 2004) and will be discussed again in this paper. The perfect ensemble prediction system is expected to have a similar spread to the ensemble's mean error (or high correlation between the ensemble spread and ensemble mean error) in the long term statistics. How much does the ensemble spread represent the forecast uncertainty in the real atmosphere? This cannot be answered quantitatively. It depends on the sizes of the spread and the error, the distribution of the error, etc. In fact, the skill of the ensemble forecast is greatly improved when comparing the ensemble mean forecast to a deterministic forecast after a short lead-time. The ensemble mean forecast for a short lead-time is degraded due to the introduction of initial perturbations (error) for both the NCEP EPS and the ECMWF EPS. The other reason is predictability. Knowing the future has always been a practical and spiritual need for people. The ultimate goal of all scientific work has also been successful prediction. The success of our prediction efforts depends on two main factors: (1) our understanding and knowledge of natural processes, and (2) the nature of these processes to be predicted. Increases in forecast predictability always correspond to decreases in forecast uncertainty. The reduction of forecast error from the ensemble forecast greatly increases the predictability. In addition, when considering the forecast itself and the user community, one of the goals of the United States National Weather Service for 2000–2005 is to provide weather, water and climate forecasts in probabilistic terms by the year 2005 (NWS, 1999), which is most achievable and practical with an ensemble forecast. In the past, there were many methods to generate the probabilistic forecast, but the ensemble model forecasts can achieve this goal easily and accurately. As expected, the probabilistic forecast, such as a spaghetti diagram (to describe uncertainty), the PQPF (to tell the probabilistic forecast) (Zhu et al., 1998; Zhu and Toth, 1999; Zhu, 2004, 2005), the RMOP (related to predictability) (Toth et al., 2001), and the ensemble spread (similar to spaghetti diagram, but more completely), has been more popular to users and the public in recent years.

3. Methodologies of the ensemble forecast

As noted earlier (Buizza et al., 2005), there are two major methods to generate an ensemble model forecast in use around the world in meteorological centers. One of them is the initial perturbation method, which adds small perturbations to the initial analysis, such as NCEP's breed mode method (Toth and Kalnay, 1993; Tracton and Kalnay, 1993; Toth et al., 1997) and ECMWF's singular vector method (Palmer et al., 1992; Molteni et al., 1996). The NCEP and ECMWF methods assume the forecasting model is perfect, and they assimilate the initial (observation and analysis/data assimilation) uncertainty by using a small and random initial perturbation. There are 10 (5 pairs) ensemble runs for each assimilation cycle (0000 UTC, 0600 UTC, 1200 UTC and 1800 UTC) in NCEP, and 50 (25 pairs) ensemble runs for 1200 UTC only in ECMWF. However, another set of ensemble forecasts is produced by using different numerical models (the spectrum model and the grid model) and different physical packages in the CMC (Houtekamer and

Derome, 1995; Houtekamer et al., 1996). Eight different physical packages are used in CMC's global EPS. There are, in total, 16 (2 models, 8 different physical packages) ensemble runs from 0000 UTC. These are used to assimilate the initial (by different models) and forecast (by different physical schemes) uncertainties. Moreover, in both research and development centers, many other ensemble forecasts have been studied from statistic post processes such as super-ensembles (Krishnamurti et al., 1999), the poor-man's ensemble (Ebert, 2001), Monte Carlo or lagged average forecast (LAF) ensembles for climate study, etc. In section 6, we will discuss the multi-center ensemble forecast by the combination of the NCEP EPS and the ECMWF EPS, as well.

