What is an Optimum Configuration for Operational Ensemble Forecast System?

Juhui Ma[1,2,3], Yuejian Zhu[2] and Panxing Wang[1]

1) Key Laboratory of Meteorological Disaster of Ministry of Education, NUIST, Nanjing 210044, China

2) Environmental Modeling Center/NCEP/NOAA, Camp Springs, MD 20746, USA

3) UCAR, Boulder, CO 80307, USA

Manuscript submitted to

Monthly Weather Review

7 March 2011

Corresponding author address:
Juhui Ma
Environmental Modeling Center/NCEP/NOAA
W/NP2 NOAA WWB #207
5200 Auth Road
Camp Springs, MD 20746
(301)7638000 ext. 7025
Email: juhui.ma@noaa.gov

ABSTRACT

Numerical Weather Prediction (NWP) 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 better represent forecast uncertainties with limited computational resources? What is the relationship between resolution and ensemble size? This paper starts from ensembles of the Lorenz 96 model generated using the ensemble transform with rescaling (ETR) initial perturbation method for over 200 members. The results are contrasted with tests based on the NCEP Global Ensemble Forecast System (GEFS) with different ensemble sizes 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 Northern Hemisphere (NH) and Southern Hemisphere (SH) extra-tropics. Results indicate that 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 Pattern Anomaly Correlation score (PAC), Continuous Ranked Probabilistic Score (CRPS) and statistical significance testing of their difference, indicates that increasing model resolution is more (less) beneficial than increasing ensemble size for short (long) lead times.

1. Introduction

Since 1992, the European Center for Medium-Range Weather Forecasting (ECMWF) and the National Centers for Environmental Prediction (NCEP) implemented operational global ensemble forecast systems (the system is called GEFS for short at NCEP). Almost two decades later, GEFS has been operationally implemented in many numerical weather prediction centres around the world such as CMC (Canada), Met Office (UK), JMA (Japan) and so on. Ensemble based probabilistic forecast is a feasible method to estimate forecast uncertainty through perturbed initial conditions (or perturbed observation for assimilation) which greatly improve and extend numerical forecast skill by comparing with deterministic forecast (Zhu and Ma, 2010). The operational initial perturbation used to represent initial uncertainty has undergone great development from the Singular Vector (SV) method (Buizza and Palmer, 1995; Molteni et al., 1996), the breeding method (Toth and Kalnay, 1993, 1997) and the Perturbed Observation (PO) method (Houtekamer et al., 1996) to the Ensemble based data assimilation and Singular Vector (EDA-SV) method (Buizza et al., 2008, 2010), the Ensemble Transform with Rescaling (ETR) method (Wei et al., 2008) and the Ensemble Kalman Filter (EnKF) method (Houtekamer and Mitchell, 2005; Houtekamer et al., 2007), which can improve the representation of the uncertainties in analysis. Though operational ensemble forecast systems focused only on assessing the initial uncertainty at first, several attempts have been made to account for model-related uncertainty, such as multi-model and multi-parameterization (implemented in CMC since 1998, Houtekamer et al., 1996), Stochastic Physics Parameterization Tendencies (implemented in ECMWF since 1998, Buizza et al., 1999a, Palmer et al., 2009), SPectral stochastic Backscatter Scheme (implemented in ECMWF since 2010, Shutts 2004, 2005; Berner et al., 2009), Stochastic Total Tendency Perturbation (implemented in NCEP since 2010, Hou et al., 2006, 2008, 2010), perturbed surface parameters (Eckel and Mass, 2005), coupling to ocean ensemble (Holt et al., 2009) and so on.

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., 1999b; Richardson, 2001; Mullen and Buizza, 2002). However, the limited

computational resources constrain model resolution and ensemble size. Therefore, when designing an effective operational ensemble prediction system, there are two main questions we seek answers to, which are 1) how many ensemble members do we need to better represent forecast uncertainties with limited computational resources? And 2) what is the relative impact of increasing model resolution and increasing ensemble size on forecast skill? In this study, the two questions above will be analyzed by using both the Lorenz 96 model (Lorenz, 1996) and the NCEP GEFS.

The famous Lorenz models are similar to other nonlinear dynamical models of atmospheric system. They have been widely used in many ensemble forecast studies. Anderson (1997) compared the performance of ensembles based on random perturbations, bred vectors and SVs with the Lorenz 63 model. Bowler (2006) compared initial perturbation methods including ensemble Kalman filter, bred vectors and SVs using the Lorenz 96 model. It is extremely expensive and complex to carry out experiment with operational forecast models enlarging the ensemble size for the purpose of expanding the sample of numerical model's phase space; however, it is feasible in the Lorenz model due to simple dynamical system. Experiments with large ensemble size attained using the Lorenz model can give a theoretical instruction in this study with less computational cost. It is recognized that the Lorenz model has limitations to represent the complexity of the realistic atmospheric system. Furthermore, the assimilation data used for this study's experiment are synthetic observations generated from random number that limits the reliability of the conclusion. To verify the conclusions obtained based on the Lorenz 96 model, this study uses realistic operational ensemble forecast system for relatively small ensemble sizes. Buizza and Palmer (1998a) analyzed the impact of 2, 4, 8, 16 and 32-member on the performance of the ECMWF Ensemble Prediction System (EPS) for 500hPa geopotential height field. Mullen and Buizza (2002) assessed the effect of horizontal resolution and ensemble size on the ECMWF EPS for 24-h accumulated precipitation. The comparisons of T_L159M51, T_L255M51, T_L319M51, T_L255M15 and T_L319M15 ("M" refers to the number of ensemble members) are shown in that paper. Reynolds et al. evaluated the impact of resolution versus ensemble size tradeoffs on the U.S. Navy global ensemble performance using resolution of T119, T159 and T239, with

33, 17 and 9 ensemble members. In this study, NCEP operational GEFS is employed and ensemble size will increase to 80-member.

