ADVANCES IN ATMOSPHERIC SCIENCES, VOL. 29, NO. 4, 2012, 782–794

An Effective Configuration of Ensemble Size and Horizontal Resolution for the NCEP GEFS

MA Juhui (麻巨慧)*[1,2,3], Yuejian ZHU[2], Richard WOBUS[4], and Panxing WANG[1]

[1] *Key Laboratory of Meteorological Disaster of Ministry of Education,*
Nanjing University of Information Science & Technology, Nanjing 210044

[2] *Environmental Modeling Center/NCEP/NOAA, Camp Springs, MD* 20746*, USA*

[3] *UCAR, Boulder, CO* 80307*, USA*

[4] *I.M. Systems Group, Inc.* (*IMSG*) *at the Environmental Modeling Center/NCEP/NOAA,*
Camp Springs, MD 20746*, USA*

(Received 30 November 2011; revised 12 January 2012)

ABSTRACT

Two important questions are addressed in this paper using the Global Ensemble Forecast System (GEFS) from the National Centers for Environmental Prediction (NCEP): (1) How many ensemble members are needed to better represent forecast uncertainties with limited computational resources? (2) What is the relative impact on forecast skill of increasing model resolution and ensemble size? Two-month experiments at T126L28 resolution were used to test the impact of varying the ensemble size from 5 to 80 members at the 500-hPa geopotential height. Results indicate that increasing the ensemble size leads to significant improvements in the performance for all forecast ranges when measured by probabilistic metrics, but these improvements are not significant beyond 20 members for long forecast ranges when measured by deterministic metrics. An ensemble of 20 to 30 members is the most effective configuration of ensemble sizes by quantifying the tradeoff between ensemble performance and the cost of computational resources. Two representative configurations of the GEFS—the T126L28 model with 70 members and the T190L28 model with 20 members, which have equivalent computing costs—were compared. Results confirm that, for the NCEP GEFS, increasing the model resolution is more (less) beneficial than increasing the ensemble size for a short (long) forecast range.

Key words: NCEP operational GEFS, ensemble size, horizontal resolution, ensemble mean forecast, probabilistic forecast

Citation: Ma, J. H., Y. J. Zhu, R. Wobus, and P. X. Wang, 2012: An effective configuration of ensemble size and horizontal resolution for the NCEP GEFS. *Adv. Atmos. Sci.*, **29**(4), 782–794, doi: 10.1007/s00376-012-1249-y.

1. Introduction

Ensemble-based probabilistic forecasting, a feasible method to estimate forecast uncertainties, greatly improves and extends numerical forecast skill compared with deterministic forecasts (Zhu and Ma, 2010). Since 1992, the European Center for Medium-Range Weather Forecasting (ECMWF) and the National Centers for Environmental Prediction (NCEP) have implemented operational global ensemble forecast systems (GEFS). Nearly two decades later, Ensemble Prediction Systems (EPS) have been operationally implemented in many numerical weather prediction centers around the world, such as Canadian Meteorological Center, Met Office (UK), Japan Meteorological Agency and so on.

The operational ensemble forecast systems have undergone great developments. Besides the development of initial and model perturbation methods, it is also important to optimize the ensemble configura-

*Corresponding author: MA Juhui, juhui.ma@noaa.gov tion, including the optimization of ensemble size and the tradeoff between model resolution and ensemble size, to improve ensemble performance (Du et al., 1997; Buizza and Palmer, 1998; Buizza et al., 1999; Richardson, 2001; Mullen and Buizza, 2002) under the constraint of limited computational resources. Therefore, when designing an effective operational ensemble prediction system, there are two main questions we seek to answer, which are: (1) How many ensemble members do we need to better represent forecast uncertainties with limited computational resources? (2) What is the relative impact on forecast skill of increasing model resolution and ensemble size? In this study, these two questions will be investigated using the NCEP GEFS model.

In the past decade, some studies have addressed related topics. Buizza and Palmer (1998) analyzed the impact of 2, 4, 8, 16 and 32-member ensembles on the performance of the ECMWF EPS at the 500-hPa geopotential height. The authors concluded that any increase of ensemble size is strongly beneficial to forecasting. However, they anticipated that further improvement may be achieved with a higher ensemble size because their experiments were limited to 32 members. Mullen and Buizza (2002) assessed the effect of horizontal resolution and ensemble size on the ECMWF EPS for 24-h accumulated precipitation. Through the comparisons of $51\mathrm{T_L}159$, $51\mathrm{T_L}255$, $51\mathrm{T_L}319$, $15\mathrm{T_L}255$ and $15\mathrm{T_L}319$ ("51" or "15" refers to the number of ensemble members; "$\mathrm{T_L}$" represents spectral triangular truncation with Linear grid), they found that ensemble size is more important than model resolution for probabilistic precipitation forecasts, particularly for heavy precipitation. However, a decade ago the skill of precipitation forecasting was very limited, especially for longer lead times. On the other hand, Reynolds et al. (2011) evaluated the impact of resolution versus ensemble size tradeoffs on the performance of tropical winds and tropical cyclone tracks based on the Navy's Operational Global Atmospheric Prediction System (NOGAPS) ensemble forecast system using resolutions of T119, T159 and T239, with 33, 17 and 9 ensemble members, respectively. The authors found that increasing the resolution has a small impact on root mean square error (RMSE) of ensemble mean for wind speed, but improves Brier scores for 10-m wind speed. Buizza (2010) analyzed the impact of horizontal resolution increases from T95 to T799 with four ensemble members on the error growth of ECMWF forecasts. Results indicated that the effect of model resolution on forecast skill is strong in the short forecast range, weaker in the medium range and undetectable in the long range. However, all these studies used older numerical model systems or smaller ensemble sizes than we can run today. Moreover, each model behaves differently and a similar study with recent versions of NCEP GEFS has not been carried out. Therefore, it is helpful to study these questions using the current NCEP GEFS with large ensemble sizes, in order to plan its future development. In this study, the NCEP operational GEFS implemented in February 2010 was employed with up to 80 ensemble members to explore the impact of ensemble size and the relative impact of ensemble size and horizontal resolution on ensemble performance, including ensemble mean and probabilistic forecasting.

The purpose of this paper is to determine an effective configuration of ensemble size and resolution for the NCEP operational GEFS. Section 2 describes the models and experimental design. Section 3 introduces the verification procedures. In section 4, the impact of ensemble sizes on ensemble skill in the NCEP GEFS is examined. Section 5 determines the most effective ensemble size for the NCEP GEFS by quantifying the tradeoff between ensemble performance and the cost of computational resources. Section 6 analyzes the tradeoff between increasing the model resolution and ensemble size through experiments with the NCEP GEFS. And finally, section 7 sets out our conclusions based on this study.