4. The skill of an ensemble forecast

Before we discuss the skill of the ensemble forecast, let us review the effect of model initial conditions and model resolutions. By running a one-year statistical average (1 June 2003– 31 May 2004) of verification scores, the pattern anomaly correlations (PAC) of the NCEP Global Forecasting System (GFS: high resolution control, T254L64 from 0–84 hours, T170L64 from 84–180 hours, T126L64 from 180–384 hours), the NCEP ensemble control (CTL: low resolution control/ensemble control, T126L28 from 0–180 hours, T62L28 from 180–384 hours) and the NCEP 10 ensemble members (5 pairs of initial perturbations with the same resolution as the ensemble control) are calculated. When comparing GFS and CTL to 10 individual ensemble members, a score of 100% will be awarded if GFS/CTL is better than all individual ensemble members, otherwise, 0% will be given if GFS/CTL is worse than all members, 50% will be added if GFS/CTL is better than 5 out of the 10 ensemble individual members (randomly). The result is shown in Fig. 1 for up to 15 days lead-time, 500 hPa geopotential height for the Northern Hemisphere latitude band (20°–80°N). For the short lead-time (0–96 hours), the high resolution GFS is the best, and the individual ensemble perturbation forecasts are far behind either the GFS (due to the resolution and initial error) or CTL (due to the initial error). After a short lead-time (120 hours), the model resolution is not as important as the first 96 hours to improve the model forecast skills; as unexpected, CTL is slightly better than GFS from 144 hours to 264 hours lead-time in this experiment period. The differences between GFS/CTL and individual ensemble members are reduced. After 11 days lead-time, the PAC is very low (less than 50%) which will not be considered as a skillful forecast for the synoptic system, and thus GFS is better than CTL again. Interestingly, both GFS and CTL are still better than any of the individual ensemble members. As noted earlier, the difference between CTL and the ensemble members is only the initial conditions. The difference between CTL and GFS is only the resolution. The 50% line is a reference to consider the equality of GFS/CTL and the 10 individual ensemble members. Based on the results presented in this section, we should point out that both the resolution and the initial conditions are very important to model forecasts. The resolution plays a key role in the success of the short-range forecasts while the influence of the resolution is much smaller than that of the initial conditions for medium-range forecasts.

The forecast skills could simply be measured in terms of pattern anomaly correlation (PAC) scores (which depends on climatological information) and root mean square (RMS) errors of the 500-hPa (or other levels) geopotential height (or other variables) by considering any deterministic (control) forecast and ensemble's mean. The assessment of the past status (summer of 2002) for three major global ensemble prediction systems was presented by Buizza et al. (2005). The objective evaluation of the present status is done here by using similar methodologies. Figure 2 shows the Northern Hemisphere extra-tropical (20°N–80°N) 500-hPa geopotential height PAC scores of three different EPSs' means (solid lines, considering the first

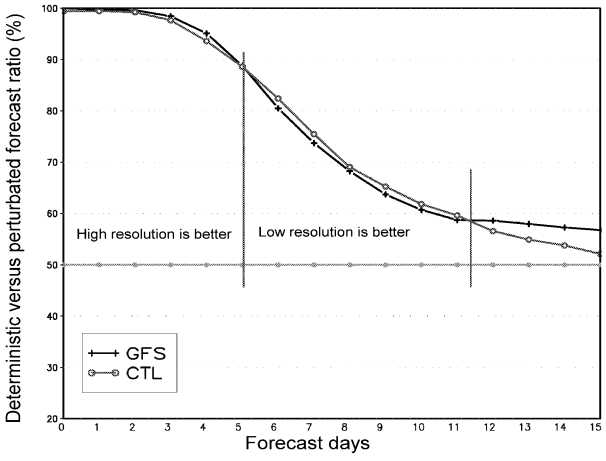

Fig. 1. 1 June 2003–31 May 2004 (1-year) daily PAC scores for the NCEP/GFS (high resolution control, crosses) and the NCEP ensemble control (the same resolution as the ensemble members, open circles) versus the NCEP 10 individual ensemble members. The 50% line is a reference (closed circles) to represent that the GFS/CTL is at the median of the ensemble members. Values (PAC scores) refer to the 500-hPa geopotential height over the Northern Hemisphere latitude band 20°–80°N.

