ADVANCES IN ATMOSPHERIC SCIENCES, VOL. 30, NO. 5, 2013, 1287–1302

Effect of Doubling the Ensemble Size on the Performance of Ensemble Prediction in the Warm Season Using MOGREPS Implemented at the KMA

Jun Kyung KAY[1], Hyun Mee KIM*[1], Young-Youn PARK[2], and Joohyung SON[2]

[1] *Department of Atmospheric Sciences, Yonsei University, Seoul, Republic of Korea,* 120–749

[2] *Korea Meteorological Administration, Seoul, Republic of Korea,* 156–720

(Received 30 April 2012; revised 30 October 2012; accepted 10 December 2012)

ABSTRACT

Using the Met Office Global and Regional Ensemble Prediction System (MOGREPS) implemented at the Korea Meteorological Administration (KMA), the effect of doubling the ensemble size on the performance of ensemble prediction in the warm season was evaluated. Because a finite ensemble size causes sampling error in the full forecast probability distribution function (PDF), ensemble size is closely related to the efficiency of the ensemble prediction system. Prediction capability according to doubling the ensemble size was evaluated by increasing the number of ensembles from 24 to 48 in MOGREPS implemented at the KMA. The initial analysis perturbations generated by the Ensemble Transform Kalman Filter (ETKF) were integrated for 10 days from 22 May to 23 June 2009. Several statistical verification scores were used to measure the accuracy, reliability, and resolution of ensemble probabilistic forecasts for 24 and 48 ensemble member forecasts. Even though the results were not significant, the accuracy of ensemble prediction improved slightly as ensemble size increased, especially for longer forecast times in the Northern Hemisphere. While increasing the number of ensemble members resulted in a slight improvement in resolution as forecast time increased, inconsistent results were obtained for the scores assessing the reliability of ensemble prediction. The overall performance of ensemble prediction in terms of accuracy, resolution, and reliability increased slightly with ensemble size, especially for longer forecast times.

Key words: ensemble prediction, ensemble size, ensemble transform Kalman filter

Citation: Kay, J. K., H. M. Kim, Y.-Y. Park, and J. Son, 2013: Effect of doubling the ensemble size on the performance of ensemble prediction in the warm season using MOGREPS implemented at the KMA. *Adv. Atmos. Sci.*, **30**(5), 1287–1302, doi: 10.1007/s00376-012-2083-y.

1. Introduction

The atmosphere is a chaotic system; therefore, small errors in initial conditions can grow fast and eventually hinder the skill of forecasts (Lorenz, 1963). The imperfection of the numerical model is associated with model error (i.e. imperfect physical parameterizations of subgrid-scale motions and insufficient model resolution) and initial condition uncertainties. These two types of uncertainties limit the predictability of a deterministic forecast, so it is necessary to consider the information associated with both uncertainties. If we have a perfect model and know the probability distribution of initial uncertainties, the temporal evolution of uncertainties described by a probability density function (PDF) can be estimated by the Liouville equation (Ehrendorfer, 1994a, b). However, this approach is not feasible in current numerical weather prediction systems due to the enormous computational resources that would be required and our lack of knowledge about uncertainty sources.

Ensemble prediction was devised to quantify the statistical sample of forecast uncertainties that are represented by a spread of finite ensembles. Different initial conditions drawn from the probability distribution of initial states are integrated through a numerical forecast model to estimate the uncertainties of the future state of the atmosphere. However, the limited

*Corresponding author: Hyun Mee KIM, khm@yonsei.ac.kr

size of ensemble members—in the order of $O(10^1$–$10^2)$ in a typical operational ensemble prediction system (EPS)—causes potential problems in representing the statistics of the model state in the order of $O(10^7)$. The finite size of the ensemble introduces sampling error that is roughly inversely proportional to the size N of the ensemble (specifically, $N^{-1/2}$) (Casella and Berger, 1990). This sampling error will lead to inbreeding or underestimation of the forecast error covariance (Ehrendorfer, 2007). As the ensemble size is small in most operational EPSs, physically meaningless correlations between state components that are far from each other increases in the forecast error covariance. In other words, long-range spurious correlations develop (Anderson, 2001). In addition, fewer members than the degrees of freedom of the forecast model cannot approximate the full space composed of model states and there is rank deficiency in estimating the forecast error covariance.

Even though it may be feasible to increase the number of ensemble members in operational EPSs with improvements in computational resources, it is necessary to determine whether the impact of a larger ensemble is significantly effective enough to warrant the additional computations that need to be performed in the operational EPS. Therefore, several studies have investigated the impacts of ensemble size on the performance of the EPS. Using the EPS of the European Centre for Medium-Range Weather Forecasts (ECMWF) that uses the dominant singular vectors (SV) of the model to produce initial perturbations, Buizza and Palmer (1998) and Mullen and Buizza (2002) showed that the benefits of increasing ensemble size from 10 to about 30 are significant, but further increases in ensemble size decrease the impacts of an increased ensemble size. It is necessary to increase the ensemble size to predict rare phenomena, but the impact of increasing ensemble size is smaller if ensemble forecasts are first post-processed (i.e. calibrated using climatological observations) (Wilks, 2002). Atger (1999) showed that increasing ensemble size improves ensemble spread, but the effect on the accuracy of the ensemble mean is not significant in the ECMWF EPS system. Using version 3 of the Community Climate Model (CCM3), Wang and Bishop (2003) compared the performance of 8- and 16-member ensembles generated by the Ensemble Transform Kalman Filter (ETKF) method and showed that a larger ensemble size improves the estimate of analysis error variance and reduces spurious long-distance correlations.

Even though the abovementioned studies investigated the effect of ensemble size using specific EPSs, it has not been investigated in the Met Office Global and Regional Ensemble Prediction System (MOGREPS). The effect of ensemble size on the performance of ensemble prediction may be different in different EPSs with different specific configurations (i.e. different ensemble generation methods, models etc.). Further, because the Korea Meteorological Administration (KMA) has implemented the Unified Model (UM) and related pre-/post-processing system imported from the United Kingdom Meteorological Office (UKMO) operationally since 2011, it was necessary to investigate the effect of doubling ensemble size on operational ensemble prediction using MOGREPS implemented at the KMA (hereafter KMA MOGREPS). The results of doubling ensemble size would be used to determine which member size is appropriate for better sampling of initial perturbations for the EPS at the KMA. Therefore, in this study, the effect of doubling ensemble size on the performance of ensemble prediction was evaluated in KMA MOGREPS. Because the operational ensemble size of KMA MOGREPS is 24, the effect of doubling the ensemble size from 24 to 48 was evaluated. Section 2 describes KMA MOGREPS, including the perturbation method used to generate the initial conditions of the ensembles, inflation, experimental design, and the various verification methods used in the study. In section 3, the EPS performance of both the 24- and 48-ensemble-size experiments are compared and the effect of inflation on both experiments discussed. Finally, in section 4, a summary and discussion are provided.

2. Methodology

2.1 *The MOGREPS system*

MOGREPS comprises global and regional ensemble systems (Bowler et al., 2008). Because the global component of MOGREPS is used in this study, only the global component of the system is discussed. The global component of MOGREPS has 24 ensemble members (23 perturbed members and one unperturbed control member) with a horizontal resolution of around 40 km and 50 vertical levels (N320L50). The initial perturbations are generated by local ETKF (Bowler et al., 2009). The 23 initial perturbations are added to the analysis that is derived from the four-dimensional variational data assimilation (4DVAR) system of the UM, and these are then integrated forward for 10 days using a deterministic forecast model (UM) with the same resolution. Model uncertainties in MOGREPS are addressed by stochastic-physics schemes consisting of "random parameters" and "stochastic convective vorticity" schemes (Bowler et al., 2008). The variables used in MOGREPS are the horizontal components of wind (u' and v'), potential temperature (θ'), exner pressure (π'), and specific humidity (q') perturbations. KMA MOGREPS, used in this study, is basically the same as the original MOGREPS.

The overall performance of MOGREPS relative to other EPSs was shown in Park et al. (2008) which compared the performance of eight centers' EPSs using THORPEX Interactive Grand Global Ensemble (TIGGE) data. In terms of the RMSE of the ensemble mean forecast, the ECMWF was the best, followed by the UKMO and Japan Meteorological Agency (JMA). The difference between the ensemble spread and RMSE of the ensemble mean forecast, a measure of reliability, showed that the ECMWF, the Meteorological Service of Canada (MSC), and the UKMO ensembles were superior than other EPSs. For probabilistic prediction, the ECMWF ensemble showed the highest ranked probability skill score (RPSS), followed by the other four centers [MSC, UKMO, JMA and NCEP (National Centers for Environmental Prediction)].