2. Experimental design

The current NCEP operational GEFS runs 20 ensemble perturbation forecasts and one control forecast four times per day (0000 UTC, 0600 UTC, 1200 UTC and 1800 UTC) at T190 horizontal resolution and 28 hybrid vertical levels. The forecast output data are interpolated to 1° lat ×1° lon resolution from 0 to 384 forecast hours at 6-h intervals. The analysis is truncated from the T382L64 analysis provided by the NCEP Global Data Assimilation System (GDAS)/Gridded Statistical Interpolation (GSI). The initial perturbations are generated using the Ensemble Transform with Rescaling (ETR) method (see Appendix A).

The impacts of different ensemble sizes (up to 80) on the NCEP GEFS were studied. In order to run a relatively larger ensemble size at similar computation cost to higher horizontal resolution, the model horizontal resolution was reduced to T126. The (2-month) experiment runs from 1 December 2009 to 31 January 2010 and long forecasts (up to 16 days) were made once per day (the computing time was enough to complete the 16.5 years of single 16-day T126L28 forecasts). The perturbations were updated by ETR every six hours. At each cycle, the ETR orthogonalized and centralized the 80 perturbations. There were six variables selected for the evaluation: 500-hPa and 1000-hPa geopotential heights; 850-hPa and 2-m temperature; and 10-m u- and v-components of wind. However, in this paper we only present the results for 500-hPa geopotential height over the Northern Hemisphere (NH, 20°–80°N) and the Southern Hemisphere (SH, 20°–80°S).

The relative impact of increasing ensemble size and horizontal resolution on the NCEP GEFS was also assessed by comparing 70-member ensembles at T126 resolution (about 100 km) (70T126) with 20-member ensembles at T190 resolution (about 70 km) (20T190). The vertical resolution for both configurations was 28 levels. Initial analyses for both T126 and T190 were all truncated from T382/L64 GDAS/GSI analysis. Initial perturbations produced by ETR were centered on the initial analysis. The sample period and domain of variables were the same as in the experiment of ensemble sizes.

3. Verification methodology

The verification methods employed in this study include Root Mean Square Error (RMSE), ensemble spread (SPREAD) (Toth et al., 2003) and Pattern Anomaly Correlation (PAC) (Zhu, 2005; Wilks, 2006), which are used to assess the skill of the ensemble mean forecast, as well as probabilistic measures such as the Brier Skill Score (BSS) (Wilks, 2006) and Continuous Ranked Probabilistic Skill Score (CRPSS) (Toth et al., 2003; Wilks, 2006). The t-test method is used to estimate the statistical significance of differences among these scores for different configurations.

3.1 *Root Mean Square Error* (*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 a perfect ensemble forecast system, in which all of the uncertainties associated with initial errors and model errors are represented, the verifying analysis is statistically indistinguishable from the ensemble members and the SPREAD would be equal to RMSE.

To detail the impact of increased ensemble size on the ensemble mean forecast, we decompose the Mean Square Error (MSE), the square of RMSE, into the systematic mean square error term and the random mean square error term:

$$\begin{aligned}\text{MSE} &= \left\langle (F-A)^2 \right\rangle \\ &= \left\langle (\overline{F}-\overline{A})^2 \right\rangle + \left\langle (F'-A')^2 \right\rangle \ . \qquad (1)\end{aligned}$$

Here, F is the ensemble mean forecast and A is the verifying analysis. The angled brackets represent the mean that is taken over all points within the verification domain and the sample period. The overbar stands for a time mean and the prime denotes a deviation from the time mean.

3.2 *Pattern Anomaly Correlation* (*PAC*)

PAC measures the ability of the 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.

3.3 *Brier Skill Score* (*BSS*)

Brier score (BS) measures the mean-square error between the probabilistic forecasts and the subsequent categorical observation. BS can be decomposed into reliability (REL), resolution (RES) and uncertainty (UNC) components (Murphy, 1973). Reliability measures the statistical consistency between *a priori* predicted probabilities and *a posteriori* observed frequencies of the occurrence of the event. For a perfectly reliable system, REL should be 0. Resolution measures how the different forecast events are classified by a forecast system. The larger RES is, the better the forecast system is at identifying whether an event is likely to occur in the future. Uncertainty is the variance of the observations.

BSS expresses the percentage improvement in the BS relative to a reference forecast (usually a climatological forecast). A value of 1 (0 or negative) indicates a perfect system (an unskillful system). BSS can also be decomposed into the reliability skill score (RELSS) and the resolution skill score (RESSS).

3.4 *Continuous Ranked Probabilistic Skill Score* (*CRPSS*)

Continuous Ranked Probabilistic Score (CRPS) is used to measure the reliability and resolution of ensemble-based probabilistic forecasts by calculating the distance between the predicted and the observed cumulative distribution functions of scalar variables. CRPSS is the percentage improvement in the CRPS relative to a reference forecast (generally a climatological forecast). A value of 1 (0 or negative) indicates a perfect system (an unskillful system).

3.5 *Statistical significance testing*

In this study, the t-test is employed to assess whether the differences in the skills for two ensemble configurations are statistically significant. Suppose a sample size n is taken from a population with mean μ and standard deviation σ. Assume the population is a normal distribution and σ is unknown such that we can only estimate it with the sample standard deviation S. For a $100\times(1-\alpha)\%$ confidence interval, the difference is statistically significant if

$$\overline{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}} \geqslant 0$$

or

$$\overline{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}} \leqslant 0 ,$$

in which $\bar{X}$ is the sample mean of the data and $t_{\alpha/2}(n-1)$ is the critical value of the t distribution with (n-1) degrees of freedom. In this study, we have chosen α=0.05 to specify the 95% confidence interval.

4. Impact of ensemble size on ensemble performance

The different ensemble sizes were tested with the NCEP GEFS running at T126L28 resolution. Ensemble forecast scores were computed for ensemble sizes up to 80 members for the period 1 December 2009 to 31 January 2010. It is important to note that the initial perturbations were all from the ETR cycling of 80 members; the ETR was not run separately for small ensemble sizes (5, 10, 20, etc.). The results from different ensemble sizes may have been underestimated (previous study by Toth, personal communication), but the conclusions should remain valid. As a combination of a set of deterministic forecasts and as one of the most widely used techniques for producing probability forecasts, ensemble forecasts can not only be used to generate a deterministic forecast by averaging, but also converted into probabilistic products. In this study, the impact of increasing ensemble size on the ensemble mean forecast and probability forecast was assessed for the 500-hPa geopotential height over the NH and SH, from several scores in the NCEP standard verification package (Zhu et al., 1996, 2002; Zhu, 2004; Zhu and Toth, 2008), including RMSE, SPREAD, PAC, BSS and CRPSS.

4.1 *Impact of ensemble size on the ensemble mean forecast*

The ensemble mean forecast is usually superior to the control forecast because it provides a nonlinear filter that removes part of the forecast error that is due to initial error uncertainty, as long as the ensemble perturbations are representative of the initial uncertainties of the analysis (Leith, 1974; Toth and Kalnay, 1997). The main goals of this subsection are (1) to explore whether increasing the ensemble size has a positive effect on improving the capability for filtering of the ensemble mean forecast; and (2) to investigate the reasons for these results.