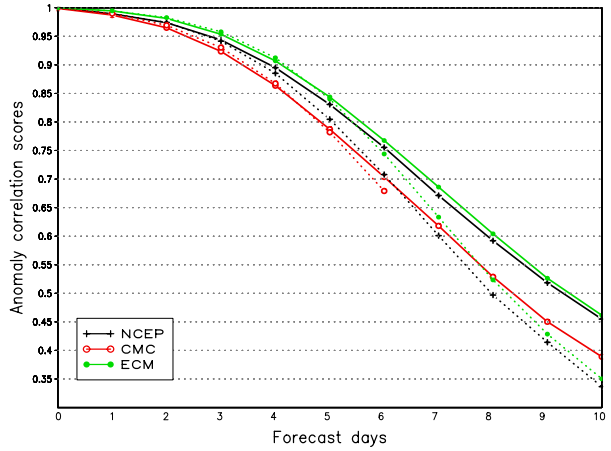

Fig. 2. June–August 2004 average PAC scores for the control (dotted lines) and the ensemble means (solid lines) of the NCEP-EPS (crosses), CMC-EPS (open circles) and ECMWF-EPS (closed circles). Values refer to the 500-hPa geopotential height over the Northern Hemisphere latitude band 20°–80°N.

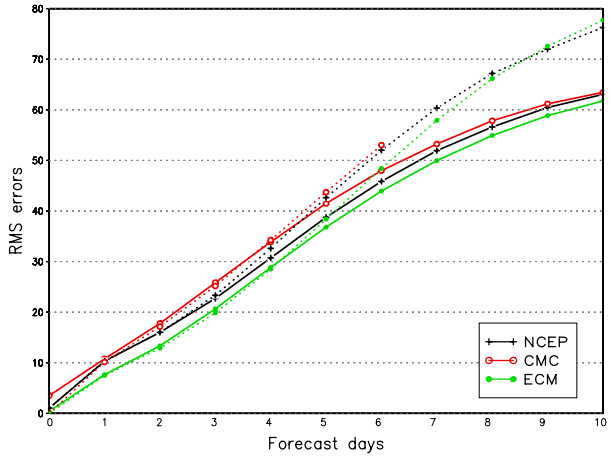

Fig. 3. June–August 2004 average RMS errors for the control (dotted lines) and the ensemble means (solid lines) of the NCEP-EPS (crosses), CMC-EPS (open circles) and ECMWF-EPS (closed circles). Values refer to the 500-hPa geopotential hight over the Northern Hemisphere latitude band 20°–80°N.

10 ensemble members only for each center, using NCEP/NCAR reanalysis data as the climatology) compared to their own deterministic/control forecast (dotted lines) for June–August 2004. The verified analysis is taken from their own data assimilation systems in this experiment. Similar results were obtained by Buizza et al. (2005) except for the improvement of all three systems. The means of the ensemble forecasts do not show any advantage for the first 3 (up to 5, depends on the model and season) days due to the introduced initial errors by NCEP and ECMWF. There is a similar result in CMC's ensemble for a very short lead-time because it uses one verified analysis for two different initial conditions (note: only one analysis is available to the verification). However, after 3 (up to 5) days of lead-time, the ensemble means have a 6-hour to 24-hour (or longer) advantage than their own deterministic forecast. Unfortunately, there are only 6 days of lead-time available for the evaluation to CMC's control forecast. The differences between the ensemble mean and their own deterministic forecasts are very similar for all three EPSs. Therefore, from these experimental results, we see that the improvement of the ensemble forecast mostly depends on its analysis and forecast model. The skill of ECMWF's deterministic forecast is slightly better than the other deterministic forecasts, and the PAC scores of the ECMWF ensemble mean lead for all of these forecasts, too. Of course, the costs of these three EPSs are slightly different. Less computation time is needed for NCEP's breeding method; in contrast, it is more difficult to maintain and develop CMC's method if resources are limited.

When considering a skillful forecast, usually defined as 65% and above for PAC scores based on the synoptic scale forecast (short-medium-range), the NCEP ensemble mean offers 7 days and 8 hours of useful forecast instead of NCEP GFS (deterministic) which has 6 days and 14 hours of skillful forecast (see Fig. 2) by using approximately the same amount of computation resources. There is an 18 hour improvement when considering the ensemble mean only in this three-month summer period, which is a huge gain compared to improvements from the observation system, data assimilation and forecast models.

