What is a reasonable ensemble size for operational ensemble forecast?

Juhui Ma* and Yuejian Zhu
*UCAR visiting scientist at Environmental Modeling Center, NCEP/NWS
Environmental Modeling Center, NCEP/NWS, Camp Springs, MD 20746
E-Mail: Juhui.Ma@noaa.gov

1. Introduction

Increases of model resolution and ensemble size are beneficial for the improvement of ensemble performance (Du et al., 1997; Buizza and Palmer, 1998a; Buizza et al., 1998b; Buizza et al., 1999; Richardson, 2001; Mullen and Buizza, 2002). However, the limited computational resources constrain model resolution and ensemble size. Therefore, when design an effective operational ensemble prediction system, there are two main questions we are looking for answers which are 1). how many ensemble members we need to have better representing forecast uncertainties with limited computational resources? And 2). what is a relative impact for increasing model resolution and increasing ensemble size? In this study, the two questions above will be analyzed by using Lorenz 96 model and NCEP GEFS.

The famous Lorenz models are similar to other nonlinear dynamical models of atmospheric system. Increasing the ensemble size for expanding the sample of numerical model's phase space is extremely expensive and complex for operational forecast models, however, it is feasible in the Lorenz model due to its simple dynamical system. The experiment with large ensemble size attained from Lorenz model can give a theoretical instruction for this study with less cost.

After all Lorenz model is a simple experiment to assimilate the complexity of real atmospheric system, so the relative small ensemble sizes should be applied to complicated operational ensemble forecast system to verify the conclusion obtained from using Lorenz 96 model. In this study, NCEP operational GEFS is employed and ensemble size will increase to 80-member.

2. Experimental design

2.1 Lorenz 96 model and its application

a. Lorenz 96 model

The Lorenz 96 model (Lorenz, 1996) is given by the following differential equations

$$\frac{dX_i}{dt} = (X_{i+1} - X_{i-2})X_{i-1} - X_i + F ,$$

Where $i = 1,2,...,N$ with cyclic boundary condition, i.e., $X_{-1} = X_{N-1}$, $X_0 = X_N$ and $X_1 = X_{N+1}$. The magnitude of the forcing is set to $F = 8$ which is well into the chaotic regime (Lorenz, 1996) and the system's size is chosen $N = 1000$. A fourth-order Runge-Kutta integration scheme is employed with a fixed time step of 0.05 which corresponds to approximately 6-hour in the real atmosphere. The first 1000 time steps are used for the system to spin-up.

b. Initial perturbation method

The analysis is obtained by using ensemble Kalman filter (EnKF) method (Evensen, 1994) which also provides analysis error covariance for the Ensemble Transform with Rescaling (ETR) initial perturbation method.

Initial perturbations are generated by using ETR based perturbation (Wei et al., 2006 and 2008) with 10, 20, 40, 60, 80, 100 and 200 ensemble members in this experiment. In ETR scheme, the basic perturbations for best analysis are generated from 6-hour forecasts

on through an ensemble transformation $\mathbf{T}$ as follows

$$\mathbf{Z}^a = \mathbf{Z}^f \mathbf{T}.$$

And then the perturbations should proceed to be centralized and rescaled.

2.2 Real atmospheric (NCEP GEFS) model and its application

The current NCEP operational GEFS (based on GFS v8.0) runs 20 ensemble member forecasts and one control forecast at T190 horizontal resolution, 28 hybrid vertical levels 4 times (00UTC, 06UTC, 12UTC and 18UTC) per day. The forecast output data is interpolated to $1^\circ \times 1^\circ$ lat/lon resolution from 0 to 384 forecast hours at 6-hour intervals. The initial perturbations are generated by ETR method. A Stochastic Total Tendency Perturbation (STTP) scheme is applied in the forecast integration to simulate random model errors.

The impact of different ensemble sizes (80, 60, 40, 20, 10 and 5) on NCEP GEFS performance is studied in this paper. In order to consider both of running relative larger ensemble size and computation costs, the GEFS model resolution is reduced to T126 for this experiment. The experiment runs from December 1st, 2009 to January 31st, 2010, longer forecasts are made once per day, ETR cycling are every 6 hours. At each cycle, orthogonalization and centration are carried out for all 80 perturbations. Verifications are processed to 60, 40, 20, 10 and 5 ensemble members which are randomly chosen from 80-member.