The RMSE and SPREAD as a function of ensemble size for forecast days 3, 7 and 13 over the NH are shown in Fig. 1a. The increases in ensemble sizes up to 20 members produced an obvious improvement of RMSE at all forecast ranges over the NH, and the improvement became larger with increasing forecast length. However, the improvements with further increases in ensemble size were negligible. SPREAD was insensitive to changes in ensemble sizes. Figure 1b shows a similar trend over the SH, but both RMSE and SPREAD were smaller than over the NH, which may have been due to seasonal variations of circulation patterns. The PAC score (not shown) revealed that there were about 18 hours (from hour 228 to hour 246, approximately) gained by increasing the ensemble size from 5 to 80 over the NH, if we consider a 60% PAC score as a useful skill level for large-scale weather forecasts.

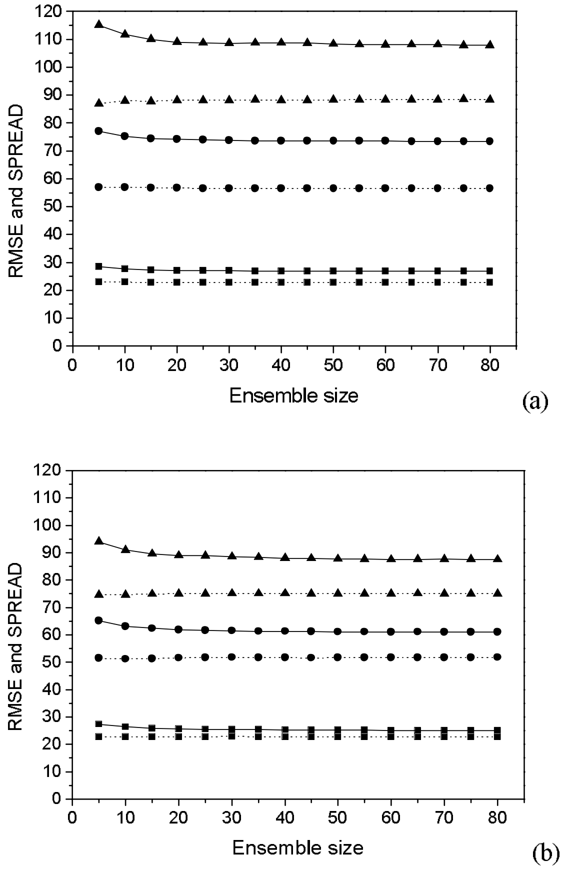

Fig. 1. RMSE (solid line) and SPREAD (dotted line) as a function of ensemble size at forecast day 3 (squares), day 7 (circles) and day 13 (triangles) at the 500-hPa geopotential height from 1 Dec 2009 to 31 Jan 2010 over (a) the NH and (b) SH.

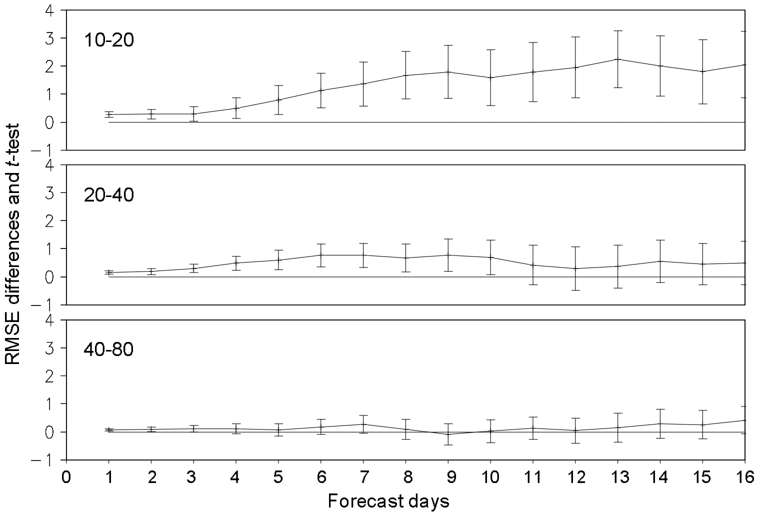

Fig. 2. The differences in RMSE between the different ensemble sizes. The vertical bars around the difference (solid line) are the 95% confidence limits.

Figure 2 shows whether the differences of RMSE between the different ensemble sizes were statistically significant. A vertical bar represents the 95% confidence interval. For example, the top panel shows the difference between 10 and 20 ensemble members. A positive value means 10 members had a larger RMSE value than 20 members. The bars do not cross zero, suggesting that these differences were significant at the 5% level. RMSE for 20 members differed significantly from 40 members for short lead times (less than 10 days, approximately). The difference between 40 and 80 was significant for shorter lead times.

To explore the reasons for error reduction with increases in ensemble size as shown above, we decomposed the mean square error into systematic mean square error and random mean square error components. The systematic mean square error term quantifies the difference between model and reality, so varying the ensemble size cannot be expected to remove this part of the error from the ensemble mean. Figure 3a confirms this point. To compare the systematic error components with different ensemble sizes, the systematic mean square errors of ensembles of different sizes are presented as their ratios to the systematic mean square error of the 80-member ensemble. We can see that the curves are hard to distinguish from one another. Therefore, only a small part of the improvement in the ensemble mean forecast with increasing ensemble size comes from differences in systematic error. The ratios of random mean square error in Fig. 3b show that a 10-member (20-member) ensemble had a 1.1 (1.05) times larger random error than an 80-member ensemble, and the error of a 40-member ensemble was

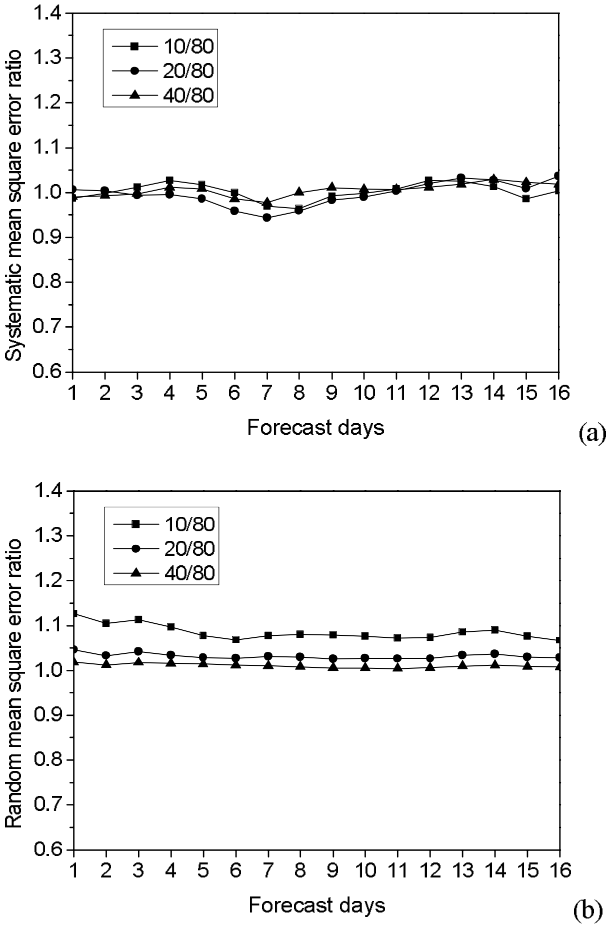

Fig. 3. (a) Systematic mean square errors and (b) random mean square errors for different ensemble sizes, expressed as ratios of the error for 80 members.

close to that of an 80-member ensemble. The RMSE gain was mainly due to a reduction of the random error component.