The RMS error is another measurement, which does not depend on the climatology. The results of the same period (June–August, 2004) for the Northern Hemisphere (NH) 500-hPa geopotential height are shown in Fig. 3. The solid lines are for the ensemble means, dotted lines are for the ensemble controls (or deterministic forecasts). ECMWF's control forecast has a smaller error for first 4 days, after that, ECMWF's ensemble mean is better than the ensemble control. It is interesting to note that in the NCEP forecast, for either the ensemble mean or the ensemble control, the forecast errors increase very rapidly in the first 24 hours, after which the error growth rates are very similar (or close) to ECMWF's. Does this indicate something we need to work on in the future?

Another important way to evaluate ensemble forecasts is to use probabilistic methods, such as the Brier score (BS) (considering both the resolution and reliability), rank probability score (RPS), potential economic value (EV), hitting rate (HR) and false alarm rate (FAR), relative operating characteristics (ROC)

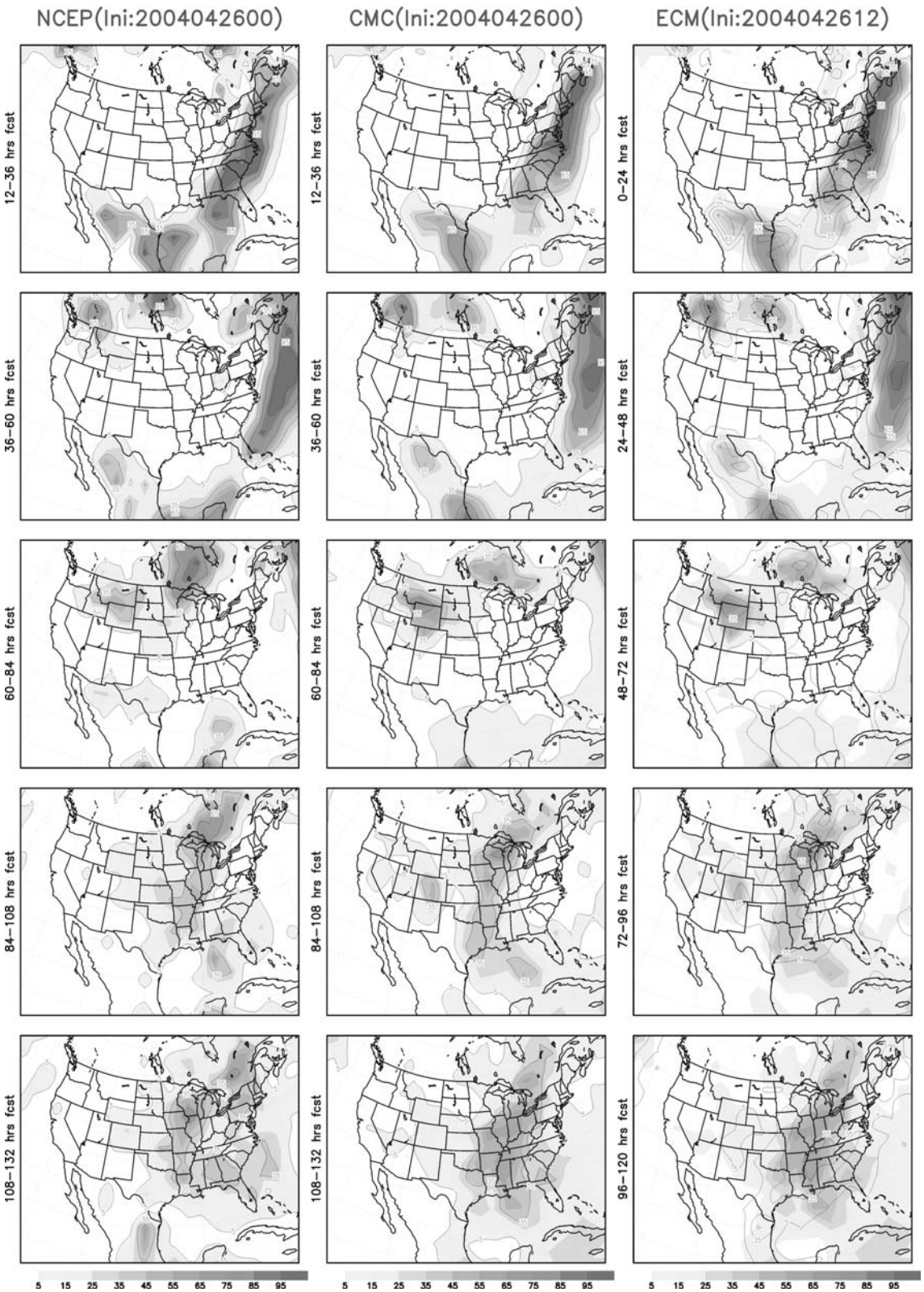

Fig. 4. The probabilistic quantitative precipitation forecast (PQPF) for 24-hours amounts exceeding 6.35 mm. The initial times are 0000 UTC 26 April for the NCEP 11 ensemble members and CMC 17 ensemble members, and 1200 UTC 26 April for the ECMWF first 11 ensemble members. The gray-scale bar indicates the probabilities from 0 to 100 in percentage.

area, etc.. (Wilks, 1995; Zhu et al., 1996, 2002; Toth et al., 2003; Zhu, 2004; and Buizza et al., 2005). In considering distributions, the Talagrand (or histogram) distribution, outliers and consistency are very useful measures to evaluate the EPS, too (Toth et al., 2003). However, all these probabilistic/distribution evaluations cannot be compared to their own deterministic forecast directly.

5. Applications of the ensemble forecast

Many new products have been generated since global ensemble forecasts started. The typical example of an early ensemble graphical application is the spaghetti diagram (Toth et al., 1997). Later, the PQPF for different threats [such as 0.1 mm (24 h)$^{-1}$, 2 mm (24 h)$^{-1}$ and so on] has been used in operational applications since 1997 at NCEP (Zhu et al., 1998; Zhu and Toth, 1999; Zhu, 2004). The calibrated QPF and PQPF were implemented on 4 May 2004, and