2.2 *Generating initial perturbations*

2.2.1 *Local ETKF*

ETKF is a family of ensemble square root filters (Tippett et al., 2003) that can be used to calculate the initial analysis perturbations from forecast perturbations without updating the ensemble mean fields. The initial analysis perturbations $\boldsymbol{X}_{\rm a}$ can be written as

$$\boldsymbol{X}_{\rm a} = \boldsymbol{X}_{\rm f}\boldsymbol{T}\boldsymbol{\Pi} \ , \tag{1}$$

where $\boldsymbol{X}_{\rm f}$ is a matrix that has forecast perturbations as its components, $\boldsymbol{T}$ is a transform matrix that combines $\boldsymbol{X}_{\rm f}$ to $\boldsymbol{X}_{\rm a}$ linearly, and $\boldsymbol{\Pi}$ is a matrix consisting of inflation factors. The two perturbation matrices $\boldsymbol{X}_{\rm f}$ and $\boldsymbol{X}_{\rm a}$ are the square root vectors of the forecast and analysis covariance matrices, respectively, with the same rank as the covariance matrices. More details of ETKF formulation are given by Wang and Bishop (2003) and Bowler et al. (2008).

The resulting initial analysis perturbations normalized by the square root of observation error are orthogonal to each other in the observational space (Wang and Bishop, 2003) and can approximate the uncertainty of the states by reflecting both the forecast error statistics and observational information.

All observations used in the 4DVAR data assimilation system of the UM are used to calculate the transform matrix in MOGREPS. The observations used in calculating the initial perturbations are the conventional observations collected for the KMA's operational data assimilation system. For the period considered in this study, these observations were taken from the Met Office because the assimilation system of the KMA UM was under trial operation at that time.

Ensemble-based data assimilation systems use the covariance localization to decrease long-range spurious correlations that occur due to limited ensemble size. In MOGREPS, the localization of ETKF is realized by dividing the globe into 92 centers of approximately equal distance; ETKF is then calculated using the observations within the specific radius of influence from each center and this improves the spread of ensemble forecasts as a function of latitude (Bowler et al., 2009). For vertical localization in ETKF, the 50 vertical levels are divided into four bands that cover the PBL, troposphere, the stratosphere up to about 45 km altitude, and the stratosphere from 45 km to 60 km altitude. Transform matrices and inflation factors are calculated for four bands using the observations in corresponding levels.

2.2.2 *Inflation factor*

When the ensemble size is much smaller than the degrees of freedom of the model state, the total analysis error covariance is underestimated because the forecast error covariances are not fully estimated by the ensemble members. To remedy this problem, MOGREPS uses the inflation factor $\boldsymbol{\Pi}$, by the elements of which transformed forecast perturbations are increased. The inflation factor is defined as

$$\boldsymbol{\Pi}_n = \boldsymbol{\Pi}_{n-1}\sqrt{\frac{\mathrm{trace}(\boldsymbol{d}_n\boldsymbol{d}_n^{\rm T}) - \mathrm{trace}(\boldsymbol{R})}{\mathrm{trace}(\boldsymbol{H}(\boldsymbol{X}_{{\rm f},n})\boldsymbol{H}(\boldsymbol{X}_{{\rm f},n})^{\rm T})}} \ , \tag{2}$$

where the subscript n represents the time step, $\boldsymbol{d}_n = \boldsymbol{y}_n - \overline{\boldsymbol{H}(\boldsymbol{x}_{f,n})}$ is an innovation vector (i.e. differences between observations $\boldsymbol{y}_n$ and the 12-h ensemble-mean forecast $\boldsymbol{x}_{f,n}$ in observational space verified at observation time), and $\boldsymbol{H}$ represents the observational operator.

The inflation factor ensures that the sum of the ensemble 12-h forecast spread matches the sum of the error variation of the ensemble-mean 12-h forecast in the observational space over the target region for cycle n (Bowler et al., 2008, 2009). In practice, three steps are generally required to implement the inflation factor in MOGREPS. First, using Eq. (2), inflation factors for two categories are calculated using radiosonde and Advanced TIROS (Television Infrared Observation Satellites) Operational Vertical Sounder system (ATOVS) observations respectively, with 12-h ensemble forecasts. Second, the inflation factors of the two categories are combined to give only one inflation factor that is applied to transformed perturbations. Third, to prevent the initial analysis perturbations from being too large, the square root of the sum of the squares of the transform matrix elements (hereafter the magnitude of the transform matrix) and the magnitude of inflation factor elements are compared. It is expected that the magnitude of the initial analysis perturbations will be smaller than that of the forecast perturbations, because in relation to the data assimilation scheme, the uncertainty of analysis error is estimated by correcting the forecast error with optimally weighted observational error information. If the magnitude of inflation factor elements derived from the second step is larger than the magnitude of the transform matrix that decreases the magnitude of the forecast perturbations, the inverse of the magnitude of the transform matrix is substituted for the inflation factor element. In this way, the initial analysis perturbations have a similar magnitude to the forecast perturbations instead of being too large. The above-mentioned three steps are described in the Appendix in more detail.

2.3 *Experimental design*

The experiments using 24 and 48 (47 perturbed forecasts and one unperturbed control forecast) ensemble members are referred to as M24 and M48, respectively. For a fair comparison, the configurations of both experiments, including the algorithm to construct the ETKF, were the same except for the randomly modulated parameters used in the stochastic physics scheme. The 10-day forecasts of M24 and M48 were implemented and compared for the one-month period from 22 May to 23 June 2009, using the CRAY X1E supercomputer of the KMA. To perform 10-day ensemble forecasts, 24 processors were dedicated per one ensemble member, and almost 10 terabytes of data were stored for both the M24 and M48 experiments. Due to the limitation of computational resources and storage, it took about 60 days to complete the M48 and M24 experiments.

The initial perturbation generation and 10-day ensemble forecast processes are shown in Fig. 1. The KMA analysis with the same resolution as MOGREPS was used as the verification reference and the verification was done for the 500-hPa height variable over the latitudinal region of 20°–90°N in the Northern Hemisphere (NH) and 20°–90°S in the Southern Hemisphere (SH). The verification results over the tropical region (20°S–20°N) are not shown because the results were similar to those obtained over the SH region. The verification for the 850-hPa temperature also shows similar results as that for the 500-hPa height. Even though the global ensemble runs twice a day at 0000 and 1200 UTC, only the ensemble forecasts at 0000 UTC were used in this study due to limited computational resources and storage space.

2.4 *Verification methods*

The assessment of ensemble prediction is related to the verification of ensemble-based probabilistic forecasts, because ensemble prediction is implemented primarily to evaluate the probabilistic forecast (Atger, 2004). Ensemble prediction can be verified considering two statistical aspects: reliability and resolution (Leutbecher and Palmer, 2008). Reliability implies statistical consistency between the ensemble forecast and observations. Sampling error due to finite ensemble size in realistic systems decreases reliability (Richardson, 2001). Although reliability is a critical

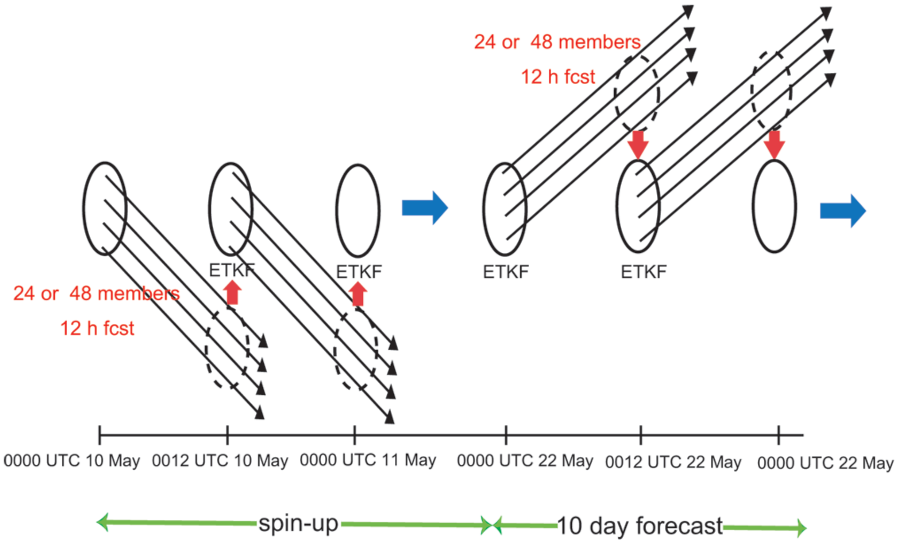

Fig. 1. A schematic diagram showing the initial perturbation generation using ETKF and 10-day ensemble forecast processes.

aspect of probabilistic forecasts, it is not a sufficient condition for a high-quality ensemble prediction. A system that always forecasts the climatological probability of the atmospheric state is perfectly reliable, but it is not useful for weather forecasting. Therefore, an additional property of probabilistic forecasts, referred to as resolution, should also be considered. Resolution is the degree to which ensemble forecasts separate the different observed events. In other words, resolution measures the variability of the frequency of observed events when the given forecast varies.