4.2 *Impact of ensemble size on probability forecast*

Buizza and Palmer (1998) found that an increase in ensemble size is beneficial to forecasting, but the extent of the improvement is strongly dependent on the measure used. This result is due in part to the negative bias in probabilistic verifications arising from the discretization and squaring measure, especially with ensemble sizes smaller than 10 (Richardson, 2001; Muller et al., 2005), and these effects become negligible when the ensemble size increases beyond 40. In most previous studies, the impacts have actually been overestimated due to the negative bias with relatively limited-sized ensembles (fewer than 40 members).

In this study, we used BSS and CRPSS as the probabilistic verification methods. Here, the definition of BSS is multi-category in terms of climatologically equal bins as thresholds, instead of the two-category (event occurs or does not occur) BSS used in Buizza and Palmer (1998). The multi-category BSS can reduce the negative bias to some extent. CRPSS extends the discrete category verifications to continuous (all-inclusive) measures, which can avoid this kind of bias arising from the discretization. BSS and CRPSS are shown against ensemble size for forecast days 3, 7 and 13 in Figs. 4a, b. BSS and CRPSS improved most rapidly when the ensemble size was smaller than 20, and the improvement became larger with an increasing lead time. For day 13, the 30-member ensemble just achieved a positive skill level in BSS, while smaller ensembles remained less skillful than climatology. The tendency over the SH was similar to the NH.

To detail the impact of ensemble size on probabilistic forecasts, we show in Fig. 5 the evolution in time of BSS (top panel) and its reliability (middle panel) and resolution (bottom panel) components, which measure two main attributes of the probabilistic system. For all tested ensembles over the NH, BSS decreased numerically with an increasing lead time, which was mainly due to the degradation of the resolution component, whereas the variation of the reliability component with lead time was significantly smaller. Figure 6a shows that the improvement of BSS for increasing the ensemble size from 10 to 20, from 20 to 40, and from 40 to 80, were all significant at the 95% confidence level for all forecast ranges. This is different from the results for ensemble mean verifications, although the extent of the improvement decreased as the ensemble size increased from 20 to 40 and from 40 to 80. The improvement of BSS for the short lead time came from the resolution gain, which slowly decreased with an increasing lead time. That may be because these ensembles all approached the climatological probability distribution with lead time, so that the differences in the ability to distinguish events in which forecast frequency differed from sample climatology (resolution) narrowed considerably, though they were still significant (Fig. 6c). For a long lead time, 40 and 80-member ensembles performed quite reliably, and the improvement in BSS came from this reliability gain. At these forecast ranges, the small size ensembles (fewer than 20) already had no skill (BSS<0), which was mainly due to their lower reliability. It is evident that ensemble systems should have more than 20 members. The situation over the SH was similar to the NH.

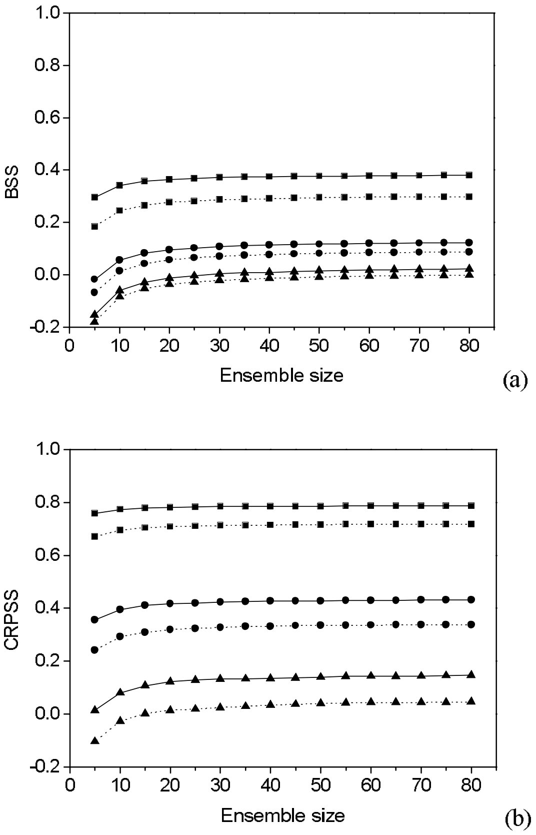

Fig. 4. (a) BSS and (b) CRPSS as a function of ensemble size for forecast day 3 (squares), day 7 (circles) and day 13 (triangles) at the 500-hPa geopotential height from 1 Dec 2009 to 31 Jan 2010 over the NH (solid line) and SH (dotted line).

5. The effective configuration of ensemble size for the NCEP GEFS

From section 4, we can conclude that for probabilistic forecasts, increases in ensemble size lead to significant performance improvements for all forecast