3. Impact of ensemble size on ensemble skill in ideal model

To assess the performance of the Lorenz 96 model experiments, RMS error of ensemble mean (RMSE) and ensemble spread (SPREAD) (Toth et al., 2003) are used (Figs.1). It shows that 1). The SPREAD is closer to RMSE; 2). The forecast error is saturated at about 60 integrated time steps (corresponding to 15 days, 6 hours for each time step). By comparing RMSE for different ensemble sizes, Fig.1 shows that the improvement is more significant for enlarging the ensemble size from 10 to 20 (double) and from 20 to 40 (double) than for further increasing the ensemble size. This conclusion is corroborated in Figs.2 by using 200 members as an optimum reference to calculate RMSE ratios to other memberships. It should be noticed that the differences of all ensemble sizes are quite small at early lead-time (less than day 3), and at longer lead time, the 99% errors could be represented by 40 ensemble members only, but 96% errors are only represented by 10 ensemble members if assuming 200 members is a perfect ensemble size. Continuous Ranked Probabilistic Score (CRPS; Toth et al., 2003) is used to measure the reliability and resolution of ensemble based probabilistic forecast. The tendencies of CRPS curves shown in Figs.3 are similar to RMSE. However, for detail shown in Fig.4, the improvements of increasing ensemble size on the representativeness of errors are larger than RMSE shown in Fig.2. 10-member represents less than 96% errors at short lead times, which decreases to 92% for long lead times. When the sizes increase to more than 40 members, the ratios as a function of lead time have few changes which maintain more than 98% errors for all lead times, and for further increasing ensemble sizes, this percentage improves more obvious than RMSE ratios.

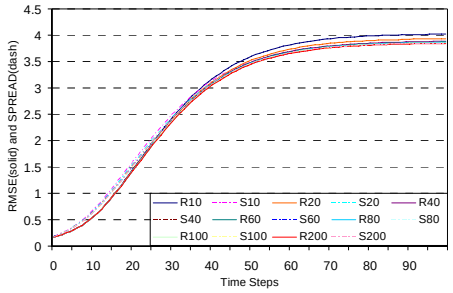

Fig.1 RMSE and SPREAD for different ensemble members.

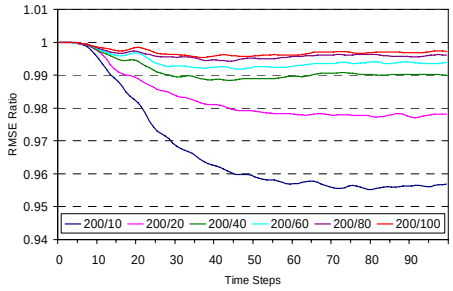

Fig.2 RMSE ratios of 200-member ensemble mean to others.

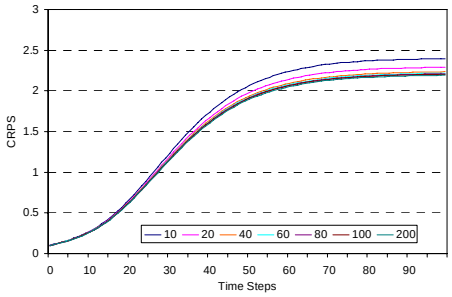

Fig.3 CRPS for different ensemble members.

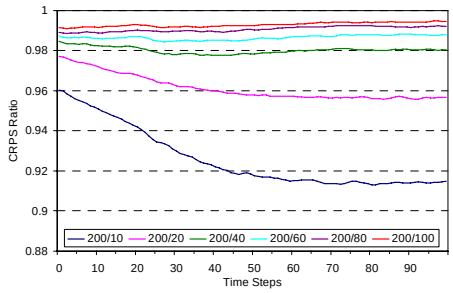

Fig.4 CRPS ratios of 200-member to other sizes.

4. Impact of ensemble size on ensemble skill in real atmospheric model

The benefits of increasing ensemble size are evaluated for 500hPa geopotential height over the NH extra-tropics based on NCEP standard probabilistic verification package (Zhu et al., 1996; Zhu and Toth, 2008) which includes RMSE, SPREAD, and CRPS.

4.1 RMSE and SPREAD

In general, ensemble SPREAD is equal to RMSE for a perfect forecast ensemble system in which the analysis is statistically indistinguishable from ensemble members. Fig.5 shows that the increases of ensemble sizes from 5 to 10, from 10 to 20 and from 20 to 40 produce statistically significant improvements of RMSE at all lead times. However, the improvements are very small when further increasing the ensemble size. SPREAD is smaller than RMSE and it isn't sensitive to increase ensemble sizes as RMSE, because increasing ensemble size have slightly effect on capturing the model error which mainly cause SPREAD underestimation. Fig.6 shows whether the differences of RMSE for ensemble sizes are statistical significance. A vertical bar represents a 95% of standard deviation. For example, the top panel shows the difference between 10 and 20 ensemble members. A positive value means 10 members have larger RMSE value than 20 members. The bars don't overlap with zero line which indicates differences significant at the 5% confidence level. RMSE for 20 members differs significantly from 40 members for short lead times (about less than 7 days), but the difference between 40 and 80 isn't significant for all lead times.