The purpose of this paper is to determine a reasonable (or optimal) ensemble size and the relationship with resolution for operational ensemble prediction system. Section 2 will describe the models and experimental design. In section 3 and 4, the impacts of ensemble sizes on ensemble skill, using Lorenz 96 model and NCEP GEFS, are examined respectively. Section 5 compares a relative trade-off of increasing model resolution versus increasing ensemble size through NCEP GEFS experiments. Section 6 provides a summary and conclusions.

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 set of differential equations

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

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. Analysis method

The truth run for all 1000 variables is obtained by integrating the Lorenz 96 model from randomly generated initial fields. The observations $\mathbf{y}$ are the fields after perturbing the truth run with an error standard deviation of 0.2 at each time step. Here, the ensemble mean of analysis $\overline{\mathbf{x}}^a$ is considered to be the best estimate of analysis. At each time step, observations $\mathbf{y}$ are assimilated by ensemble mean of forecast $\overline{\mathbf{x}}^b$ using equation (2) to update $\overline{\mathbf{x}}^a$, and then obtain analysis-error covariance by equation (3) (Evensen, 1994).

$$\overline{\mathbf{x}}^a = \overline{\mathbf{x}}^b + \mathbf{P}^b\mathbf{H}^T(\mathbf{H}\mathbf{P}^b\mathbf{H}^T + \mathbf{R})^{-1}(\mathbf{y} - \mathbf{H}\overline{\mathbf{x}}^b), \tag{2}$$

$$\mathbf{P}^a = \mathbf{P}^b - \mathbf{P}^b\mathbf{H}^T(\mathbf{H}\mathbf{P}^b\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{P}^b, \tag{3}$$

where $\mathbf{P}^a$ is the analysis-error covariance; $\mathbf{P}^b$ is the background-error covariance; $\mathbf{H}$ is observation operator (assume the forecasts and observations correspond to the same points, so the elements of the diagonal matrix are equal to 1); $\mathbf{R}$ is the observation-error covariance (assume the elements of the diagonal matrix that are equal to 0.03). $\overline{\mathbf{x}}^b$ is equal to $\frac{1}{m}\sum_{k=1}^{m}\mathbf{x}_k^b$ in which m is chosen 40 and initial condition of $\mathbf{x}_k^b$ is obtained by adding paired random perturbations with amplitude of 0.2 to the analysis of previous time step, except random perturbations is added to the truth run at time 0. The background-error covariance estimate is generated from this 40-member ensemble using equation

$$\mathbf{P}^b = \frac{1}{m-1}\mathbf{X}'^b\mathbf{X}'^{bT}, \tag{4}$$

where $\mathbf{X}'^b$ is defined as a matrix formed from the ensemble of perturbations $\mathbf{X}'^b = (\mathbf{x}_1'^b, ..., \mathbf{x}_m'^b)$ in which $\mathbf{x}_k'^b = \mathbf{x}_k^b - \overline{\mathbf{x}}^b$.

c. Initial perturbation method

Initial perturbations are generated 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 through an ensemble transformation $\mathbf{T}$ as follows

$$\mathbf{Z}^a = \mathbf{Z}^f\mathbf{T}. \tag{5}$$

$\mathbf{T}$ can be constructed by solving eigenvalue $\mathbf{\Gamma} = diag(\lambda_1, \lambda_2, ...\lambda_m)$ and eigenvector $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, ...\mathbf{c}_m]$ of $\mathbf{Z}^{f^T}\mathbf{P}^{a^{-1}}\mathbf{Z}^f$,

$$\mathbf{Z}^{f^T}\mathbf{P}^{a^{-1}}\mathbf{Z}^f = \mathbf{C}\mathbf{\Gamma}\mathbf{C}^{-1}, \tag{6}$$

where $\mathbf{P}^a$ is obtained from analysis process or data assimilation.

Suppose $\mathbf{G} = diag(\lambda_1, \lambda_2, ...\lambda_{m-1}, \alpha)$, where α is a non-zero constant. The analysis perturbations can be generated through $\mathbf{T}_p = \mathbf{C}\mathbf{G}^{-1/2}$ as

$$\mathbf{Z}_p{}^a = \mathbf{Z}^f\mathbf{T}_p = \mathbf{Z}^f\mathbf{C}\mathbf{G}^{-1/2}. \tag{7}$$

After transformation, the analysis perturbations are orthogonal, but not centered. If the initial perturbations are centered around the analysis, the performance of ensemble mean will be better. Equation (8) can be used to centralize the perturbations.

$$\mathbf{Z}^a = \mathbf{Z}_p{}^a\mathbf{C}^T = \mathbf{Z}^f\mathbf{C}\mathbf{G}^{-1/2}\mathbf{C}^T. \quad (8)$$

Though the initial perturbations are not orthogonal anymore after centralizing, the more ensemble members we have, the more orthogonal the perturbations will become.