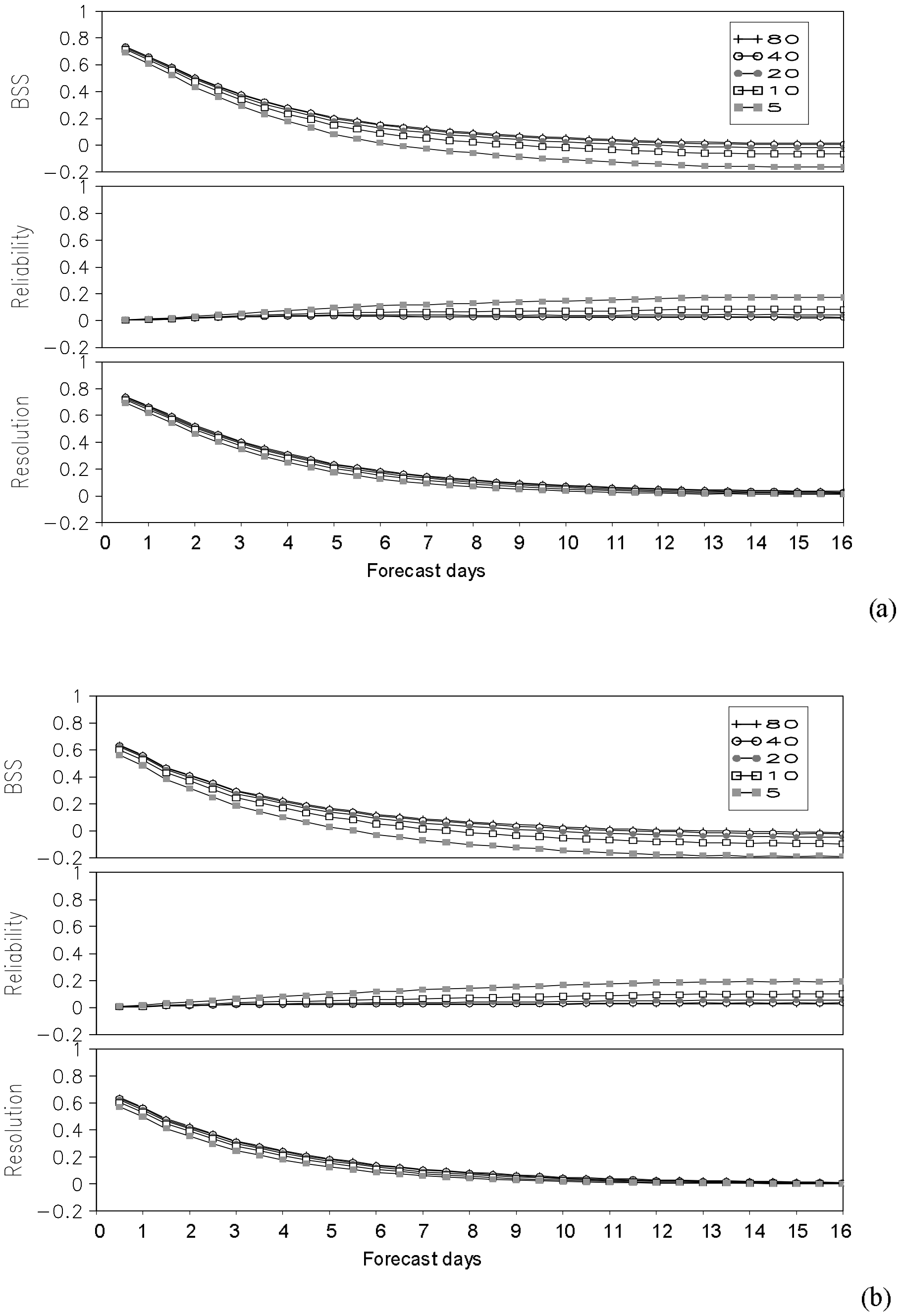

Fig. 5. BSS, RELSS and RESSS of different ensemble sizes at the 500-hPa geopotential height from 1 Dec 2009 to 31 Jan 2010 over (a) the NH; and (b) SH.

ranges, but improvement in the ensemble mean forecast is not significant beyond 20 members for long forecast ranges. These conclusions are based only on the impact of the ensemble performance, without considering the cost of computational resources. The optimal configuration may depend on the application. We can determine the most effective ensemble size for the NCEP operational GEFS by quantifying the tradeoff between ensemble performance and the cost of computational resources.

To remove the effect of different dimensions, we used the standardized RMSE and BSS as a function of ensemble size for every lead time to evaluate the impact of ensemble size on the ensemble performance. To estimate the cost of computational resources, the CPU time was used, which is proportional to the ensemble size (M) and the square of the spectral truncation (N), and is inversely proportional to the time

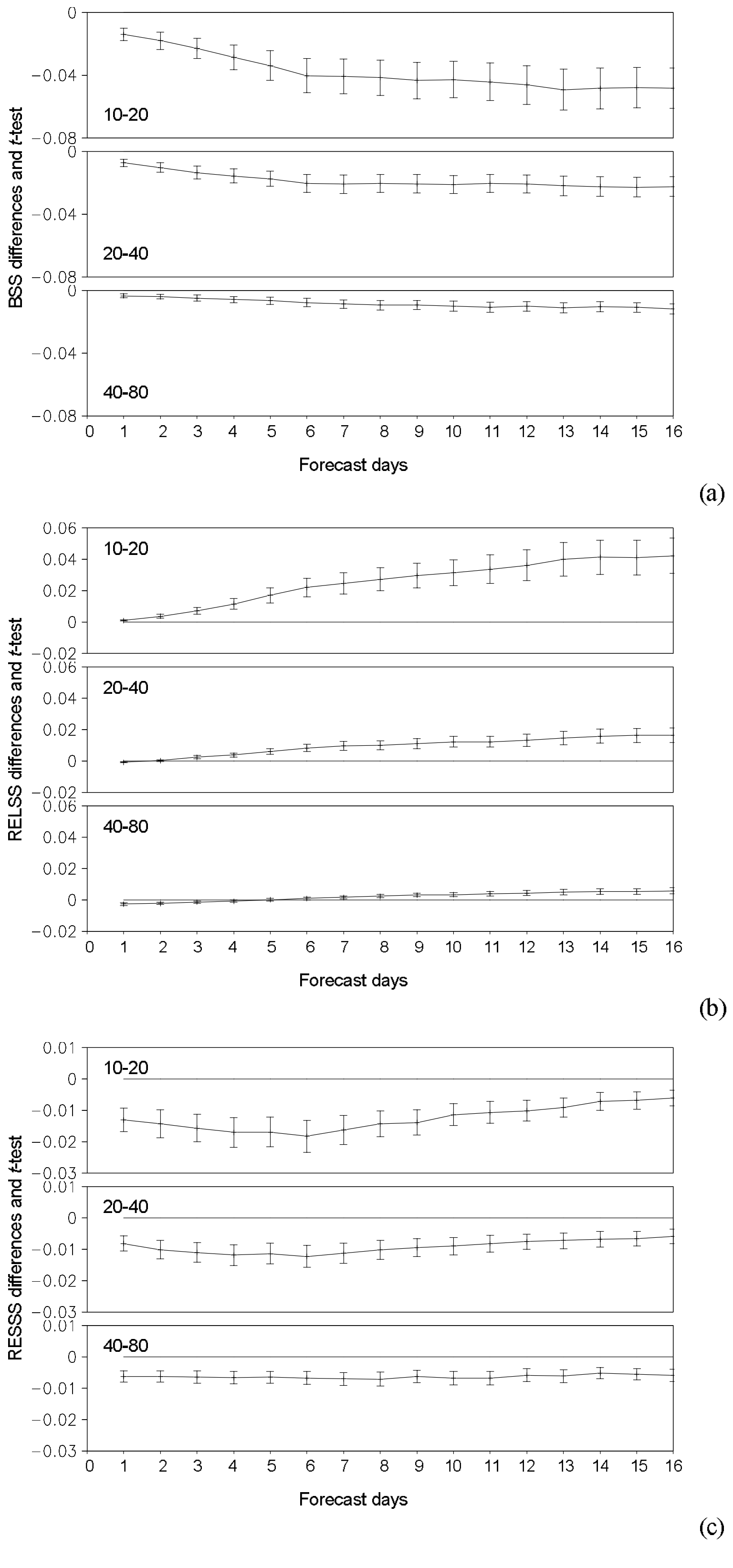

Fig. 6. The differences in (a) BSS; (b) RELSS; (c) RESSS between the different ensemble sizes. The vertical bars around the difference (solid line) are the 95% confidence limits.