To verify the reliability and resolution of an ensemble prediction, it is appropriate to use a suite of verification measures and to understand the results considering the emphases different scores have, because the different verification methods inform different aspects of the ensemble forecast (Wei et al., 2008). The necessary verification scores and diagrams to assess the quality of ensemble prediction are probabilistic skill scores (such as the Brier score, Brier skill score, ranked probability score, and ranked probability skill score), reliability diagrams, relative operating characteristic (ROC) curves, and rank histograms (Hamill et al., 2000). The Brier Score (BS), reliability diagram, and ROC curves are for probability forecasts of dichotomous events. Conversely, the Ranked Probability Score (RPS) is for probability forecasts of continuous variables treated as categorical forecasts. BS and RPS measure the degree of reliability and resolution of probability forecasts, but RPS is the average of the BS for the priori defined thresholds (Candille and Talagrand, 2005). Reliability diagrams are conditioned on the forecast and divide the reliability from the resolution (Hacker et al., 2011). ROC curves only measure the resolution and often provide similar results with the resolution term of the BS qualitatively. In terms of reliability, additional measures such as the rank histogram and relation between ensemble mean error and spread, can describe more detailed characteristics and causes of reliability of the probabilistic forecast rather than the reliability diagram or the reliability term of the BS (Jolliffe and Stephenson, 2003).

2.4.1 *Brier Score*

The BS measures the mean squared error of the probability forecast of the occurrence of a dichotomous event as follows (Wilks, 2006):

$$\mathrm{BS}=\frac{1}{M}\sum_{k=1}^{M}(z_k-o_k)^2\ , \tag{3}$$

where z_k indicates the forecast probability, k denotes an index of M forecast event pairs, and the observational probability o_k is defined as o_k=0 if the event does not occur and o_k=1 if the event does occur. The BS can be decomposed into reliability and resolution components (Murphy, 1973). While the reliability term of the BS will be small if the forecast system has good reliability, the resolution term will be large if the forecast sorts events well. The skill score of BS [the Brier Skill Score (BSS)] is positively oriented; that is, a forecast of good quality has a score of one (Wilks, 2006). For reference probabilistic forecasts, we used the NCEP–NCAR 40-Year Reanalysis (Kalnay et al., 1996).

2.4.2 *Ranked Probability Score*

The RPS is equivalent to the BS, but it measures the accuracy of probability forecasts when there are multiple probability categories (Epstein, 1969; Murphy, 1971) as follows:

$$\mathrm{RPS}=\frac{1}{M}\sum_{k=1}^{M}\left[\sum_{c=1}^{J}\left(\sum_{j=1}^{c}z_j-\sum_{j=1}^{c}o_j\right)^2\right]_k\ , \tag{4}$$

where c is any number of J categories over which to distribute the probability.

The skill score of the RPS [Ranked Probability Skill Score (RPSS)] is computed from the climatological probabilities and is positively oriented. The reference climatological data were calculated by using NCEP–NCAR 40-Year Reanalysis data (Kalnay et al., 1996).

2.4.3 *Rank histograms*

Rank histograms of an ensemble prediction measure how well the spread of the ensemble forecast reflects the observed probability distribution (Anderson, 1996; Talagrand et al., 1997). If the forecast is perfectly reliable and has the correct spread, an observation is equally likely to be placed in any quantile of the distribution estimated by the ensemble prediction, and the histogram would be flat (Wilks, 2006).

The reliabilities of different ensemble sizes in terms of a rank histogram can be compared by evaluating the score defined by Candille and Talagrand (2005) to measure the degree of flatness of a rank histogram. If the number of verification samples is M, the deviation of the rank histogram from flatness is measured as

$$\Delta=\sum_{k=1}^{N+1}(s_k-\frac{M}{N+1})^2\ , \tag{5}$$

where s_k is the number of elements in the kth interval of the rank histogram, and $M/(N+1)$ represents the expectation of s_k when the verifications are expected to be equally placed at each interval in a reliable system. The ratio defined as

$$\delta=\frac{\Delta}{MN/(N+1)}\ , \tag{6}$$

is used to measure the reliability of two experiments, where the denominator, $MN/(N+1)$, is the expectation of Δ. The ratio δ would approach one in the case of a perfectly reliable system (Candille and Talagrand, 2005).

In the rank histogram, the number of outliers is also calculated, as in Arribas et al. (2005). The number of outliers is a useful diagnostic tool to assess the reliability of an EPS and is defined as the number of occasions that the verifying analysis lies outside the ensemble range. The number of outliers is calculated by counting the verifying analysis that is larger than the largest ensemble and smaller than the smallest ensemble in the histogram. If the EPS is perfectly reliable, the expected number of outliers is equal to $2/(N+1)$ for one verifying sample.

2.4.4 *Reliability diagrams*

A reliability diagram compares the predicted probabilities of dichotomous events with their observed frequencies (Wilks, 2006). The conditional distribution of observations given each allowable value of the forecast against the forecast probability is plotted. The reliability and resolution components of the BS can be measured graphically using reliability diagrams.

2.4.5 *Relative Operating Characteristic curves*

A ROC curve is based on signal detection theory (Stanski et al., 1989) and it measures the resolution of probabilistic forecasts; that is, the ability of the forecast to discriminate dichotomous events (Wilks, 2006).

3. Results

3.1 *The ensemble-mean skill and spread*

The ensemble-mean skill is an overall measure of ensemble performance. It is known that the ensemble-mean forecast is more accurate than that of individual ensemble members due to the filtering of unpredictable scales of motion in the ensemble mean (Leith, 1974). However, the benefit of the ensemble mean is limited when the flow regime of the forecast changes (Palmer, 1993), and the ensemble mean does not measure the degree of uncertainty associated with ensemble prediction directly. Forecast uncertainties can be measured by the ensemble spread, which represents the differences among the members in an ensemble forecast. Further, ensemble spread is considered to be a predictor of ensemble-mean skill, especially when the variation in the spread is extreme (Whitaker and Loughe, 1998).

Figure 2a shows the ensemble spread and the RMSE of the ensemble mean for the 500-hPa height variable in the NH region. The ensemble spreads for M24 and M48 were both smaller than their corresponding RMSEs in the NH region after short forecast times, which implies that the ensemble spread was not large enough to cover uncertainties. The average RMSE of the ensemble mean of M48 was lower than that of M24 for most forecast times, but not much different at the 90% confidence level, in terms of the T^2 test (Wilks, 2006). Conversely, the M48 ensemble produced a slightly lower spread than M24 until the 4-day forecast at the 90% confidence level. While the spread of M48 grew faster than M24 during the first four days, the growth rates of the spreads for M24 and M48 were almost the same after four days.

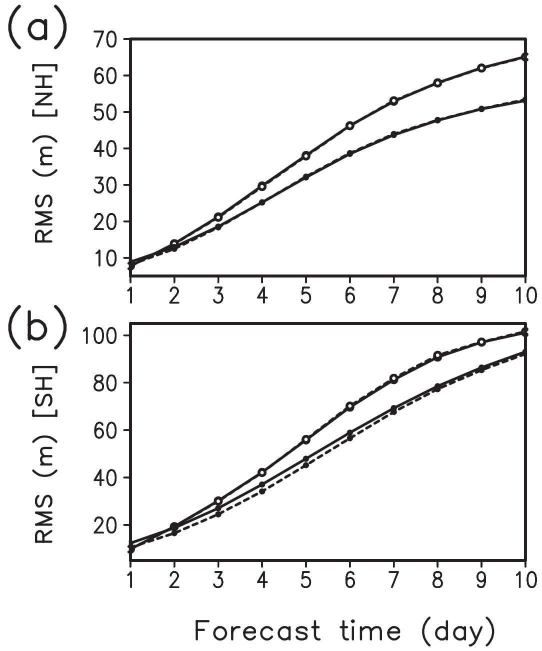

Fig. 2. The average spread (closed circles) and RMSE (open circles) of the ensemble mean forecast at a 500-hPa height for M24 (solid) and M48 (dotted) in the (a) NH region and (b) SH region.

The small spread difference between the two ensemble size experiments in this study contradicts the results reported in previous studies. In perfect or realistic forecast system experiments, an increase in ensemble size makes the average ensemble spread larger to be consistent with the average error of the ensemble mean because of better sampling and improved distribution (Buizza et al., 1998; Atger, 1999; Leutbecher and Palmer, 2008). The reason for the discordant results between our study and previous studies with regard to ensemble spread may be the inflation factors that were applied to the transform matrices in ETKF. For short forecast times (during the first 1.5 days), the magnitude of the ensemble spread was larger than the RMSE of the ensemble mean for M24, while the magnitude of the ensemble spread and RMSE of the ensemble mean were almost the same for M48. As a result, the spread skill score, which is the ratio of the error of the ensemble forecast to the ensemble spread, was 0.874 for M24 and 0.965 for M48 for the one-day forecast. Because the inflation factor reflects the relationship between the ensemble spread and ensemble mean forecast according to Eq. (2), the inflation factor of M48 was expected to be greater than that of M24. However, the inflation factor of M48 calculated from Eq. (2) was very large compared to the magnitude of the transform matrix, so it was substituted by the inverse of the magnitude of the transform matrix, as described in section 2.2.2 and the Appendix, to prevent the amplitude of the initial perturbations of M48 from being too large. Consequently, the average inflation factor elements of M24 was about 1.6 times larger than that of M48 in this study, which resulted in the ensemble spread of M48 being smaller than that of M24 for short forecast times (i.e. 1–3 days).

