Evaluation of the NMC Regional Ensemble Prediction System During the Beijing 2008 Olympic Games*

LI Xiaoli† (李晓莉), TIAN Hua (田 华), and DENG Guo (邓 国)

The Center of Numerical Weather Prediction, China Meteorological Administration, Beijing 100081

(Received August 16, 2010; in final form May 29, 2011)

ABSTRACT

Based on the B08RDP (Beijing 2008 Olympic Games Mesoscale Ensemble Prediction Research and Development Project) that was launched by the World Weather Research Programme (WWRP) in 2004, a regional ensemble prediction system (REPS) at a 15-km horizontal resolution was developed at the National Meteorological Center (NMC) of the China Meteorological Administration (CMA). Supplementing to the forecasters' subjective affirmation on the promising performance of the REPS during the 2008 Beijing Olympic Games (BOG), this paper focuses on the objective verification of the REPS for precipitation forecasts during the BOG period. By use of a set of advanced probabilistic verification scores, the value of the REPS compared to the quasi-operational global ensemble prediction system (GEPS) is assessed for a 36-day period (21 July–24 August 2008). The evaluation here involves different aspects of the REPS and GEPS, including their general forecast skills, specific attributes (reliability and resolution), and related economic values. The results indicate that the REPS generally performs significantly better for the short-range precipitation forecasts than the GEPS, and for light to heavy rainfall events, the REPS provides more skillful forecasts for accumulated 6- and 24-h precipitation. By further identifying the performance of the REPS through the attribute-focused measures, it is found that the advantages of the REPS over the GEPS come from better reliability (smaller biases and better dispersion) and increased resolution. Also, evaluation of a decision-making score reveals that a much larger group of users benefits from using the REPS forecasts than using the single model (the control run) forecasts, especially for the heavy rainfall events.

Key words: regional ensemble prediction, ensemble verification, probabilistic scores

Citation: Li Xiaoli, Tian Hua, and Deng Guo, 2011: Evaluation of the NMC regional ensemble prediction system during the Beijing 2008 Olympic Games. *Acta Meteor. Sinica*, **25**(5), 568–580, doi: 10.1007/s13351-011-0503-z.

1. Introduction

The errors existing in the initial conditions (ICs) of the numerical prediction models, along with the errors in forecast models due to the approximate nature of the parameterization schemes, cause the predictability problem of the numerical models at either short range or medium range. Based on the rationale of uncertainties caused by these errors, the ensemble prediction is an efficient method to take these sources of forecast errors into consideration by representing the atmospheric circulation through appropriate probability density function (Buizza et al., 2005). During the past decades, the global ensemble prediction systems (GEPSs) with different initial perturbation methods have been developed at several major meteorological centers, and have been successfully applied to operational weather forecasting, e.g., the European Center for Medium-range Weather Forecasts (ECMWF), the National Centers for Enviromental Prediction (NCEP), the Meteorological Service of Canada (MSC). The development of the GEPS at the National Meteorological Center (NMC) of the China Meteorological Administration (CMA) started in the late 1990s and has gone through the upgradation of model system and experimented with different initial perturbation methods (Li and Chen, 2002). Since 2007, the forecasts from 10 GEPSs including GEPS from CMA have been in the collection of the TIGGE (THORPEX Interactive Grand Global Ensemble) and

*Supported by the Scientific Fund for Chinese Returnees of the Ministry of Human Resources and Social Security of China and the Special Public Welfare Research Fund for Meteorological Profession of China Meteorological Administration (GYHY201006015).

†Corresponding author: lixl@cma.gov.cn.

have been providing products to the users worldwide for research and application purposes (Park et al., 2008).