4.2 CRPS

The comparison of CRPS in the Fig.7 shows that the increase of ensemble size improves the probabilistic forecast skill, especially when the size is smaller than 40-member. The improvement is significantly larger than the forecast skill for ensemble mean evaluation.

This result has been confirmed by statistical significance test (Fig.8). The differences of CRPS for 10-20, 20-40 and 40-80 ensemble members are all significant at the 5% confidence level for all lead times which is different from ensemble mean verification, although they decrease greatly when the sizes increase from 20 to 40 and from 40 to 80.

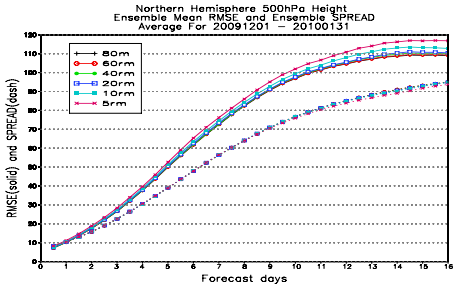

Fig.5 RMSE and SPREAD of different ensemble sizes for 500hPa geopotential height from 1 Dec. 2009 to 31 Jan. 2010 over the NH extra-tropics.

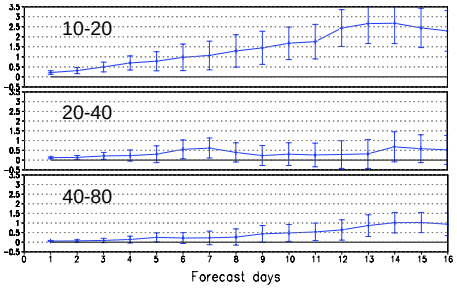

Fig.6 The differences of RMSE for 10-20, 20-40 and 40-80 ensemble members respectively. The Blue bars around the difference (blue line) are 95% confidence intervals.

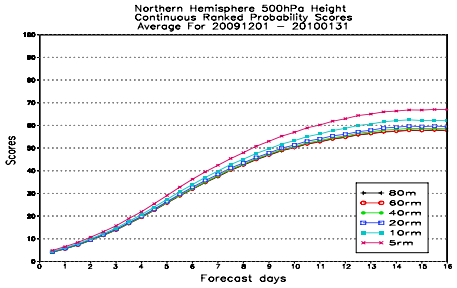

Fig.7 CRPS of different ensemble sizes for 500hPa geopotential height from 1 Dec. 2009 to 31 Jan. 2010 over the NH extra-tropics.

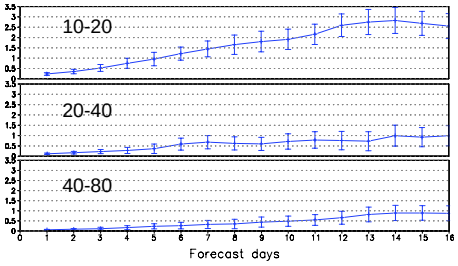

Fig.8 The differences of CRPS for 10-20, 20-40 and 40-80 ensemble members respectively. The Blue bars around the difference (blue line) are 95% confidence intervals.

5. The relative impact of increasing model resolution and increasing ensemble size in real atmospheric model

By comparing 70 members at T126L28 resolution with 20 members at T190L28 resolution which are using the equivalent computation resource and the same model physics, the relative impact for both increasing resolution and ensemble size has been assessed. All of the comparisons for PAC and CRPS scores (the top figures of Figs.9 and 10) seem similar that increasing model resolution (T190) is more (less) beneficial than increased ensemble size for short (long) lead times. The statistical significance testing (the bottom figures of Figs.9 and 10) confirms this conclusion. Table 1 summarizes the statistical significant forecast time at which one forecast configuration performs significantly better than the other one by using 95% confidence interval. We can clearly find that the resolution plays more important than ensemble size when the forecast time is less than 5d, however, large ensemble size is significantly superior to higher resolution when the forecast time exceed 12d, which means more ensemble members will benefit the extend forecast. Therefore, it is a trade-off between these resolution and ensemble membership configuration. The optimal configuration may be depended on the practical application. In

this experiment period, for 6-10 days forecast lead times, there is no significant difference between increasing resolution and membership. In NCEP, a higher resolution may be considered to improve 1-5 days forecast since CMC's ensemble has been implemented in the NAEFS (The North American Ensemble Forecast System) to have more membership. Meanwhile, lagged ensemble could be another optimal option by constructing week-2 or extended range forecast.