The ensemble spread and RMSE of the ensemble mean for the 500-hPa height variable in the SH region are shown in Fig. 2b. The RMSEs of the ensemble means of M24 and M48 were not much different at the 90% confidence level. As Atger (1999) discussed, this result indicates that an increase in ensemble size has little effect on the ensemble mean. The ensemble spread of M48 was lower than that of M24 during the first six days at the 90% confidence level. Similar to the results in the NH region, M48 had larger inflation elements on average than M24 according to Eq. (2), but the inflation factor of M48 applied to the transform matrix was smaller than that of M24 due to the inner algorithm of MOGREPS, which limits the effect of increasing the number of ensemble members on the spread. While the ensemble spreads of the two experiments showed similar growth rates for the first five days, the ensemble spread of M48 grew faster than that of M24 after five days.

3.2 *BS and RPSS*

Figure 3a shows the BSS in the NH region. For the first four days, M24 was slightly better than M48, but after four days M48 showed better performance than M24. The resolution component of the BS, which is typically one order of magnitude larger than the reliability component of the BS, shows that M48 had slightly better performance than M24 during the entire forecast period (Fig. 3b). However, the differences of BSS and the resolution component of BS between M24 and M48 were very small in terms of the 90% confidence level. In contrast, the reliability component of the BS shows that M24 performed better than M48 for most of the forecast period at the 90% confidence level (Fig. 3c).

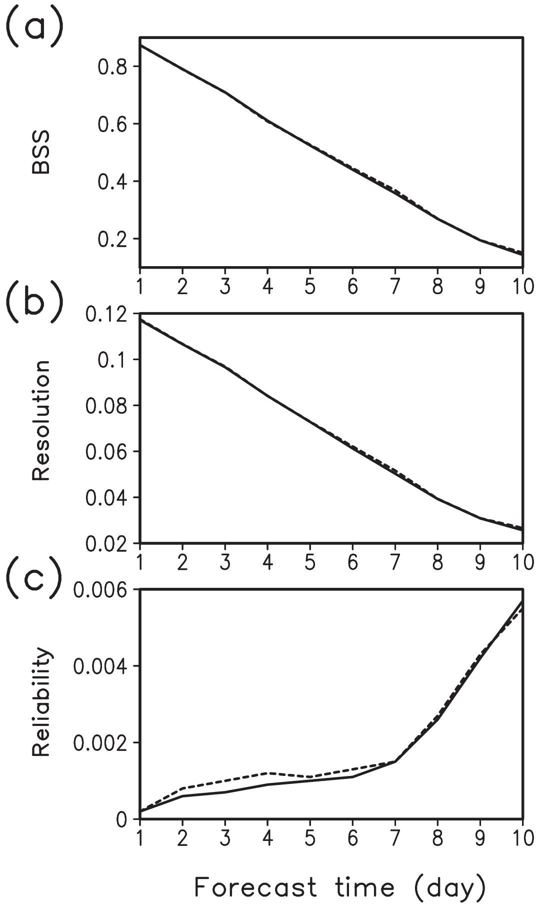

Fig. 3. (a) BSS, (b) resolution, and (c) reliability components of BSS for M24 (solid) and M48 (dotted) in the NH region.

In the SH region, the BSS of M48 was slightly better than that of M24 for the first four days (Fig. 4a). Even though M24 was better than M48 for the first five to six forecast days, M48 showed better performance with relatively larger differences as the forecast time increased to more than six days. M48 had a better resolution component of the BS than M24 for the entire forecast period (Fig. 4b). However, the BSS and resolution component of BS were not much different at the 90% confidence level, similar to the results in the NH region shown in Fig. 3. In contrast, the reliability component of the BS in the SH region was worse for M48 than for M24 at the 90% confidence level (Fig. 4c).

Overall, increasing the ensemble size had different effects on the reliability and resolution components of the BS. As Atger (1999) showed, ensemble spread is directly related to the reliability of the BS. In our study, it was difficult to identify the effect of increasing ensemble size on reliability because the inflation factor

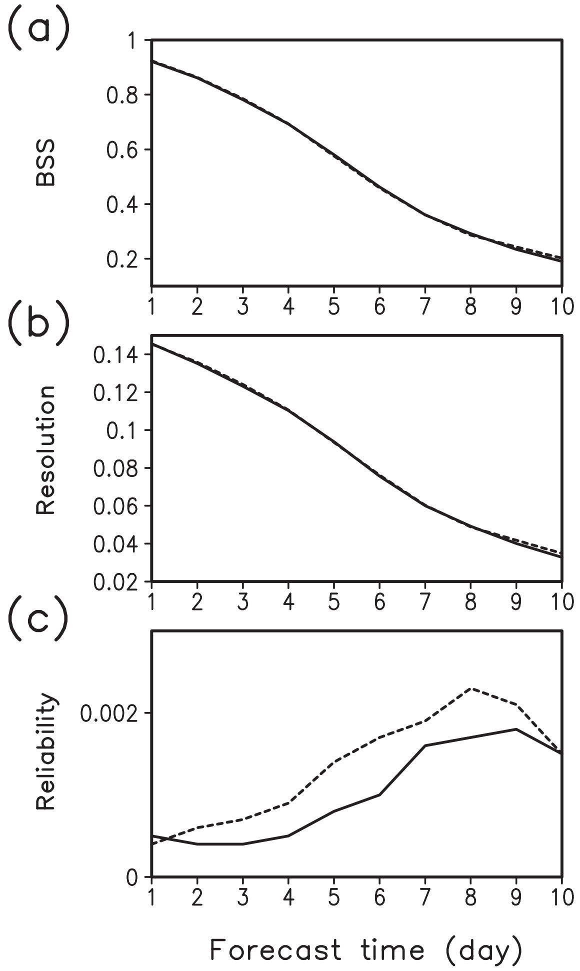

Fig. 4. The same as Fig. 3, but for the SH region.

calculation algorithm in MOGREPS forced the ensemble spread of M48 to decrease. The results of the present study indicate that increasing the ensemble size from 24 to 48 may not improve reliability if the spread is not maintained properly. The resolution component of the BS improved slightly when the ensemble size was increased from 24 to 48, even though it was not significant, because the resolution of the BS is determined not by the average amplitude of the ensemble spread, but by the daily variation in ensemble spread, which was improved by increasing the ensemble size, as mentioned in Atger (1999). There is another factor to be considered in assessing the reliability and resolution of the BS in terms of ensemble size. As already discussed by Candille and Talagrand (2004), the results in this subsection show that the effect of ensemble size on the reliability of the BS is affected by the number of verification samples; that is, increasing the ensemble size without increasing the sample number improves resolution, but decreases reliability.

Figure 5 shows the RPSS in the NH and SH region.

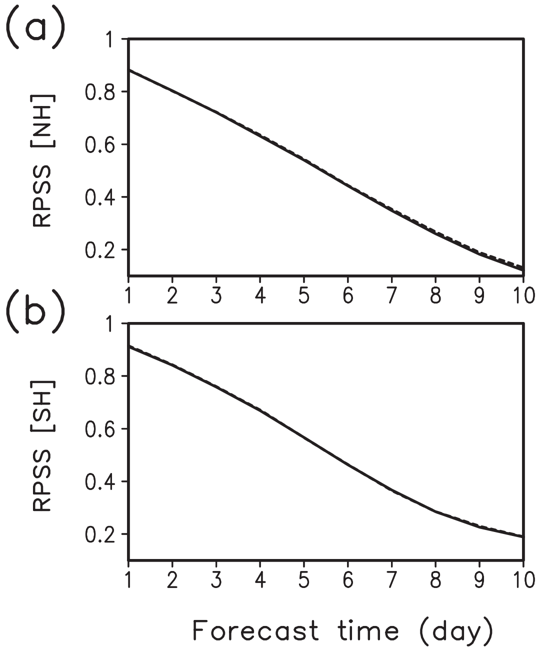

Fig. 5. RPSS for M24 (solid) and M48 (dotted) in the (a) NH region and (b) SH region.

In both the NH and SH regions, M48 and M24 were not very different at the 90% confidence level.