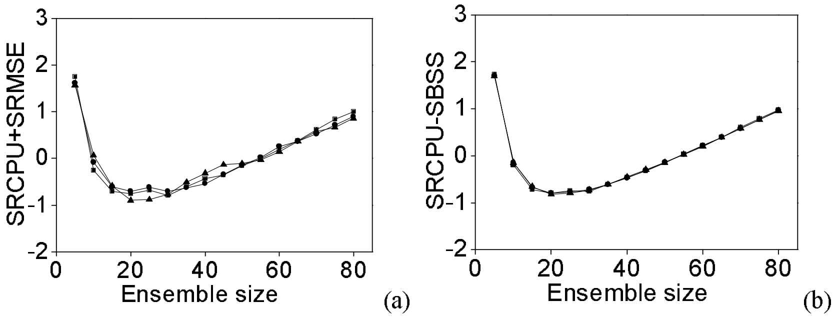

Fig. 7. (a) SRCPU+SRMSE and (b) SRCPU-SBSS as a function of ensemble size for forecast day 3 (squares), day 7 (circles) and day 13 (triangles) at the 500-hPa geopotential height from 1 Dec 2009 to 31 Jan 2010.

step (Δt) (Buizza, 2010). In this experiment, configurations with different ensemble sizes used the same spectral truncation (N=126) and the same time step (Δt=600 s), so the differences in their computational resources were only related to the ensemble size. The CPU time that each configuration requires was calculated relative to the 5-member configuration; for example, the configuration with 80 members required 16 times the CPU time of the 5-member ensemble, so its cost is stated as 16. Following this the standardization was carried out for the relative CPU time of these configurations. In this way we were able to define the cost of each configuration in terms of the Standardized Relative CPU time (SRCPU), which is dimensionless. The smaller the SRCPU is, the less computation resources cost, and vice versa.

For the ensemble mean forecast, the Standardized RMSE (SRMSE) was used to determine how many ensemble members we need to better represent forecast uncertainties with less computational resources, exploring the minimum of the sum of SRCPU and SRMSE. Figure 7a shows that numbers between 15 and 30 were the most effective ensemble sizes, jointly considering both ensemble performance and computer cost. The 5-member ensemble was more costly because forecast error was high, and ensembles of 50 or more members were costly in terms of CPU time.

The Standardized BSS (SBSS) measures the skill of probabilistic forecasts. The larger the SBSS is, the more skillful the system. The optimum is the minimum of the difference between SRCPU and SBSS. As in Fig. 7a, Fig. 7b shows that numbers between 15 and 30 were also the most effective ensemble sizes for probabilistic forecasts. However, from section 4 we know that ensembles with fewer than 20 members already have no skill, so we are able to conclude that 20 to 30 members were the most effective configurations of ensemble sizes.

6. The relative impact of ensemble size and horizontal resolution on ensemble performance

The relative impact of increasing the model resolution versus increasing the ensemble size was assessed by comparing 70 members at T126L28 resolution (70T126) with 20 members at T190L28 resolution (20T190). Since the CPU time is proportional to the ensemble size (M) and the square of the truncation (N), and inversely proportional to the time step (Δt) (Buizza, 2010), these configurations (70T126, N=126, M=70, Δt=600 s and 20T190, N=190, M=20, Δt=400 s) used equivalent computation resources.

From the comparison of RMSE in Fig. 8, we can see clearly that with similar computer resources and model physics, the model resolution played a more important role than ensemble size when the forecast lead time was less than 4 days, whereas a large ensemble size was significantly superior to a higher resolution when the forecast lead time exceeded 13 days. This means that using more ensemble members will benefit extended range forecasts. For 5–12-day forecast lead times, there was no significant difference between increasing the resolution and ensemble size.

The comparisons of BSS (Fig. 9a) indicate that increasing the ensemble size led to a significant improvement in probabilistic forecasts compared to increasing the model horizontal resolution, except for days 3, 4 and 5. For days 1 and 2, the advantage of the large ensemble was entirely due to the resolution component of BSS, and for the week-2 forecast range, the improvement for 70T126 came from joint gain in both reliability and resolution components. For days 3, 4 and 5, the reliability for 20T190 contributed to its improvement in BSS (Figs. 9b, c).

During the first two days, since 70T126 could sam-

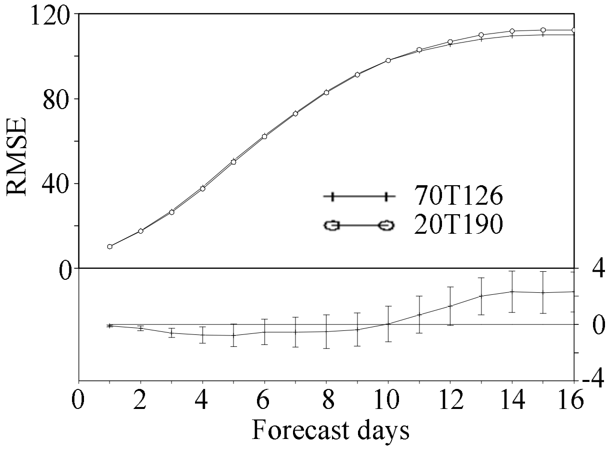

Fig. 8. RMSE (top) for 70T126 (crosses) and 20T190 (open circles) at the 500-hPa geopotential height from 1 Dec 2009 to 31 Jan 2010 over the NH. Differences in RMSE (bottom) between 70T126 and 20T90, with vertical bars representing the 95% confidence limits.

ple more initial uncertainties than 20T190 with the ETR initial perturbation method, whereas 20T190 could induce model uncertainty related error reduction, the skill of the two were basically equivalent. During the subsequent few days, an adequate model resolution was more crucial for model integration: 20T190 performed better than 70T126. However, as the lead time increased and nonlinearities became important, the effect of model-related uncertainty decreased and eventually became insignificant. Therefore, the ensemble size played a more important role than the model resolution.