To make the initial spread be similar to the analysis-error covariance, it is rescaled using factor γ which is defined as the ratio of the square root of $\mathbf{P}^a$ and the square root of $\frac{1}{m}\sum_{k=1}^{m}\mathbf{z}_k{}^{a2}$.

2.2 NCEP GEFS model and its application

The current NCEP operational GEFS (based on GFS v8.00 which was in operation since December 15th, 2009) 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 are interpolated to $1°\times1°$ lat/lon resolution from 0 to 384 forecast hours at 6-hour intervals. The initial perturbations, around the analysis provided by GDAS/GSI, are generated using the ETR method which is the same as in the Lorenz 96 experiment. 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. To be able to run a relatively larger ensemble size at similar computation costs, the GEFS model resolution is reduced to T126 and STTP is not applied for this experiment. The experiment runs from December 1st, 2009 to January 31st, 2010, and long forecasts are made once per day. ETR cycling are every 6 hours. At each cycle, both equations (7) and (8) are used to orthogonalize and centralize all 80 perturbations. Verifications are processed for 60, 40, 20, 10 and 5 ensemble members which are randomly chosen from 80-member.

2.3 Verification methodology

a. RMS error of ensemble mean (RMSE) and ensemble spread (SPREAD)

RMSE of ensemble mean measures the distance between forecast and analyses. SPREAD measures the deviation of ensemble members from the ensemble mean. In general, SPREAD is equal to RMSE for a perfect forecast ensemble system in which the verifying analysis is statistically indistinguishable from ensemble members. In other words, ensemble forecast could represent all the uncertainties associated with initial errors and model errors. (Toth et al., 2003)

b. Pattern Anomaly Correlation score (PAC)

PAC measures the ability of ensemble mean to represent weather patterns which is defined as the correlation between the predicted anomaly and the observed anomaly with respect to their corresponding climatology. The maximum value of 1.0 indicates a perfect depiction of the patterns. (Zhu, 2005)

c. Continuous Ranked Probabilistic Score (CRPS)

CRPS is used to measure the reliability and resolution of ensemble based probabilistic forecast by calculating the distance between the predicted and the observed cumulative density functions of scalar variables. The high (low) value indicates a low (high) skill of the forecast system. (Toth et al., 2003)

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

To explore what could be a reasonable ensemble size, the performance of relative large ensemble size (greater than 100 members) should be studied. However, it is extremely expensive and complex for running operational forecast models as mentioned in Section 1. Therefore, most studies assessing the performance of ensemble sizes focused on a limited membership for realistic atmospheric models. In this study, the Lorenz model is employed to demonstrate this issue through a theoretical study and a numerical application with less computational cost. Initial perturbations in this ideal experiment are generated using the ETR method. This method was implemented for operations at NCEP since 2006. Wei et al. (2006, 2008) compared the results for different ensemble forecast systems based on BM, ET, ETR and ETKF. Magnusson et al. (2009) compared SV and ET using ECMWF IFS-model. McLay et al. (2008) found that ET perturbation, with a

finite ensemble size, is too small in the tropics and too large in the midlatitudes. But there are few studies that relate to the impact of ensemble sizes with the ETR initialization.

To assess the performance of the Lorenz 96 model experiments, RMSE, SPREAD and CRPS are used. Figs.1 show that 1) The SPREAD is close to RMSE; 2) The forecast error is saturated at about 60 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 among all ensemble sizes are quite small at early lead-time (less than day 3), and at longer lead time, if assuming 200 members is a perfect ensemble size which can represent all errors, the 99% errors could be represented by 40 ensemble members, but 96% errors are only represented by 10 ensemble members. 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 obviously than RMSE ratios.

4. Impact of ensemble size on ensemble skill in a realistic atmospheric model

The different ensemble sizes are tested on NCEP global ensemble forecast system (GEFS) running at T126L28 resolutions. Ensemble forecast skill scores have been computed for ensemble sizes of 5, 10, 20, 40, 60 and 80 members for the period December 1st, 2009 to January 31st, 2010. The improvements of increasing ensemble size are assessed based on 500hPa geopotential height over the NH and SH extra-tropics from the NCEP standard probabilistic verification package (Zhu et al., 1996, 2002; Zhu, 2004; Zhu and Toth, 2008), which includes RMSE, SPREAD, PAC and CRPS.

4.1 RMSE and SPREAD

Fig.5a shows that SPREAD is smaller than RMSE due to not accounting for all possible sources of model related uncertainty. If STTP is applied, the situation will be improved to some extend. 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 over the NH extra-tropics. However, the improvements are very small when further increasing the ensemble size. SPREAD is insensitive to increases in ensemble sizes. Fig.6 shows whether the differences of RMSE for ensemble sizes are statistically significant. 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 do not cross with the zero line, suggesting significant differences at the 5% confidence level. RMSE for 20 members differs significantly from 40 members for short lead times (about less than 7 days). The difference between 40 and 80 is not significant for short lead times and is significant after about 10 days. The situation over the SH extra-tropics is similar to the NH extra-tropics (see Fig.5b), but both RMSE and SPREAD are smaller than over the NH extra-tropics because of seasonal variations of circulation patterns and predictability.