3.3 *Rank histograms*

Figure 6 shows the rank histograms for M24 and M48 for days 1, 5, and 10 in the NH region. For the entire forecast period, M24 and M48 showed similar rank histogram patterns in the NH region. At day 1, there were overforecasting biases; that is, the verifications were put too frequently on the side of the smallest ensemble forecasts. These unconditional biases for short forecast times decrease statistical consistency and reliability between ensemble members and verification. M48 showed a relatively more flat histogram than M24. As forecast times lengthened, the rank histograms of both ensemble sizes became flatter and adopted a U-shape, which indicates underdispersion of the ensemble forecast. Each ensemble was similar, so verifications tended to reside at both ends of the ensemble members, which caused the number of outliers over the largest ensemble forecast, as well as the smallest ensemble forecast, to increase in both experiments. Figure 7 shows the time evolution of the number of outliers in the NH region. The outliers of M24 and M48 grew very fast during the first two days. This is because the ensemble spread, which was artificially amplified by the inflation factor at the initial time, did not increase as fast as the error of the ensemble mean forecast. After two days, M48 had less outliers than M24, which reflects that the directions

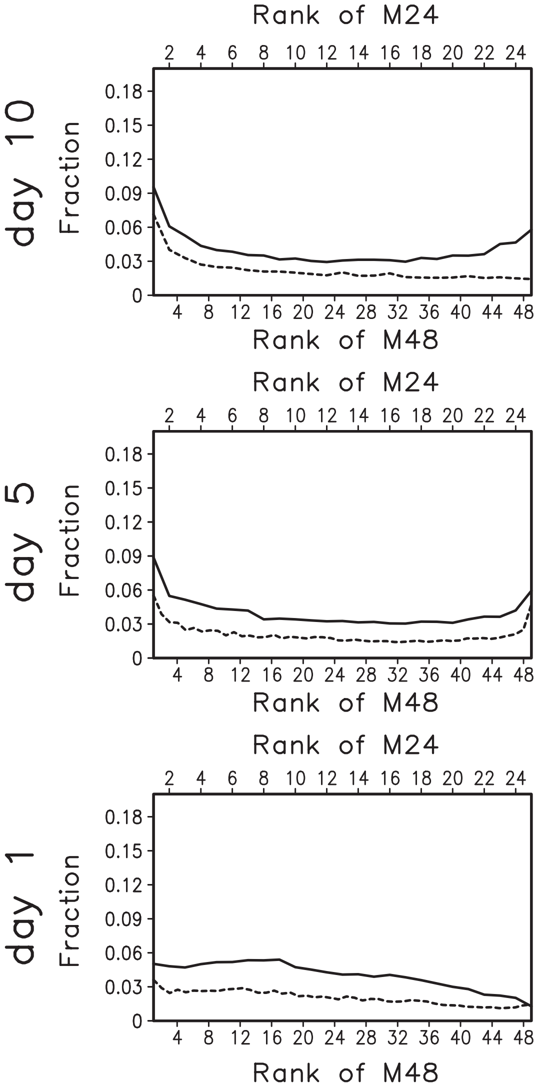

Fig. 6. Rank histograms of M24 (solid) and M48 (dotted) for 1-, 5-, and 10-day forecasts in the NH region. The abscissa at the top is for M24, and the abscissa at the bottom is for M48.

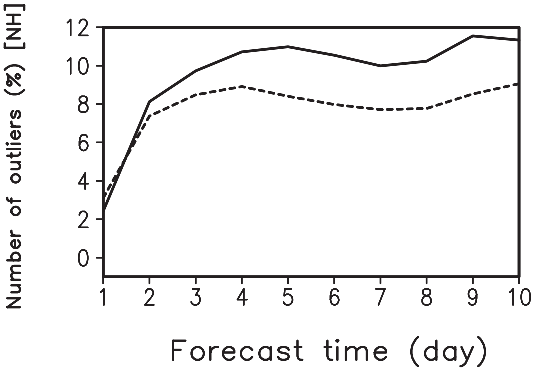

Fig. 7. Average number of outliers (%) of M24 (solid) and M48 (dotted) in the NH region.

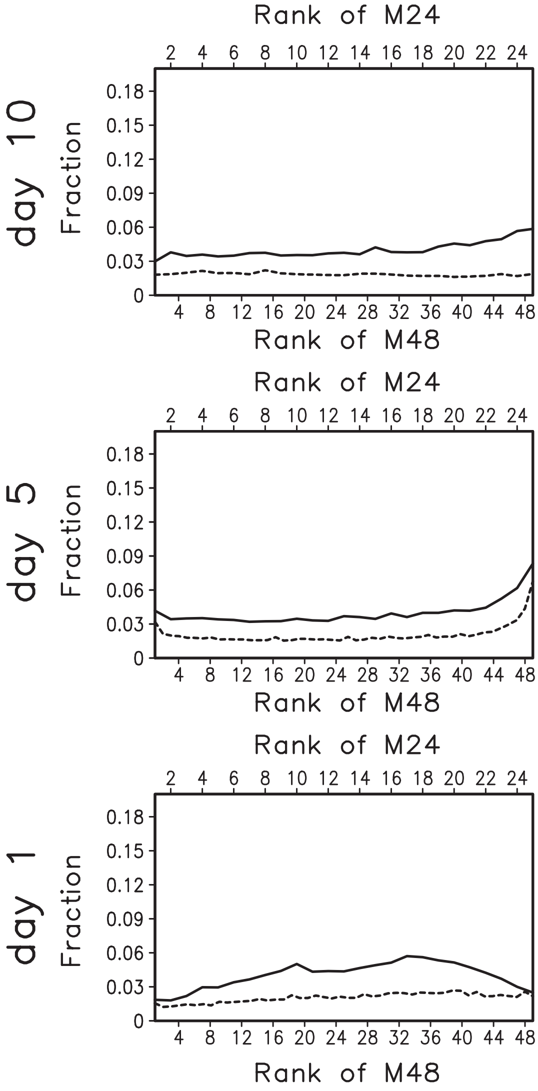

Fig. 8. The same as Fig. 6, but for the SH region.

spanning the subspace of the initial condition uncertainty were augmented by increasing the ensemble size in this system, as described by Buizza et al. (1998).

In contrast, the SH region showed an underforecasting bias rather than the overforecasting bias seen in the NH region for M48 for a short forecast period (day 1) (Fig. 8). The rank histogram of M24 had a dome shape at day 1, indicating overdispersion of the ensemble forecast. Because the ensemble spread of M24 modified by the inflation factor was too large (Fig. 2b), the verification was likely located not at extreme members, but around the center of ensemble members. As a result, M48 showed a relatively flatter histogram than M24. As the forecast time increased, the rank histograms of both experiments became flatter, and showed underforecasting biases. Figure 9 shows the time evolution of the number of outliers in the SH region. Up until day 3, the growth rates of out-

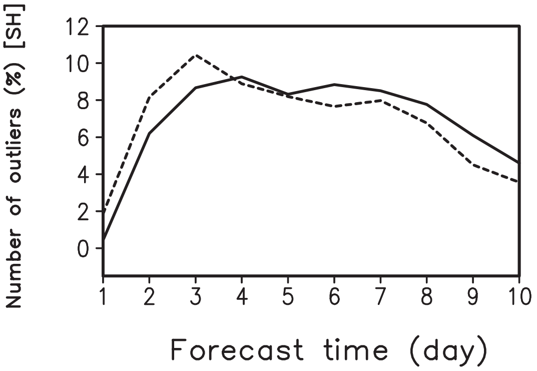

Fig. 9. The same as Fig. 7, but for the SH region.

liers for both ensemble size experiments were very large, and M48 had more outliers than M24, because the ensemble spread of M48 was more underdispersed than that of M24 during these forecast times. After day 4, the outliers of M24 and M48 began to decrease, and M48 had fewer outliers than M24. As forecast time increased, the number of outliers in M48 decreased faster than that in M24.

Figure 10a shows the time evolution of the ratio δ in Eq. (6) derived from the full rank histogram in the NH region. M48 had better reliability than M24 for the entire forecast period, even though the scores that measure how far the systems are away from reliability were greater than one. On day 1, the ratio of M48 was the smallest and significantly different from that of M24. For up to four days, M48 showed faster growth than M24. After four days, the ratios of both experiments showed similar growth rates. To assess the reliability without the outliers that were significant in the rank histograms as the forecast time increased, the time evolution of the ratio δ without the outliers was plotted and is shown in Fig. 10b. The ratio of M48 showed better reliability than that of M24 for most of the forecast times. In particular, M24 had a large value for the one-day forecast, which was consistent with the larger overforecasting bias of M24 than M48. While the ratio of M48 was almost constant for the entire forecast period, the ratio of M24 tended to increase after three days. The smaller outliers of M48 than M24 in Fig. 7 suggest that an increase in ensemble size reduces the degree of underdispersion of the ensemble forecast and improves reliability in the NH region.