Overall, based on the current NCEP GEFS, a higher model resolution is more important for short-range forecasts, while a larger ensemble size is better for longer range forecasts. Therefore, there is a tradeoff between model resolution and ensemble size configuration. These conclusions are similar to those of Mullen and Buizza (2002) and Buizza (2010) obtained from ECMWF GEPS, which we extended to larger numbers of ensemble members. To optimize the use of computational resources based on these conclusions, the Variable Resolution Ensemble Prediction System (VAREPS) was designed, which reduces the model resolution for long forecast lead times. Buizza et al. (2007) evaluated this system and concluded that VAREPS provides better forecasts in the early range without losing accuracy in the long range. Szunyogh and Toth (2002) reached a similar conclusion based on their experience with the NCEP system. Variable resolution will also be used in the next operational version of the NCEP GEFS to improve week-1 forecasts. Meanwhile, the North American Ensemble Forecast System (NAEFS) provides NCEP with an increased ensemble size. On the other hand, the use of lagged ensemble members could be an option to enhance week-2 or longer-range ensemble forecast skill. For this purpose, in the future, ensemble systems may be designed to have variable ensemble size configurations which enlarge the ensemble size for long forecast lead times.

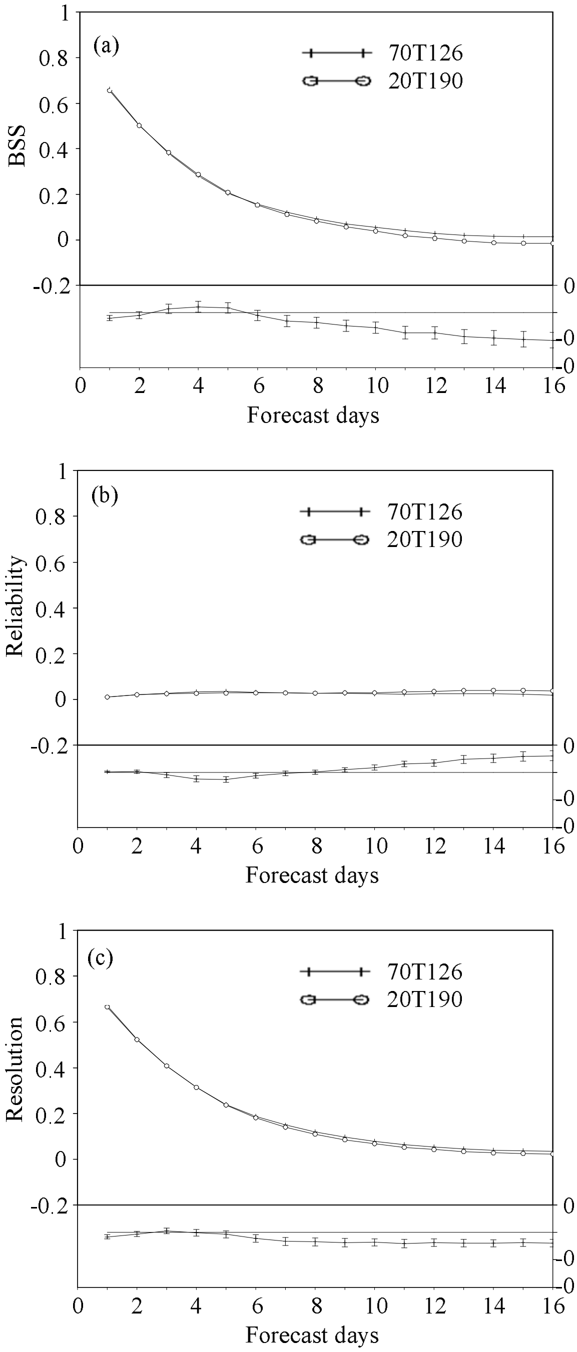

Fig. 9. As in Fig. 8 but for (a) BSS; (b) RELSS; and (c) RESSS.

7. Conclusions

This study compared the impact of different ensemble sizes and the relative performance of ensemble sizes versus model resolution to determine the effective configuration to improve the ensemble forecast skill. These issues arise every time NCEP developers plan an upgrade to their global ensemble forecast system. Although a similar topic has been studied at ECMWF (Buizza and Palmer, 1998; Mullen and Buizza, 2002; Buizza, 2010), the present work addresses the need for an up-to-date study of the current NCEP GEFS with more detailed analyses. This study, based on the current NCEP GEFS with ensemble sizes of up to 80 members, will be helpful in planning its future development.

The 2-month experiments from the NCEP GEFS at T126L28 resolution were used to test the impact of ensemble size. The measures, such as RMSE, SPREAD, PAC, BSS and CRPSS, were applied to evaluate the benefits of increasing the ensemble size at the 500-hPa geopotential height over the NH and SH. It was found that increases in the ensemble size led to significant improvements in the performance for all forecast ranges when measured by probabilistic metrics, but was not significant beyond 20 members for long forecast ranges when measured by deterministic metrics. To detail the impact of ensemble size on the ensemble mean forecast and probabilistic forecast respectively, we decomposed the mean square error into systematic mean square error and random mean square error components, and further decomposed BSS into reliability and resolution partitions. Results indicated that the reduction of ensemble mean error was mainly due to an improvement in the reduction of the random error. The improvement in probabilistic forecast skill for the short lead time came from the resolution gain, which slowly decreased with increasing lead time. For the long lead time, it came from the reliability gain. To determine the most effective ensemble size for the NCEP operational GEFS, we quantified the tradeoff between ensemble performance and the cost of computational resources. Results indicated that 20 to 30 members are the most effective configurations of ensemble sizes.

Developers of the NCEP GEFS always face the issue of optimizing computational resources. They usually compromise between increasing the model resolution and enlarging the ensemble size. The relative benefits of the T126L28 model with 70 members and the T190L28 model with 20 members, which have equivalent computing costs, were compared at the 500-hPa geopotential height over the NH. Results indicated that increasing the model resolution was more (less) beneficial than increasing the ensemble size for short (long) lead times. Therefore, a variable resolution system and the use of lagged ensemble members are options to enhance week-1 and week-2 forecast products of the NCEP GEFS, respectively.

Based on these results, we will continue our research using global analyses from future hybrid EnKF/ETR initialization in the NCEP GEFS with variable model resolutions and lagged ensembles. In doing so, we hope to show clearly whether lagged ensembles improve the statistical reliability of the week-2 forecasts. This work will also focus on the impact of ensemble size and resolution for the summer season, in order to evaluate the probabilistic quantitative precipitation forecast (PQPF) and the tropical cyclone forecast.

Acknowledgements. 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 also gratefully acknowledges the support of Dr. Stephen J. Lord and the EMC.