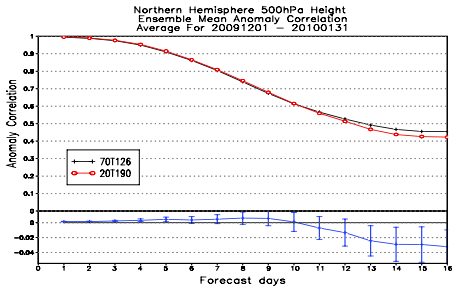

Fig.9 PAC (top) for 70T126 (black) and 20T190 (red) for NH extra-tropics 500hPa geopotential height from Dec. 1st, 2009 to Jan. 31st, 2010. The vertical bars around the RMSE difference (T190 – T126, solid line) are 95% confidence intervals (bottom).

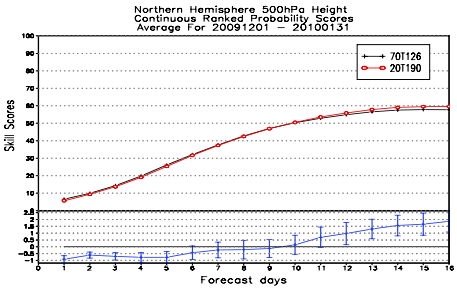

Fig.10 As in Fig.9 but for CRPS.

Table 1: Summary of statistical significant forecast time for 20T190 and 70T126

	PAC	CRPS
20T190	1-5d	1-5d
70T126	13-16d	12-16d

6. Conclusions

The numerical prediction centers around the world face the same questions when they develop (or upgrade) an ensemble forecast system. How many ensemble members do we need to have better representing forecast uncertainties with the limited computational resources? What is relationship between resolution and ensemble membership? This paper starts from Lorenz 96 model using ensemble transform with rescaling (ETR) initial perturbation method for over 200 members, then tests NCEP global ensemble forecast system (GEFS) with different ensemble size and resolution. The impact of various ensemble sizes is studied using different verification methods from December 1st, 2009 to January 31st, 2010 for 500hPa geopotential height field over the NH extra-tropics. Results indicate increasing ensemble size is beneficial to improve skill of ensemble, especially for small ensemble size (less than 40-member), and there is still significant improvement on the skill of probabilistic forecast with further increasing ensemble members. The relative benefits of T126L28 model with 70 members and T190L28 model with 20 members which have equivalent computing cost are also compared. The comparison of the two configurations, from the PAC, CRPS scores and statistical significant testing of their difference, indicates that increasing model resolution is more (less) beneficial than increasing ensemble size for short (long) lead times.

Acknowledgments: The authors thank Drs. Dingchen Hou, Mozheng Wei, Malaquias Pena and other members of Ensemble and Post Processing Team at EMC/NCEP for helpful suggestions during the course of this work. First author gratefully acknowledges the support of Dr. Stephen J. Lord and EMC.

References

Buizza, R., and T. N. Palmer, 1998a: Impact of ensemble size on ensemble prediction. Mon. Wea. Rev., 126, 2503–2518.

Buizza, R., T. Petroliagis, T. N. Palmer, J. Barkmeijer, M. Hamrud, A.Hollingsworth, A. Simmons, and N. Wedi, 1998b: Impact of model resolution and ensemble size on the performance of an ensemble prediction system. Quart. J. Roy. Meteor. Soc., 124, 1935–1960.

Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of precipitation using the ECMWF Ensemble Prediction System. Wea. Forecasting, 14, 168–189.

Du, J., S. L. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. Mon. Wea. Rev., 125, 2427–2459.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res., 99 (C5), 10143–10162.

Lorenz, E. N., 1996: Predictability: A problem partly solved. Proc. Workshop on Predictability, Vol. 1, Reading, United Kingdom, ECMWF, 1–18.

Mullen, S. L., and R. Buizza, 2002: The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF Ensemble Prediction System. Wea. Forecasting, 17, 173–191.

Richardson, D. S., 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. Quart. J. Roy. Meteor. Soc., 127, 2473–2489.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: "Probability and Ensemble Forecasts." In book of: Forecast Verification: A practitioner's guide in atmospheric science. Ed.: I. T. Jolliffe and D. B. Stephenson. Wiley, 137–163.

Wei, M., Z. Toth, R. Wobus, Y. Zhu, C. H. Bishop, and X. Wang, 2006: Ensemble Transform Kalman Filter-based ensemble perturbations in an operational global prediction system at NCEP. Tellus, 58A, 28–44.

Wei, M., Z. Toth, R. Wobus, and Y. Zhu, 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. Tellus, 60A, 62–79.

Zhu, Y., G. Iyengar, Z. Toth, S. Tracton, and T. Marchok, 1996: Objective Evaluation of the NCEP Global Ensemble Forecasting System. In Proceedings of the 15th AMS Conference on Weather Analysis and Forecasting, 19–23 August 1996, Norfolk, Virginia.

Zhu, Y., and Z. Toth, 2008: Ensemble Based Probabilistic Forecast Verification. In Proceedings of the 19th AMS Conference on Probability and Statistics, 21–24 January 2008, New Orleans, Louisiana.