The ratio δ in the SH region is shown in Fig. 11a. On day 1, M48 showed much better reliability than M24. However, M24 was better than M48 for days 2–7 because M48 rapidly increased after the first day. After seven days, M48 mostly showed better reliability than M24. Figure 11b shows the ratio δ calculated without outliers. A comparison of the ratio in the NH region (Fig. 10b) reveals a decreasing tendency to approach one in both M24 and M48. M48 showed better reliability than M24 for most forecast times. For the first three days, the ratio δ of M48 was significantly

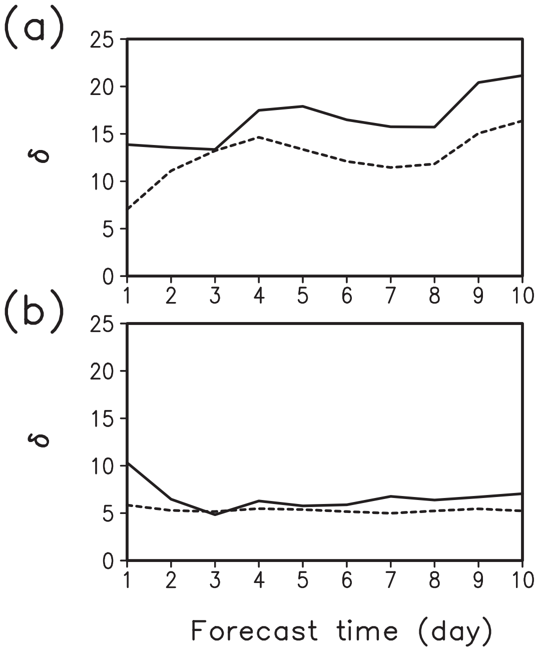

Fig. 10. The ratio δ derived from (a) the rank histogram of full elements and (b) the rank histogram excluding elements in the two extreme ranks of the NH region (M24, solid; M48, dotted).

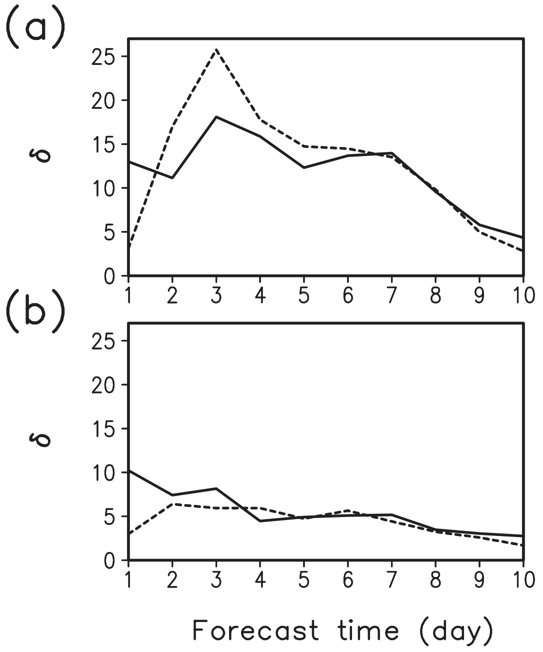

Fig. 11. The same as Fig. 10, but for the SH region.

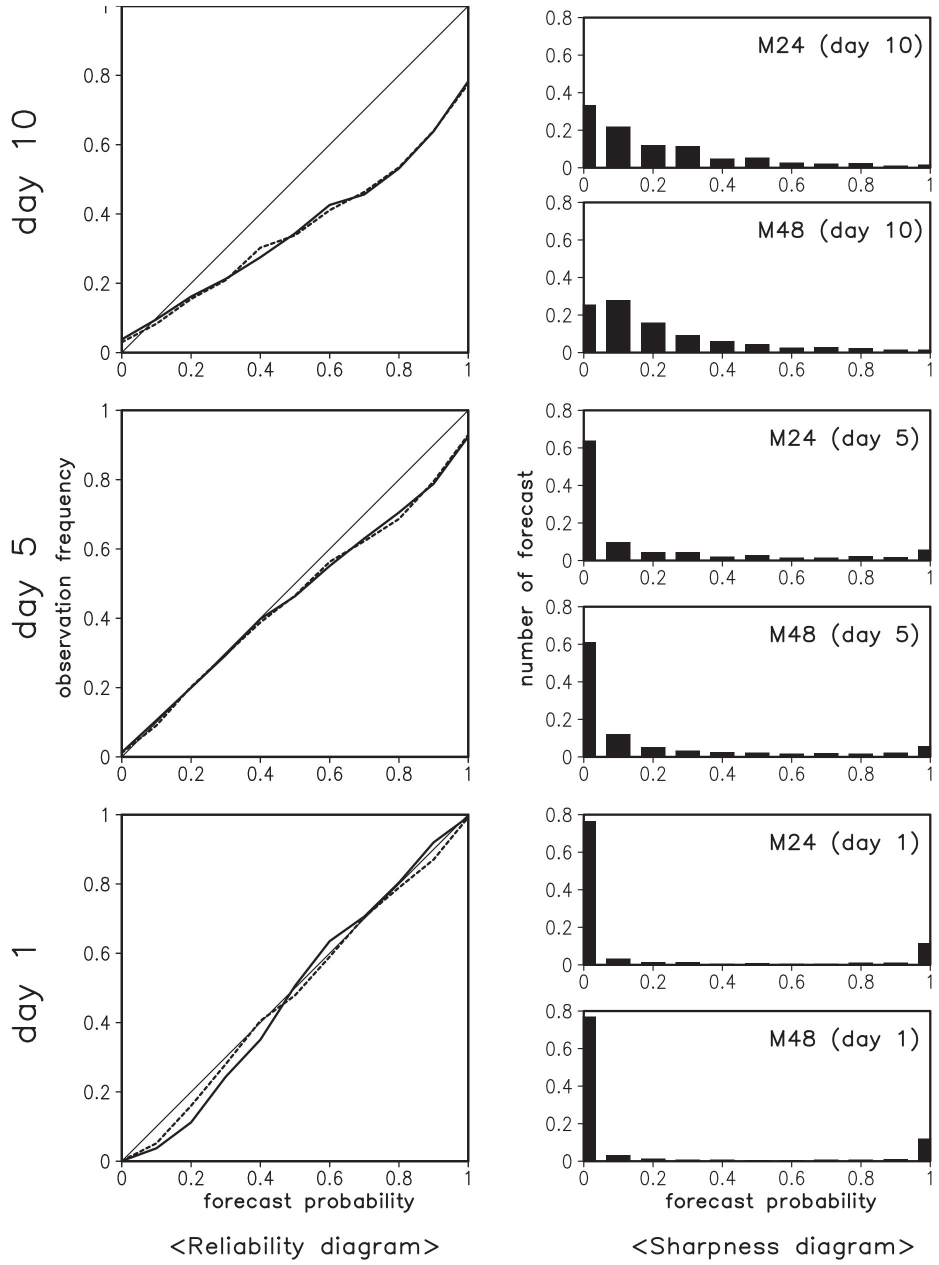

Fig. 12. Reliability diagram (left panel) and sharpness diagram (right panel) for 1-, 5-, and 10-day forecasts in the NH region. The bold solid line represents M24, the bold dotted line M48, and the thin solid line indicates the 1:1 line of the reliability diagram.

smaller than that of M24; after this period, the ratio of M48 and M24 became similar.

3.4 *Reliability diagrams*

The reliability diagram for the 500-hPa height in the NH region is shown in Fig. 12. On day 1, the reliability diagram indicates that both M24 and M48 showed good reliability and resolution (lower panel of Fig. 12). The reliability curve of M48 was closer to the 45° line than that of M24, which indicates that M48 was more reliable than M24. However, the resolution of M24 measured by calculating the mean square difference of the reliability curve to the sample climatology in relation to the resolution component of BS was better than that of M48 (0.42137 for M24 and 0.40993 for M48). The sharpness diagram shows that the extreme probability was too extreme in both M24 and M48.