4.2 PAC

If we consider 65% PAC score as a useful skill for large scale weather forecast, it is very clear to see from Fig.7a that there are about 13 hours (approximately from 214-hour extended to 227-hour) gain by increasing the ensemble size from 5 to 80 over the NH extra-tropics. It is evident that ensemble systems should have more than 20 members. The forecast performance for the SH extra-tropics is clearly lower (about 1 day difference) than for the NH extra-tropics. The difference may be due to differences in the quality of initial conditions, and different seasons. Despite these differences, the impact of increasing the ensemble size is similar in both extra-tropical hemispheres.

4.3 CRPS

The comparison of CRPS in Fig.8 shows that the increase in ensemble size improves the probabilistic forecast skill over both the NH and the SH extra-tropics, especially when the size is smaller than 40-member. The improvement is significantly larger than the forecast skill for ensemble mean evaluation. For example, over the NH extra-tropics, Fig.7a shows that PAC score for 5-member is larger than 65% before 214-hour. If we use 214-hour for 5-member as a reference to project on CRPS score map (Fig.8, corresponding to 0.53 skill score of CRPS), the skill score extends to 262hr for 80-member, so there is approximately 35-hour more gain than the PAC score for increasing the ensemble size from 5 to 80-member. This result has been confirmed by statistical significance test (Fig.9). 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.

5. Model resolution versus ensemble size in a realistic atmospheric model

The relative impact of increasing model resolution versus increasing ensemble size is assessed by comparing 70 members at T126L28 resolution with 20 members at T190L28 resolution. These two configurations take equivalent computation resources and use the same model physics.. The comparisons of PAC and CRPS scores (the top figures of Figs.10 and 11) indicate similarly that increasing model resolution (T190) is more (less) beneficial than increased ensemble size for short (long) lead times. A statistical significance test (the bottom figures of Figs.10 and 11) 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 under similar computer resources and model physics, the resolution plays a more important role than ensemble size when the forecast lead time is less than 5 days, whereas large ensemble size is significantly superior to higher resolution when the forecast lead time exceeds 12 days, which means more ensemble members will benefit the extended range forecast. Therefore, there is a trade-off between model resolution and ensemble membership configuration. The optimal configuration may depend on the

application. In this experiment period, for 6-10 days forecast lead times, there is no significant difference between increasing resolution and membership. At NCEP, a higher resolution may be considered to improve 1-5 days forecast. Meanwhile, the approach of using lagged ensemble members could be an option to enhance week-2 or longer range ensemble forecast skill.

6. Conclusions

This study compares the impact of different ensemble sizes, and the relative performance of ensemble sizes versus model resolution to investigate an optimal configuration to improve the skill of operational ensemble forecasts. These issues are considered every time a numerical center upgrades or develops its ensemble prediction system.

The study starts by using the Lorenz 96 model to obtain answers to these issues. Because the model is computationally simple, it allows to arrive to conclusions with less cost than using a realistic atmospheric model. Initial perturbations are generated by using ETR method with 10, 20, 40, 60, 80, 100 and 200 ensemble members. RMSE, SPREAD and CRPS are used to measure the ensemble performance. The results show that performances of ensemble mean improve slightly when ensemble sizes increase at early lead-time (less than day 3), and at longer lead time, if assuming 200 members is a perfect ensemble size which can represent all errors, the 99% errors could be represented by 40 ensemble members, but 96% errors are only represented by 10 ensemble members. Performances of probabilistic forecast improve more obviously with increasing ensemble size than performances of ensemble mean.

The conclusions obtained from the Lorenz model are corroborated with a more realistic model of the atmosphere. The experiments from NCEP GEFS at T126L28 resolution are used to test the impact of membership. The ensemble forecasts are generated by using 5, 10, 20, 40, 60 and 80 members from December 1st, 2009 to January 31st, 2010. A 500hPa geopotential height over the NH and SH extra-tropics has been considered as main variable for evaluation. The measures, such as RMSE, SPREAD, PAC and CRPS, are applied to evaluate the benefits of increasing ensemble size. SPREAD is not sensitive to

increase ensemble sizes. RMSE score indicates that there are statistically significant improvements in ensemble mean performance when the ensemble size is less than 40-member, however, the improvements are not significant any more with further increasing ensemble members. But a probabilistic forecast verification score (CRPS) shows that the improvements are still significant when doubling ensemble size from 40 to 80-member. Similarly, the comparison of the gains for useful skill through increasing ensemble size from 5 to 80-member evaluated by PAC and CRPS scores indicates the different impacts on the skill of ensemble mean and probabilistic forecast with increasing the ensemble size. There is about 13-hours skill gain over the NH extra-tropics shown in the PAC score, which is 35 hours less gain than in the CRPS (probabilistic measurement). Overall, increasing ensemble size is beneficial to improve skill of ensemble forecasts, especially when the ensemble size is small, and there is still significant improvement on the skill of probabilistic forecast when ensemble size becomes lager.

Numerical centers that develop global ensemble prediction systems face the issue of best use of their available computational resources. They usually compromise between increasing model resolution and enlarging ensemble size. The relative benefits of T126L28 model with 70 members and T190L28 model with 20 members which have equivalent computing cost are compared for 500hPa geopotential height from December 1st, 2009 to January 31st, 2010 over the NH extra-tropics. The comparison of 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.

Based on these experiments, we will continue our study to improve ensemble initialization from hybrid EnKF/ETR by using operational GEFS with higher resolution. We will focus on the impact of ensemble size and resolution for the summer season, and on the precipitation forecast and the storm forecast in a future study.