these applied bias removal techniques (Zhu, 2005). The ensemble mean and spread have been the standard products from NCEP since 2000. Recently, precipitation-type-based probabilistic forecasts have been implemented for every 6-hour lead times, where the products include probabilistic quantitative rain forecasts (PQRF), probabilistic quantitative snow forecasts (PQSF), probabilistic quantitative freezing rain forecasts (PQFF) and probabilistic quantitative ice pellet forecasts (PQIF). The relative measure of predictability (RMOP) values have been calculated globally since 2000 (Toth et al., 2001; and Zhu, 2004). Figure 4 shows a synoptic example of a probabilistic quantitative precipitation forecast (PQPF) of the North American (NA) area for three comparable global ensemble systems (NCEP, CMC and ECMWF EPSs). The initial time is 0000 UTC 26 April 2004 for NCEP and CMC, and 1200 UTC 26 April for ECMWF. The lead times are 12–36 (0–24), 36–60 (24–48), 60–84 (48–72), 84–108 (72–96) and 108–132 (96–120) hours for NCEP and CMC (ECMWF). The contour levels are for 5%, 35%, 65% and 95% respectively. The forecasts of the main features are very close to each other for up to 5 days. When verified with the observations (from rain gauges, not shown), we find that all of them make very good forecasts. Through a number of investigations, we are expecting to have more joint ensemble products in the future through the North American Ensemble Forecast System (NAEFS) project which was endorsed by the National Weather Service of the United States, the Meteorological Service of Canada and the National Meteorological Service of Mexico in November 2004.

6. Multi-center model ensemble

The multi-model super-ensemble for weather and climate applications was discussed many years ago (Krishnamurti et al., 1999). The study was mainly focused on climate prediction by applying statistical methods and training data. After that, a poor-man's ensemble was investigated to predict the PDF of a 1–2-day precipitation forecast (Ebert, 2001) which used a set of individual models from several operational center. Therefore, the ensemble size was limited. The experiment in this study tries to combine two similar ensemble systems from NCEP and ECMWF. The advantages of this combination might be the improvement of the forecast skill, less computation cost, a larger ensemble size and so on.