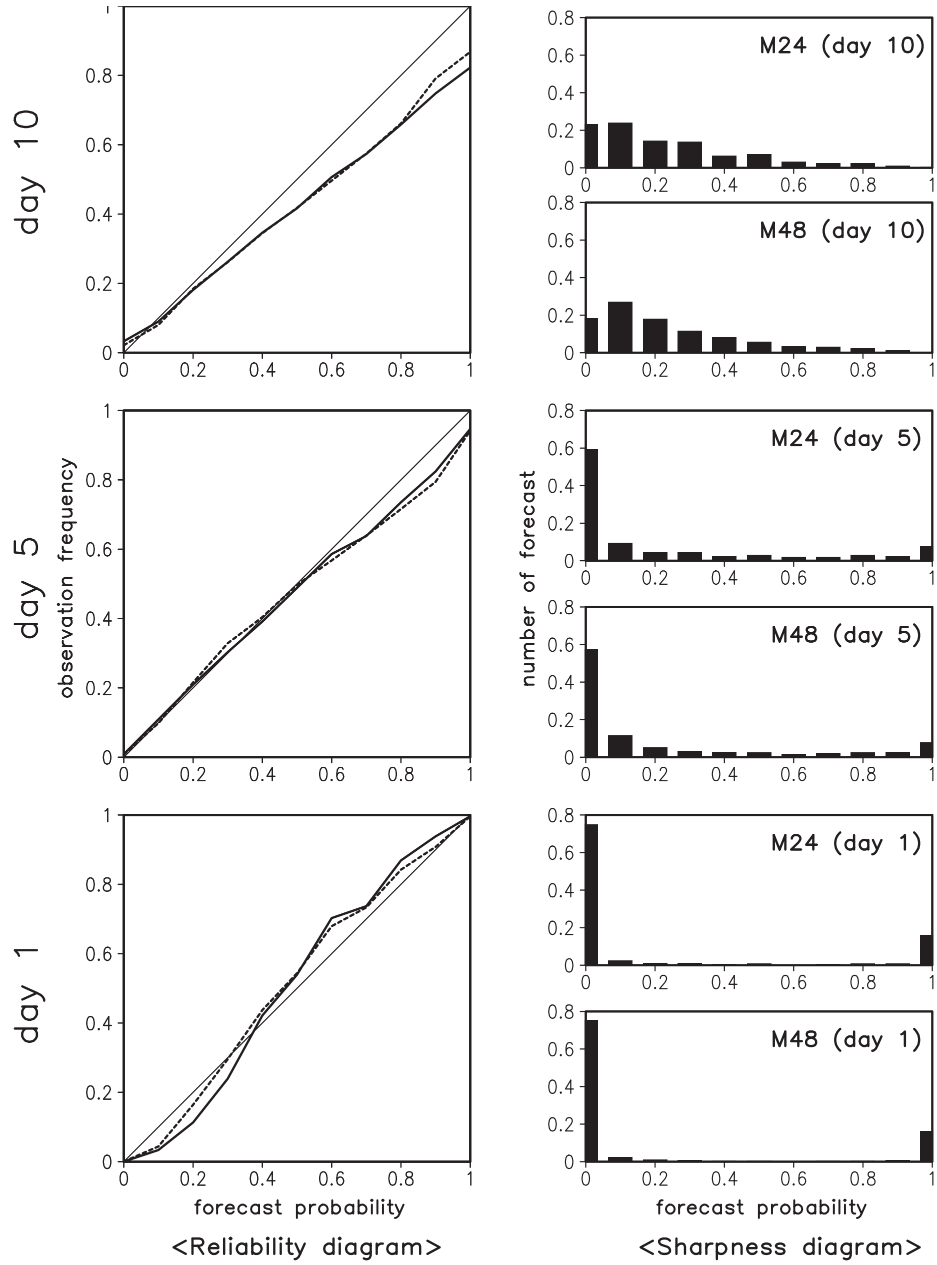

Fig. 13. The same as Fig. 12, but for the SH region.

At day 5, the reliabilities of rare events were good, but the unconditional biases of overforecasting over higher forecast probabilities resulted in deterioration of the reliability of both M24 and M48 (middle panel of Fig. 12). The reliability of M24 was better than that of M48, but the difference was negligible. However, M48 showed better resolution than M24 (3.6883 for M24 and 3.7062 for M48).

The upper panel of Fig. 12 shows the reliability diagram for day 10. The reliability curves of both experiments have slopes lower than the 45° line for all forecast probabilities, and the reliability as well as resolution of M24 and M48 are much degraded. On average, M48 shows better reliability and resolution than M24, but the two curves are almost the same. The sharpness diagram for day 10 shows that probabilities in the middle of the range are used more frequently than those at short forecast times. As the forecast time increases, more constant forecasts are performed with a loss of sharpness.

The properties of the reliability and sharpness diagrams of both M24 and M48 in the SH region are similar to those for the NH region. On day 1, both M24 and M48 show relatively good reliability and resolution, with M48 showing better reliability than M24 (lower panel of Fig. 13). As the forecast time increases, unconditional overforecasting biases occur, which decrease the reliability and resolution ability, with few differences in the diagrams of M24 and M48 (middle and upper panels of Fig. 13). M24 had better resolution than M48 for days 5 and 10, but the resolution of M48 was better than that of M24 for all other forecast times (not shown).

3.5 *ROC area*

Figure 14 shows the area under the ROC curve, which measures the degree of resolution in probabilistic forecasts. In the NH region, M24 had better resolution than M48 until day 4; after four days, M48 showed better resolution than M24 (Fig. 14a). The ROC area in the SH region showed that both ensemble sizes had very similar resolution abilities until day 7. After seven days, M48 showed better resolution than M24, but the difference between M48 and M24 was not significant. In contrast to the numerical difference of ROC area between M24 and M48, they were not much different at the 90% level in the both the NH and SH region during all forecast times.

The results from the ROC areas were similar to those observed for the resolution component of the BS. While the calculation of ROC is based on signal detection theory using the conditional distribution of the forecast probability given the observations by considering all probability thresholds over zero to one, the calculation of the BS assesses the relative accuracy of probabilistic forecasts using the conditional distribution of observations given for each forecast (Stanski et al., 1989). Although these two verification methods have different underlying assumptions, the similar verification results obtained in this study indicate the consistency of these methods for assessing the resolution ability of probabilistic forecasts.

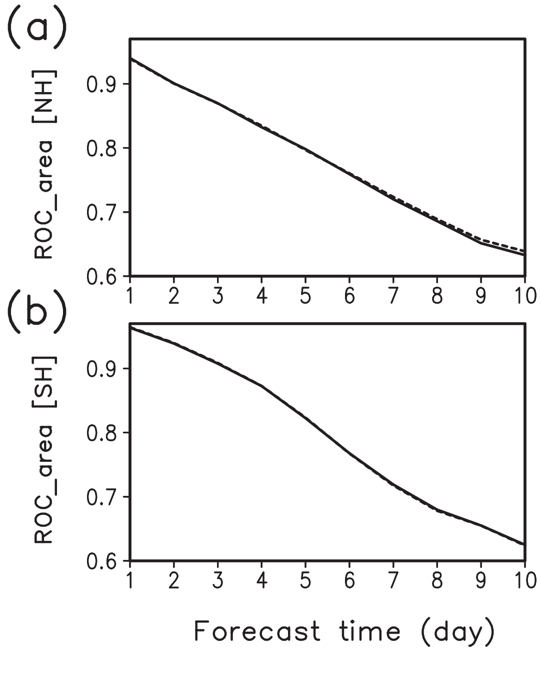

Fig. 14. Average ROC area of M24 (solid) and M48 (dotted) for the (a) NH and (b) SH regions.

4. Summary and discussion

The main purpose of ensemble prediction is to estimate the probability distribution of forecast uncertainties by conducting multiple predictions from slightly different initial conditions. The effect of doubling ensemble size on the performance of ensemble prediction was investigated in terms of probabilistic forecasts using MOGREPS implemented at the KMA for one month from 22 May to 23 June 2009. The initial ensemble perturbations generated by the ETKF technique were increased from 24 to 48 members, and then integrated for 10 days.

The performance of probabilistic forecasts is assessed by considering the statistical reliability and resolution of the ensemble forecasts as well as the accuracy of the forecasts. Because different verification scores are based on different characteristics of probabilistic forecasts, various verification scores are interpreted together. Different scores sometimes show inconsistent results for probabilistic forecasts due to the different perspectives and arithmetical approaches underlying the various verification methods.

The accuracy of ensemble prediction was measured by the RMSE of the ensemble mean forecast, the BSS, and the RPSS. The results showed that the accuracy improved slightly, but was not significant at the 90% confidence level, when the ensemble size was increased from 24 to 48, especially for longer length forecasts in the NH region. In contrast, in the SH region, the impact of ensemble size on accuracy did not show consistent results and the differences in scores between the two ensemble size configurations was negligible. These results were similar to those reported by Buizza et al. (1998), which indicated that increasing the ensemble size from 30 to 51 had little impact on accuracy in the ECMWF EPS.

The reliability of ensemble prediction was assessed using the reliability component of the BS, rank histograms with outliers, and reliability diagrams. The reliability component of the BS showed that increasing ensemble size had a negative effect on reliability, because the ensemble spread of the forecasts with larger ensemble sizes was decreased by the inflation factor in MOGREPS. However, the rank histogram results showed that increasing ensemble size improved reliability as forecast time lengthened due to the better relationship between the ensemble spread and the RMSE of the ensemble mean forecast with increasing ensemble size. The reliability diagrams also showed that reliability was improved by increasing the ensemble size for both short and long forecast times. Although each verification method yields different results, an increase in ensemble size is expected to have a positive effect on reliability considering not only the increase in consistency between ensemble spread and error ensemble forecast, but also the critical condition of sample size, which is related to model resolution and the verification period. The ROC area and resolution component of the BS were used to measure the resolution of MOGPRES. Even though the results were not significant, these scores indicated the relative advantage of increasing ensemble size, especially for longer forecast times.

Even though the inflation factor calculation algorithm in MOGREPS hindered the growth of spread of M48 and caused contradictory results in the spreads of M48 and M24 compared to previous studies, increasing ensemble size had slight advantages overall, especially for longer forecast times. However, the advantage of doubling ensemble size may be limited by the inflation factor calculation algorithm in MOGREPS. In future work, the inflation factor calculation algorithm in MOGREPS will be modified to incorporate the issues revealed by this study while maintaining its operational efficiency.