APPENDIX A

The Ensemble Transform with Rescaling (ETR) Method

In the ETR scheme (Wei et al., 2006; Wei et al., 2008), the analysis perturbations matrix $\boldsymbol{X}_{\rm a}$ are generated from the forecast perturbations matrix $\boldsymbol{X}_{\rm f}$ through an ensemble transformation matrix $\boldsymbol{T}$ as follows:

$$\boldsymbol{X}_{\rm a}=\boldsymbol{X}_{\rm f}\boldsymbol{T}\ , \tag{A1}$$

where k analysis perturbations $x'_{{\rm a}i}$ (i=1, 2, …, k) are listed as columns in the matrix $\boldsymbol{X}_{\rm a}$, and k forecast perturbations $x'_{{\rm f}i}$ (i=1, 2, …, k) are listed as columns in the matrix $\boldsymbol{X}_{\rm f}$. In the ensemble representation, the analysis covariance matrix $\boldsymbol{P}_{\rm a}$ is approximated as:

$$\boldsymbol{P}_{\rm a}=\boldsymbol{Z}_{\rm f}\boldsymbol{T}\boldsymbol{T}^{\rm T}\boldsymbol{Z}_{\rm f}^{\rm T}\ ; \tag{A2}$$

namely:

$$\boldsymbol{T}^{\rm T}(\boldsymbol{Z}_{\rm f}^{\rm T}\boldsymbol{P}_{\rm a}^{-1}\boldsymbol{Z}_{\rm f})\boldsymbol{T}=\boldsymbol{I}\ , \tag{A3}$$

where $\boldsymbol{Z}_{\rm f}=\boldsymbol{X}_{\rm f}/\sqrt{k-1}$, $\boldsymbol{I}$ is the identity matrix and superscript T represents the matrix transpose.

To obtain the solution of the transformation matrix $\boldsymbol{T}$, we need to solve the eigenvalues and eigenvectors of $\boldsymbol{Z}_{\rm f}^{\rm T}\boldsymbol{P}_{\rm a}^{-1}\boldsymbol{Z}_{\rm f}$:

$$\boldsymbol{Z}_{\rm f}^{\rm T}\boldsymbol{P}_{\rm a}^{-1}\boldsymbol{Z}_{\rm f}=\boldsymbol{C}\boldsymbol{\Gamma}\boldsymbol{C}^{-1}\ , \tag{A4}$$

where columns of the matrix $\boldsymbol{C}$ contain the orthonormal eigenvectors (c_i, i=1, 2, ..., k) of $\boldsymbol{Z}_\mathrm{f}^\mathrm{T}\boldsymbol{P}_\mathrm{a}^{-1}\boldsymbol{Z}_\mathrm{f}$, and the diagonal matrix $\boldsymbol{\Gamma}$ contains the corresponding eigenvalues $(\lambda_i, i = 1, 2, \ldots, k)$. From Wei et al. (2008), we know that the first k-1 eigenvalues are non-zero and the last eigenvalue is zero. We define a diagonal matrix $\boldsymbol{\digamma}$ by setting the zero eigenvalue in $\boldsymbol{\Gamma}$ to a non-zero constant α. Here, the diagonal matrix $\boldsymbol{P}_\mathrm{a}$ contains the analysis error variances obtained from the NCEP data assimilation system. Therefore, from (A3) and (A4), the transformation matrix $\boldsymbol{T}_\mathrm{p}$ instead of $\boldsymbol{T}$ can be constructed by:

$$\boldsymbol{T}_\mathrm{p} = \boldsymbol{C}\boldsymbol{\digamma}^{-1/2} \, . \tag{A5}$$

The perturbations $\boldsymbol{X}_\mathrm{ap}$ can be generated through $\boldsymbol{T}_\mathrm{p}$ as:

$$\boldsymbol{X}_\mathrm{ap} = \boldsymbol{X}_\mathrm{f}\boldsymbol{T}_\mathrm{p} = \boldsymbol{X}_\mathrm{f}\boldsymbol{C}\boldsymbol{\digamma}^{-1/2} \, . \tag{A6}$$

$\boldsymbol{X}_\mathrm{ap}$ are orthogonal perturbations after transformation, but not centered. If the initial perturbations are re-centered around the analysis, the performance of the ensemble mean will be better. The following can be used to center the perturbations:

$$\boldsymbol{X}_\mathrm{a} = \boldsymbol{X}_\mathrm{ap}\boldsymbol{C}^\mathrm{T} \, . \tag{A7}$$

Though centering the initial perturbations destroys their orthogonality, this effect decreases with increasing ensemble size (Wei et al., 2008).

To make the initial spread similar to the analysis error covariance, $\boldsymbol{X}_\mathrm{a}$ is rescaled using factor γ, which is defined as the ratio of the square root of kinetic energy from $\boldsymbol{P}_\mathrm{a}$ and the square root of kinetic energy from $\boldsymbol{X}_\mathrm{a}$ at each grid point.

REFERENCES

Buizza, R., 2010: Horizontal resolution impact on short and long range forecast error. *Quart. J. Roy. Meteor. Soc.*, **136**, 1020–1035.

Buizza, R., and T. Palmer, 1998: 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, 1998: 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.

Buizza, R., J.-R. Bidlot, N. Wedi, M. Fuentes, M. Hamrud, G. Holt, and F. Vitart, 2007: The new ECMWF VAREPS (Variable Resolution Ensemble Prediction System). *Quart. J. Roy. Meteor. Soc.*, **133**, 681–695.

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

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418.

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.

Muller, W. A., C. Appenzeller, F. J. Doblas-Reyes, and M. A. Liniger, 2005: A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes. *J. Climate*, **18**, 1513–1523.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595-600.

Reynolds, C. A., J. G. McLay, J. S. Goerss, E. A. Serra, D. Hodyss, and C. R. Sampson, 2011: Impact of resolution and design on the U.S. navy global ensemble performance in the tropics. *Mon. Wea. Rev.*, **139**, 2145–2155.

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.

Szunyogh, I., and Z. Toth, 2002: The effect of increased horizontal resolution on the NCEP global ensemble mean forecasts. *Mon. Wea. Rev.*, **130**, 1125–1143.

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. *Forecast Verification: A practitioner's Guide in Atmospheric Science,* Jolliffe and 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 A*, **58**, 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 A*, **60**, 62–79.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences.* 2nd ed., Academic Press, 627pp.

Zhu, Y., 2004: Probabilistic forecasts and evaluations based on global ensemble forecast system. Vol. 3, *World Scientific Series on Meteorology of East Asia*, 277–287.

Zhu, Y., 2005: Ensemble forecast: A new approach to uncertainty and predictability. *Adv. Atmos. Sci.*, **22**, 781–788.

Zhu, Y., and Z. Toth, 2008: Ensemble based probabilistic forecast verification. Preprints, *19th Conference on Probability and Statistics*, Amer. Meteor. Soc., New Orleans, Louisiana, 1–6.

Zhu, Y., and J. Ma, 2010: Predictability, probabilistic forecasting and ensemble prediction system. *Lecture Notes on Numerical Weather Prediction*, WMO Regional Training Center, NUIST, China, 43–60.

Zhu, Y., G. Iyengar, Z. Toth, S. Tracton, and T. Marchok, 1996: Objective evaluation of the NCEP Global Ensemble Forecasting System. Preprints, *15th Conference on Weather Analysis and Forecasting*, Amer. Meteor. Soc., Norfolk, Virginia, 1–4.

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