Acknowledgments: The authors thank Drs. Dingchen Hou, Mozheng Wei, Malaquias Peña 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

Anderson J. L., 1997: The impact of dynamical constrains on the selection of initial conditions for ensemble predictions: Low-order perfect model results. Mon. Weather Rev., 125, 2969–2983.

Berner, J., G. J. Shutts, M. Leutbecher, and T. N. Palmer, 2009: A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF ensemble prediction system. J. Atmos. Sci., 66, 603–626.

Bowler N. E., 2006: Comparison of error breeding, singular vectors, random perturbations and ensemble Kalman filter perturbation strategies on a simple model. Tellus A, 58, 538–548.

Buizza, R., and T. N. Palmer, 1995: The singular-vector structure of the atmospheric global circulation. J. Atmos. Sci. 52, 1434–1456.

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., M. Miller, and T. N. Palmer, 1999a: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. Quart. J. Roy. Meteor. Soc., 125, 2887–2908.

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

Buizza, R., M. Leutbecher, and L. Isaksen, 2008: Potential use of an ensemble of analyses in the ECMWF Ensemble Prediction System. Quart. J. Roy. Meteor. Soc., 134, 2051–2066.

Buizza, R., M. Leutbecher, L. Isaksen, and J. Haseler, 2010: Combined use of EDA- and SV-based perturbations in the EPS. Newsletter n. 123, ECMWF, Shinfield Park, Reading RG2-9AX, UK, pg 22–28.

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

Eckel, F. S., and C. F. Mass, 2005: Aspects of effective mesoscale, short-range Ensemble forecasting. Wea. Forecasting, 20, 328–350.

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

Holt, T., J. Pullen, and C. H. Bishop, 2009: Urban and ocean ensembles for improved meteorological and dispersion modeling of the central zone. Tellus A, 61, 232–249.

Hou, D., Z. Toth, and Y. Zhu, 2006: A stochastic parameterization scheme within NCEP Global Ensemble Forecast System. In Proceedings of the 18th AMS Conference on Probability and Statistics, 29 January –2 February 2006, Atlanta, Georgia.

Hou, D., Z. Toth, Y. Zhu, and W. Yang, 2008: Impact of a stochastic perturbation scheme on NCEP Global Ensemble Forecast System. In Proceedings of the 19th AMS Conference on Probability and Statistics, 21–24 January 2008, New Orleans, Louisiana.

Hou, D., Z. Toth, Y. Zhu, W. Yang, and R. Wobus, A stochastic perturbation scheme representing model-related uncertainties in the NCEP Global Ensemble Forecast System. Manuscript to be published.

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.

Houtekamer, P. L., and H. L. Mitchell, 2005: Ensemble Kalman filtering. Quart. J. Roy. Meteor. Soc., 131, 3269–3289.

Houtekamer, P. L., M. Charron, H. L. Mitchell, and G. Pellerin, 2007: Status of the Global EPS at Environment Canada. Proc. ECMWF Workshop on Ensemble Prediction, 7-9 November 2007, ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, UK, 57–68.

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

Magnusson, L., J. Nycander, and E. Källén, 2009: Flow-dependent versus flow-independent initial perturbations for ensemble prediction. Tellus, 61A, 194–209.

McLay, J. G., C. H. Bishop, and C. A. Reynolds, 2008: Evaluation of the ensemble transform analysis perturbation scheme at NRL. Mon. Wea. Rev., 136, 1093–1108.

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.

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.

Palmer, T. N., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. J. Shutts, M. Steinheimer, and A. Weisheimer, 2009: Stochastic parametrization and model uncertainty. ECMWF Tech. Memo. no. 598, ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, UK.

Reynolds, C. A., J. G. McLay, J. S. Goerss, E. A. Serra, D. Hodyss, and C. R. Sampson, Impact of resolution and design on the U.S. Navy global ensemble performance in the tropics. Manuscript to be published.

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.

Shutts, G. J., 2004: A stochastic kinetic energy backscatter algorithm for use in ensemble prediction systems. ECMWF Tech. Memo. no. 449, ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, UK.

Shutts, G. J., 2005: A kinetic energy backscatter algorithm for use in ensemble prediction systems. Quart. J. Roy. Meteor. Soc., 131, 3079–3102.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: the generation of perturbations. Bull. Amer. Meteror. Soc., 174, 2317–2330.

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

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., 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.

Zhu, Y., 2004: Probabilistic Forecasts and Evaluations based on Global Ensemble Forecast System. World Scientific Series on Meteorology of East Asia, 3, 277–287.

Zhu, Y., 2005: Ensemble forecast: A new approach to uncertainty and predictability. Advance in Atmospheric Sciences, 22, 781–788.

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.

Zhu, Y., and J. Ma, 2010: "Predictability, Probabilistic Forecasting and Ensemble Prediction System." In book of: Lecture Notes on Numerical Weather Prediction. Ed.: WMO Regional Training Centre, NUIST, China, 43–60.

Figure Captions

Figure 1: RMSE and SPREAD for different ensemble members.

Figure 2: RMSE ratios of 200-member ensemble mean to other sizes.

Figure 3: CRPS for different ensemble members.

Figure 4: CRPS ratios of 200-member ensemble mean to other sizes.

Figure 5: RMSE and SPREAD of different ensemble sizes for 500hPa geopotential height from 1 Dec. 2009 to 31 Jan. 2010. a, over the NH extra-tropics; b, over the SH extra-tropics.

Figure 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.

Figure 7: As in Figure 5 but for PAC.