Acknowledgements. This study was supported by the Korea Meteorological Administration Research and Development Program under Grant CATER 2012–2030. The authors would like to thank the Numerical Weather Prediction Division of the Korea Meteorological Administration and UK Met Office for providing computer facility support and resources for this study.

APPENDIX

The Three Steps to Calculate the Inflation Factor in MOGREPS

The inflation factor calculation is originally based on the innovation statistic described in Eq. (2). However, additional post processes are necessary to realize the basic idea of calculating the inflation factor in MOGREPS. The three steps to calculate the inflation factor referred to in section 2.2.2 are discussed with mathematical expressions as follows.

Step 1

According to the Eq. (2), $\boldsymbol{\Pi}_{n,i}$ at each localization region is calculated using sonde and ATOVS observations.

$$\boldsymbol{\Pi}_{n,i} = \boldsymbol{\Pi}_{n-1,i} \circ \boldsymbol{I}_{n-1,i} \, , \tag{A1}$$

where n represents the time step, i denotes the index for each observation category, $\circ$ denotes multiplication between vector elements which results in vector, and $\boldsymbol{I}$ denotes the inflation ratio as,

$$\boldsymbol{I}_{n,i} = \sqrt{\frac{\mathrm{trace}(\boldsymbol{d}_{n,i}\boldsymbol{d}_{n,i}^{\mathrm{T}}) - \mathrm{trace}(\boldsymbol{R})}{\mathrm{trace}(\boldsymbol{H}\boldsymbol{P}_{\mathrm{f},n,i}\boldsymbol{H}^{\mathrm{T}})}} \, . \tag{A2}$$

Step 2

To obtain the inflation factor that is finally used to scale the transformed perturbations, $\boldsymbol{\Pi}_{n,i}$ for two observation categories should be combined by considering the minimum discrepancy between the resulting inflation factors and $\boldsymbol{\Pi}_{n,i}$. If $\boldsymbol{\Pi}_n$ is the combined inflation factor, it will minimize the

$$\sum_{i=1}^{2} (\boldsymbol{\Pi}_n/\boldsymbol{\Pi}_{n,i} - 1)^2 \, . \tag{A3}$$

To obtain optimal $\boldsymbol{\Pi}_n$, we differentiate Eq. (A3) with respect to $\boldsymbol{\Pi}_n$, and get the following equation

$$\sum_{i=1}^{2} (2\boldsymbol{\Pi}_n/\boldsymbol{\Pi}_{n,i}^2 - 2/\boldsymbol{\Pi}_{n,i}) = 0 \, . \tag{A4}$$

From Eq. (A4), we can get the solution, $\boldsymbol{\Pi}_n$ (hereafter called the "combined inflation factor")

$$\boldsymbol{\Pi}_n = \sum_{i=1}^{2} (1/\boldsymbol{\Pi}_{n,i}) / \sum_{i=1}^{2} (1/\boldsymbol{\Pi}_{n,i}^2) \, . \tag{A5}$$

Step 3

Sometimes, the elements of the inflation factor have too large magnitudes and result in excessively large initial ensemble perturbations. The last procedure is the safety check to obtain the new inflation factors without excessive inflation. First, we derive "RawScale" $\boldsymbol{S}$ that measures how the magnitude of forecast perturbations is rescaled into the magnitude of initial perturbations by the transform matrix in ETKF as

$$\boldsymbol{S}_n = \sqrt{\frac{\sum\limits_{p,q}^{N} \boldsymbol{T}_n(p,q)^2}{N} - 1} \, , \tag{A6}$$

where p and q are the indexes for the elements of transform matrix.

Second, "ScalingFactor" $\boldsymbol{F}$ is obtaind by

$$\boldsymbol{F}_n = \boldsymbol{S}_n \circ \boldsymbol{\Pi}_n \tag{A7}$$

which measures the relative magnitudes of $\boldsymbol{S}_n$ and "combined inflation", $\boldsymbol{\Pi}_n$, obtained in Step 2. As explained in Section 2.2.2, "ScalingFactor" is expected to be less than one because the uncertainty of initial perturbations are less than that of forecast perturbations by optimally weighted observational error information through ETKF. Therefore, the final inflation factor $\boldsymbol{\Pi}_{f,n}$ is determined by

$$\boldsymbol{\Pi}_{\mathrm{f},n} = \boldsymbol{\Pi}_n \circ \min(1.0, 1.2/\boldsymbol{F}_n)\,, \tag{A8}$$

where min(a, b) is a function that chooses the smaller one between a and b. The constant, 1.2, is the predefined limit for $\boldsymbol{F}_n$. Equation (A8) implies that if the magnitude of inflation factor elements is larger than the magnitude of the transform matrix that decreases the magnitude of the forecast perturbations, the inverse of the magnitude of the transform matrix (i.e. RawScale) is substituted for the inflation factor. Otherwise, the inflation factor calculated in Step 2 is determined as the final inflation factor.

REFERENCES

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*, **9,** 1518–1530.

Anderson, J. L., 2001: An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev.*, **129**, 2884–2903.

Arribas, A., K. B. Robertson, and K. R. Mylne, 2005: Test of a poor man's ensemble prediction system for short-range probability forecasting. *Mon. Wea. Rev.*, **133**, 1825–1839.

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

Atger, F., 2004: Estimation of the reliability of ensemble based probabilistic forecasts. *Quart. J. Roy. Meteor. Soc.*, **130**, 627–646.

Bowler, N. E., A. Arribas, K. R. Mylne., K. B. Robertson, and S. E. Beare, 2008: The MOGREPS short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **134**, 703–722.

Bowler, N. E., A. Arribas, S. E. Beare, K. R. Mylne, and G. J. Shutts, 2009: The local ETKF and SKEB: Upgrades to the MOGREPS short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **135**, 767–776.

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

Buizza, R. T. Petroliagis, T. 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.

Candille, G., and O. Talagrand, 2004: On limitations to the objective evaluation of ensemble prediction systems. *Workshop on Ensemble Methods*, UK Met Offect, Exeter, October 2004.

Candille, G., and O. Talagrand, 2005: Evaluation of probabilistic prediction systems for a scalar variable. *Quart. J. Roy. Meteor. Soc.*, **131**, 2131–2150.

Casella, G., and R. L. Berger, 1990: *Statistical Inference.* Duxbury Press, 650pp.

Ehrendorfer, M., 1994a: The Liouville equation and its potential usefulness for the prediction of forecast skill. I: Theory. *Mon. Wea. Rev.*, **122**, 703–713.

Ehrendorfer, M., 1994b: The Liouville equation and its potential usefulness for the prediction of forecast skill. I: Applications. *Mon. Wea. Rev.*, **122**, 714–728.

Ehrendorfer, M., 2007: A review of issues in ensemble-based Kalman filtering. *Meteor. Z.*, **16**, 795–818.

Epstein, E. S., 1969: A scoring system for probability forecast of ranked categories. *J. Appl. Meteor.*, **8**, 985–987.

Hacker, J. P., and Coauthors, 2011: The U.S. air force weather agency's mesoscale ensemble: Scientific description and performance results. *Tellus A*, **63**, 625–641.

Hamill, T. M., S. L. Mullen, C. Snyder, Z. Toth, and D. P. Baumhefner, 2000: Ensemble forecasting in the short to medium range: Report from a workshop. *Bull. Amer. Meteor. Soc.*, **81**, 2653–2664.

Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner'S Guide in Atmospheric Science.* John Wiley and Sons, Chichester, 292pp.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteor. Soc.*, **77**, 437–471.

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

Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. *J. Comput. Phys.*, **227**, 3515–3539.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141.

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.

Murphy, A. H., 1971: A note on the ranked probability score. *J. Appl. Meteor.*, **10**, 155–156.

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

Palmer, T. N., 1993: Extended-range atmospheric prediction and the Lorenz model. *Bull. Amer. Meteor. Soc.*, **74**, 49–65.

Park, Y. Y., R. Buizza, and M. Leutbecher, 2008: TIGGE: Preliminary results on comparing and combining ensembles. *Quart. J. R. Meteorol. Soc.*, doi: 10.1002/gi.

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.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in meteorology. Research Report 89-5, Environment Canada, 114pp.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction system. *Proc. ECMWF Workshop on Predictability*, ECMWF, 1–25.

Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker, 2003: Ensemble square root filters. *Mon. Wea. Rev.*, **131**, 1485–1490.

Wang, X., and C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.*, **60**, 1140–1157.

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.

Whitaker, J. S., and A. F. Loughe, 1998: The relationship between ensemble spread and ensemble mean skill. *Mon. Wea. Rev.*, **126**, 3292–3302.

Wilks, D. S., 2002: Smoothing ensembles with fitted probability distribution. *Quart. J. Roy. Meteor. Soc.*, **128**, 2821–2836.

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