First, comparing two ensemble systems from NCEP and ECMWF, both of them use the initial perturbation method and are available for the 1200 UTC initial runs, and the overall skills are very comparable to each other. After reviewing the NCEP and ECMWF EPSs, the experiments are designed to combine the two ensembles by selecting (1) 10 members from NCEP's first 6 members and ECMWF's first 4 members (verifying this against NCEP analysis) and (2) 10 members from ECMWF's first 6 members and NCEP's first 4 members (verifying this against ECMWF analysis) in order to match/compare the NCEP 10-member ensemble at the 1200 UTC cycles. Figure 5 shows the PAC scores of the ensemble mean for up to 10 days in June–August 2004 of the NH extratropicl (20°N–80°N) 500-hPa geopotential height. The two new, combined ensembles (closed circles and open squares) are better than either the NCEP or ECMWF original 10-member ensembles. For example, there are about 8 hours of improvement for a 5-day forecast when comparing the NCEP ensemble (crosses) to the second new ensemble (open squares). Figure 6 is the same as Fig. 5 but for the ROC area verification. The ROC area is calculated based on an accumulative hit rate and false alarm rate of 10 climatologically-equally-likely intervals (Zhu et al., 1996; Mason, 2003). Both of the new ensembles are better than the NCEP ensemble for all lead times. The second new ensemble is better than the ECMWF ensemble in all ways except for its longer lead time. The first new ensemble has a mixed improvement with a short lead time but is slightly worse than the ECMWF ensemble in this experiment. A tentative explanation for this result is that the systematic errors (bias) are still in both forecasts and analyses. The probabilistic skills (include

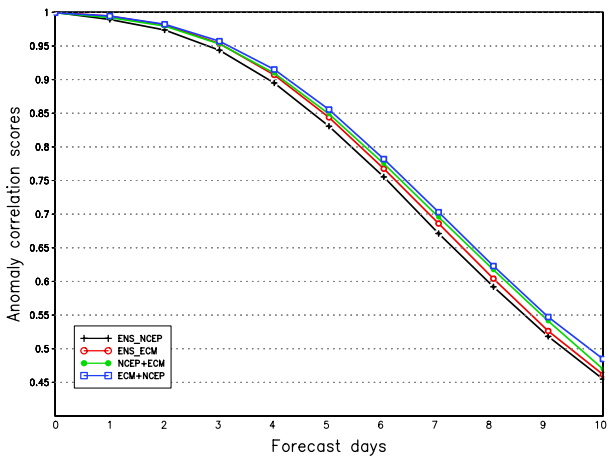

Fig. 5. June–August 2004 average PAC scores for the 10-member ensemble means of NCEP (crosses, the same as Fig. 2), ECMWF (open circles, the same as Fig. 2), NCEP (6 members) + ECMWF (4 members) (closed circles) and ECMWF (6 members) + NCEP (4 members) (open squares). Values refer to the 500-hPa geopotential hight over the Northern Hemisphere latitude band 20°–80°N.

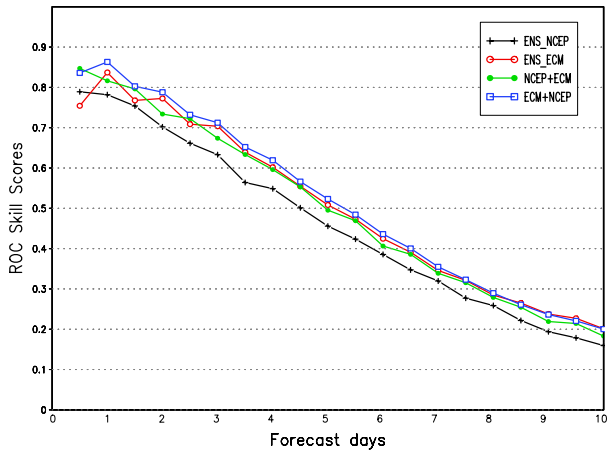

Fig. 6. June–August 2004 average ROC skill scores for the 10-member ensemble distributions of NCEP (crosses), ECMWF (open circles), NCEP (6 members) + ECMWF (4 members) (closed circles) and ECMWF (6 members) + NCEP (4 members) (open squares). Values refer to the 500-hPa geopotential hight over the Northern Hemisphere latitude band 20°–80°N.

resolution and reliability) of the new ensembles should be improved by removing bias latitude band 20°–80°N. (or related pre-processing) before they are combined. It is still questionable for the combined ensemble whether the original ensemble systems are very different, such as the system design, forecast skill, and spread. Further studies are required to answer this question.