Figure 8: As in Figure 5 but for CRPS.

Figure 9: As in Figure 6 but for CRPS.

Figure 10: 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).

Figure 11: As in Figure 10 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

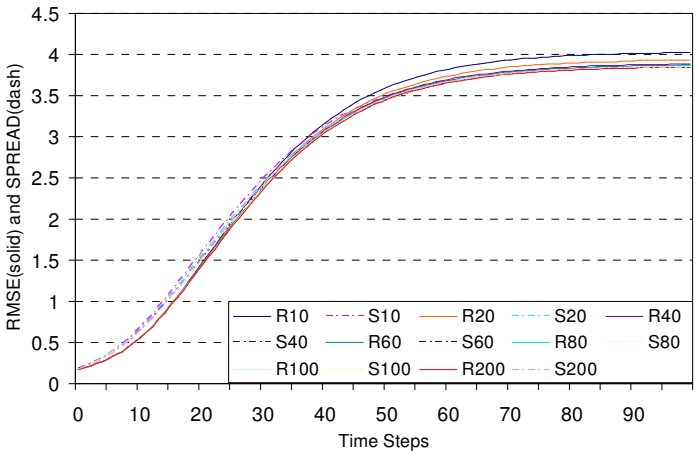

Fig.1 RMSE and SPREAD for different ensemble members.

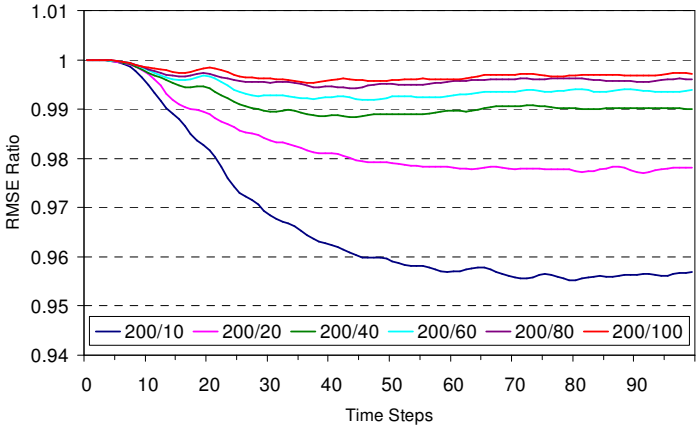

Fig.2 RMSE ratios of 200-member ensemble mean to other sizes.

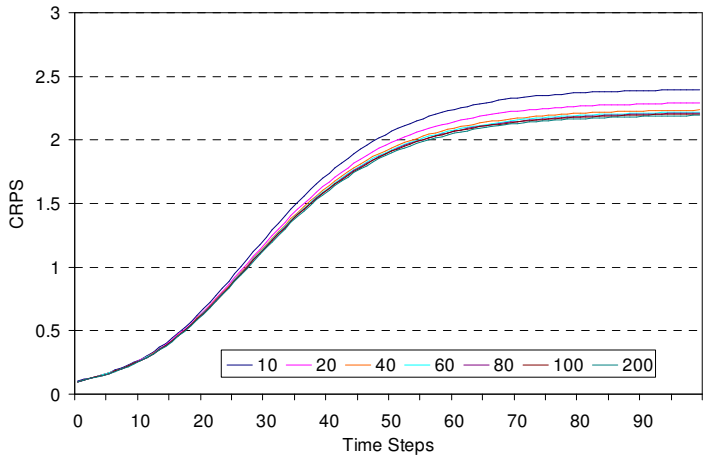

Fig.3 CRPS for different ensemble members.

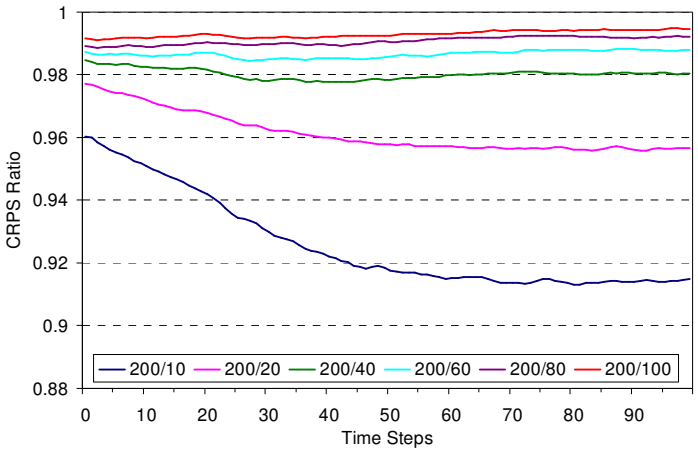

Fig.4 CRPS ratios of 200-member to other sizes.

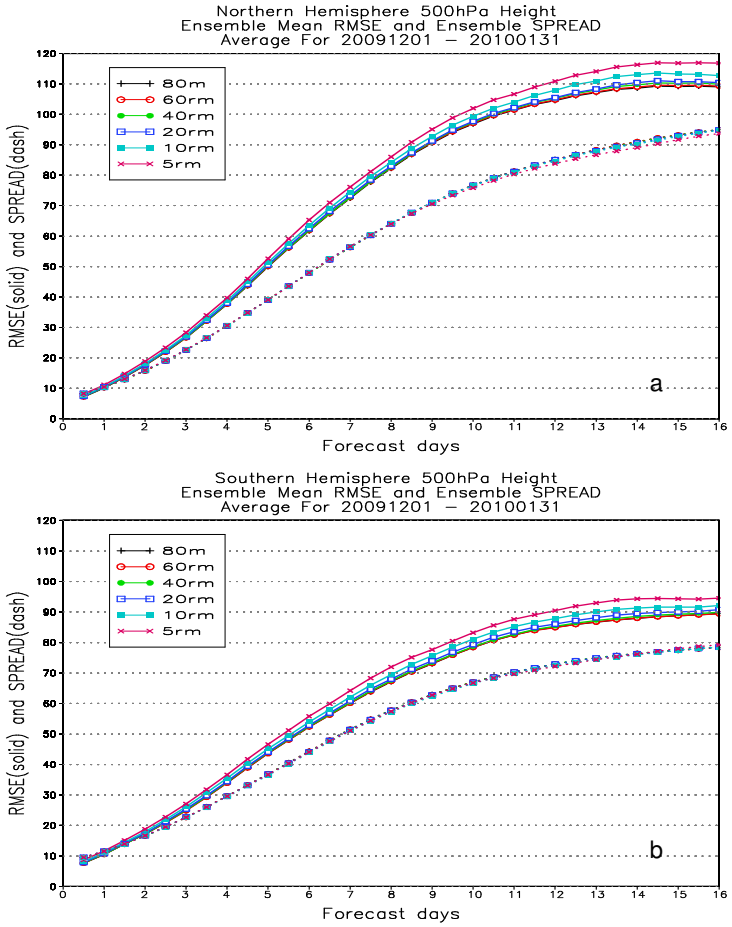

Fig.5 RMSE and SPREAD of different ensemble sizes for 500hPa geopotential height from 1 Dec. 2009 to 31 Jan. 2010.

a, over the NH extra-tropics; b, over the SH extra-tropics.

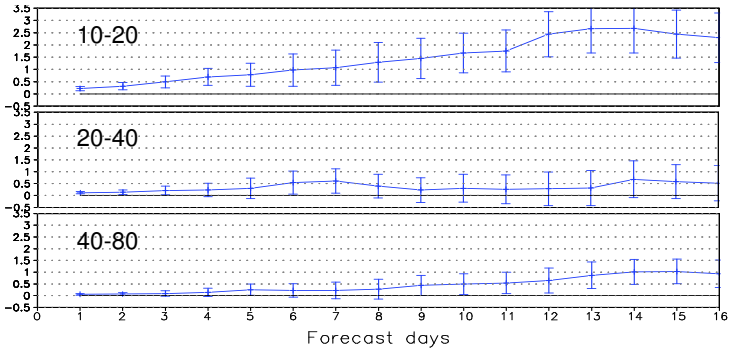

Fig.6 The differences of RMSE for 10-20, 20-40 and 40-80 ensemble members respectively. The Blue bars around the difference (blue line) are 95% confidence intervals.

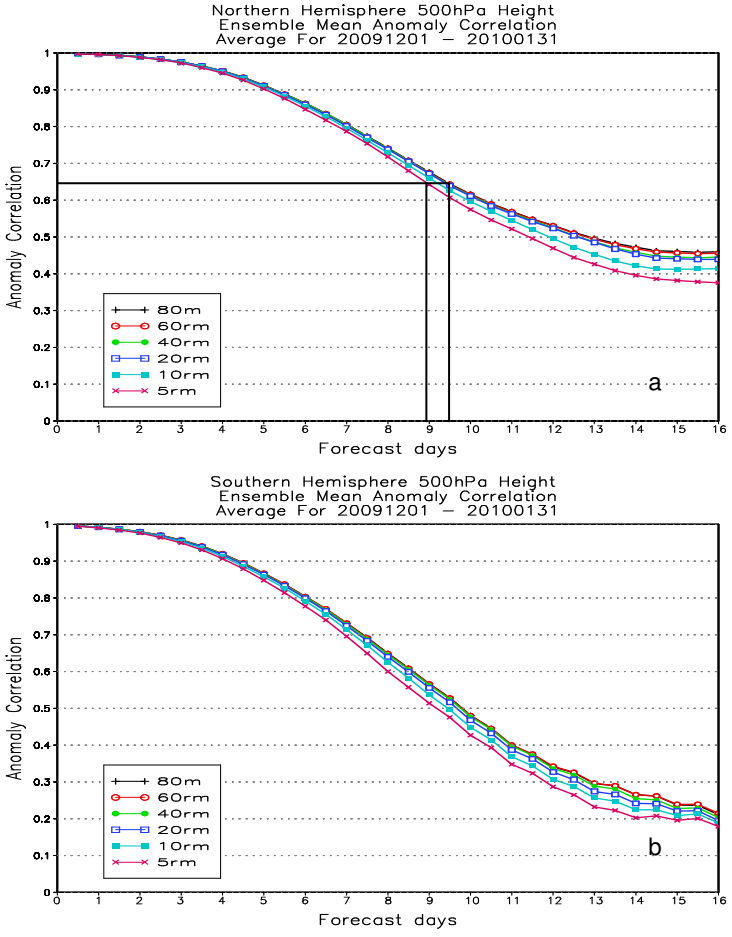

Fig.7 As in Fig.5 but for PAC.