7. Discussion and conclusion

Let us discuss the relationship between the RMS error of the ensemble mean and the ensemble spread. The RMS error is the distance measured from the ensemble mean (of the forecasts) to the truth (analysis). The spread measures the distance from the ensemble mean (of the forecasts) to each individual ensemble member. Apparently, we would expect a perfect ensemble forecast to have an ensemble spread equal (or close) to the RMS error, which means that the ensemble spread will maximally represent the forecast uncertainty. But in fact, our current ensemble prediction systems have less spread than the RMS error for medium- and extended-range forecasts (Zhu, 2004; Buizza et al., 2005), which means that the ensemble forecasts are insufficient to capture reality systematically, or that none of them is able to simulate all sources of forecast uncertainties for this chaotic system. In the practical, the medium-range forecast could be improved by reducing the RMS error and increasing the spread. The recent experiments indicate (not shown here) that the RMS error can be reduced by using statistical calibration (or bias correction), while the spread can be increased by introducing stochastic processes (or other techniques) in the NWP model.

The skill of the ensemble forecast relatively depends on the quality of our observing system, the data assimilation system (analysis/initial conditions) and the forecast model (dynamics, physical processes, etc.). When the importance of developing an ensemble prediction system is emphasized, one should not forget to pay attention to improving the basic numerical weather prediction (NWP) system, which includes the data assimilation and the forecast model. The model resolution is a key in making a superior forecast for the first 1-6 days. After that, the high resolution does not give much advantage due to the lack of predictability by the nonlinear interactions, physical parameterizations, etc. The initial conditions are the most important factor in making a good forecast in the short, medium and extended ranges.

It is very difficult to simulate all possible errors (or uncertainties) perfectly in present-day EPSs. The multi-model and multi-analysis approaches may be better, but their costs of maintenance and development are too expensive for any numerical center. Therefore, our tentative conclusions could be the following: (1) the effort to improve the analysis and the forecast model could benefit both the ensemble and deterministic forecasts; (2) ensemble post-processing

is another way to enhance a forecast skill by using statistical bias correction; (3) a combined multi-center, multi-model ensemble with bias correction could possibly approach the goal closely in the future.

Acknowledgments. The author would like to thank Dr. Zoltan Toth of NCEP/NOAA for his encouragement. I acknowledge the support of Stephen Lord, Acting Chief of the Global Climate and Weather Modeling Branch, EMC, and Director of EMC. Helpful comments and suggestions from the two anonymous reviewers resulted in an improved manuscript.

REFERENCES

Atger, F., 1999: The skill of ensemble prediction systems. *Mon. Wea. Rev.*, **127**, 1941–1953.

Buizza, R., P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu, 2005: Assessment of the status of global ensemble prediction. *Mon. Wea. Rev.*, **133**, 1076–1097.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480.

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

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, **124**, 1225–1242.

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.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289–307.

Lorenz, E. N., 1982: Atmospheric predictability experiments with a large numerical model. *Tellus*, **34**, 505–513.

Mason, Ian B., 2003: Binary events. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, 37–76.

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. 24pp. [Available from NWS, 1315 East-West Highway, Silver Springs, 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, S. M. Tracton, R. Wobus, and J. Irwin, 1997: A synoptic evaluation of the NCEP ensemble. *Wea. Forecasting*, **12**, 140–153.

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, 2003: Probability and ensemble forecasts. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, 137–163.

Toth, Z., O. Talagrand, and Y. Zhu, 2005: The attributes of a forecast system: A general framework for the evaluation and calibration of weather forecasts. *Predictability of Weather and Climate,* T. N., Palmer, and R. Hagedorn, Eds., Cambridge University Press (in print).

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 Atpmospheric 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*, Beijing, 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., 2004: Probabilistic forecasts and evaluations based on a global ensemble prediction system. *Vol. 3-Observation, Theory, and Modeling of Atmospheric Variability* World Scientific Series on Meteorology of East Asia, 277–287.

Zhu, Y., 2005: Calibration of QPF/PQPF forecast based on the NCEP global ensemble. San Diego, CA. Amer. Meteor. Soc., J3.3.

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