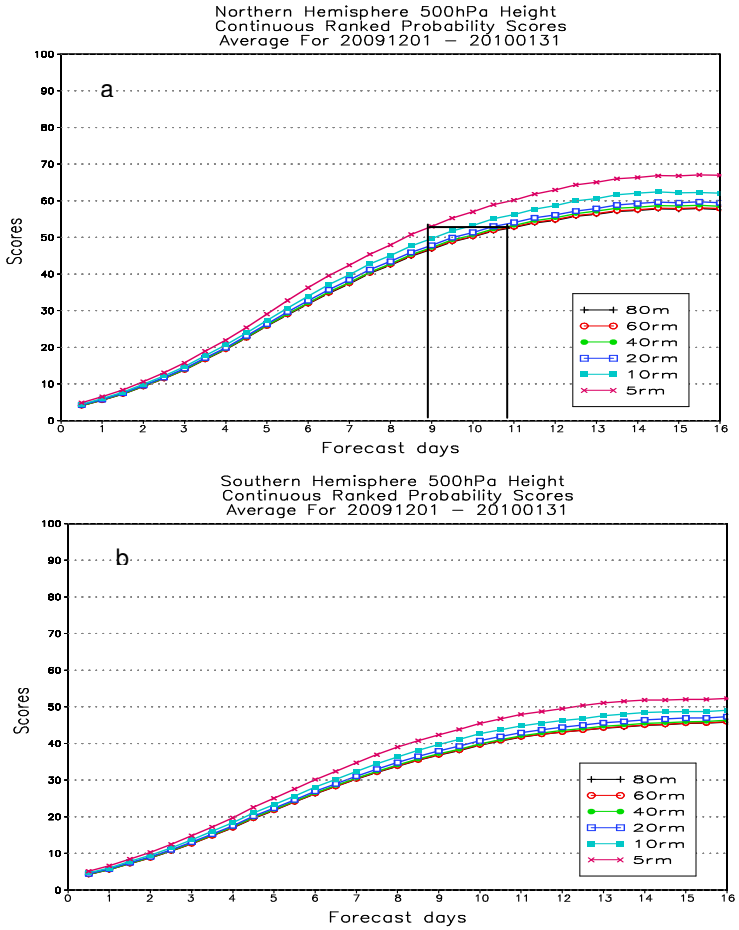

Fig.8 As in Fig.5 but for CRPS.

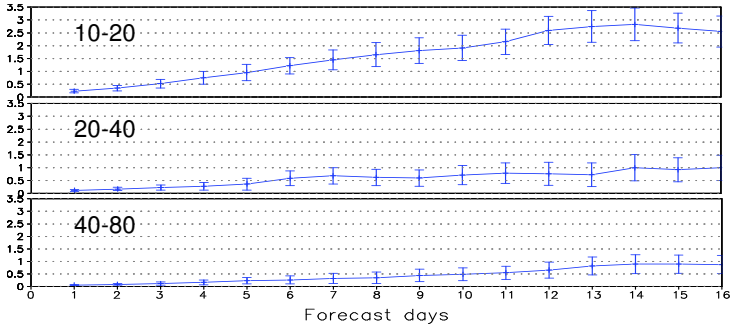

Fig.9 As in Fig.6 but for CRPS.

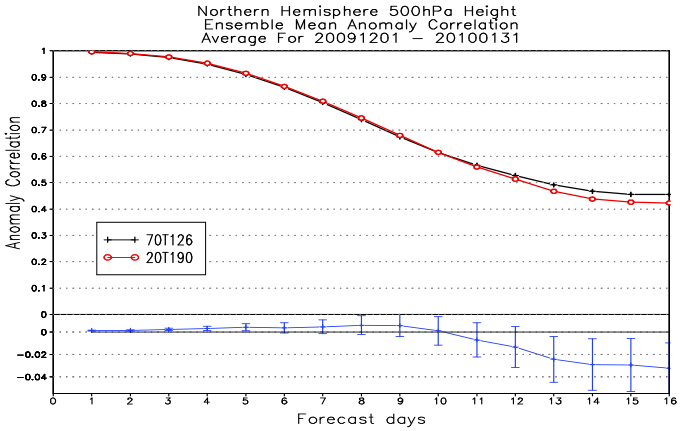

Fig.10 PAC (top) for 70T126 (black) and 20T190 (red) for NH extra-tropics 500hPa geopotential height from Dec. 1st, 2009 to Jan. 31st, 2010. The vertical bars around the RMSE difference (T190 – T126, solid line) are 95% confidence intervals (bottom).

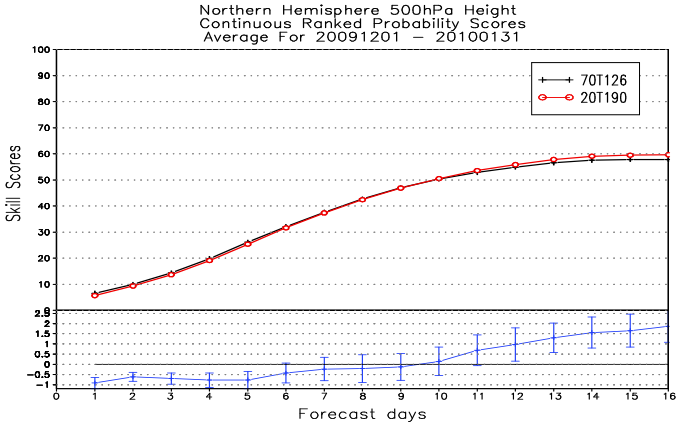

Fig.11 As in Fig.10 but for CRPS.