Encouraged by the successful application of GEPSs, the Short-Range Ensemble Forecasting (SREF; Stensrud et al., 1999; Du and Tracton, 2001) at NCEP pioneered the regional ensemble forecasting technique in addressing the errors caused by the low predictability of mososcale short-range forecasts. The studies about developing REPSs have caused great attention during the past decades. The operationally-aimed REPSs were recently being tested and run by different operational forecast centers, e.g., the ALADIN-LAEF (Wang et al., 2011) at ZAMG (ZentralAnstalt für Meteorologie und Geodynamik), the COSMO-LEPS developed within the COSMO consortium by six European countries (Montani et al., 2003; Marsigli et al., 2005), the SRNWP-PEPS (Heizenreder el al., 2006) at German Met Service, ect. In order to enhance technical support to the Beijing 2008 Olympic Games, in 2004, the World Weather Research Programme (WWRP) sponsored the B08RDP (Beijing 2008 Olympic Games Mesoscale Ensemble Prediction Research and Development Project) for undertaking the research and development of the REPSs and applying the results to the quasi real-time forecasting support. The detailed implementation of the B08RDP can be found in Duan et al. (2011). Aimed to participate in the B08RDP and to promote the research in this field, an REPS was developed and tested at NMC (Deng et al., 2010). During the Beijing 2008 Olympic Games, the REPS and GEPS at NMC both played important roles in providing useful information. Based on the subjective evaluation from forecasters, more skillful forecasts for high impact weathers (HIWs) were made by the REPS. An objective evaluation of the REPs in comparison with its single model run (control run) and the existing GEPS is necessary and is of increasing interest with the rapid development and application of REPSs (Chessa et al., 2004; Bowler et al., 2008). Moreover, it is important to understand the specific attributes related to probabilistic forecasts, such as their reliability and resolution (Jolliffe and Stephenson, 2003), through comprehensive verification measures. In this work, the evaluation will be performed on precipitation forecasts. The accumulated 6- and 24-h precipitation forecasts will be evaluated by using a number of probabilistic scores concurrently.

This paper is organized as follows. The configurations of the REPS and GEPS at NMC are introduced in Section 2. The verification data are described in Section 3. Section 4 gives the comparison results, and summary and conclusions are presented in Section 5.

2. Configurations of the REPS and GEPS

A brief description of system configurations of the REPS and GEPS at NMC is given in this section. Note that the GEPS not only acts as a comparison reference, but also provides lateral boundary conditions (LBCs) to the REPS.

2.1 Configuration of the GEPS

The GEPS at NMC is based on 15 T213L31 (spectral triangular truncation T213 with 31 vertical levels) members, and performs 10-day forecasts twice daily at 0000 and 1200 UTC. The initial perturbation of the GEPS adopts the bred-vector (BV) approach proposed by Toth and Kalnay (1993, 1997). The BV approach is based on the notion that the errors of analysis fields generated by data assimilation schemes will grow and accumulate by the virtue of perturbation dynamics. For the GEPS, 7 breeding cycles, each initialized with different random perturbation fields, are used to produce 14 initial perturbations. These 14 initial perturbations are centered as positive-negative pairs around the T213 SSI (spectral statistical interpolation) analysis fields and are used to construct 14 perturbed ensemble members of the GEPS. A regional rescaling algorithm is applied in the breeding cycles to reflect the geographically varying uncertainties in the analyses. In the breeding cycles of the GEPS, the kinetic energy of the difference field between NCEP (as independent field) and T213 analysis fields at 500 hPa is chosen to measure the uncertainties of T213 analyses. Based on the historical data in 2003, every 5-day average of kinetic energy of the difference field is obtained, and its square root is regarded as the

estimator of the uncertainties of T213 analyses. The average kinetic energy difference field is then used as a geographic mask in the rescaling step of the breeding cycle, which ensures that the spatial distribution of initial perturbations produced by the BV cycle is similar to the analysis errors.

2.2 Configuration of the REPS

To participate in the B08RDP, an REPS with 15 members was developed with the Weather Research and Forecasting (WRF) modeling system (Janjié et al., 2001) at NMC. During the implementation of B08RDP, the horizontal resolution of the REPS was 15 km, with 35 levels in the vertical. Its integration was performed up to 36 h, covering northern China. In July 2010, the REPS was upgraded by expanding the integration domain to entire China and the forecast length up to 60 h.

The BV method is used to generate the initial perturbations for the REPS. The 7 breeding cycles, included in the WRF 3D-VAR assimilation cycles, are used to produce 7 paired initial perturbations. Based on these initial perturbations, 15 ICs of the REPS (7 paired perturbed ICs and a control WRF analysis) are generated, and LBCs are provided by 15 members of the GEPS. As mentioned earlier, being the function of geographic location, the rescaling factor plays an important role in the BV cycles for determining the magnitude of the perturbations, and two methods for describing the analysis uncertainty are tested when constructing WRF-based BV cycles, and similar results are obtained (Deng et al., 2010). One is similar to the method used in the GEPS that computes the kinetic energy difference between the WRF and NCEP analyses; the other is to use the temperature difference at 850 hPa between the above analyses.

The multi-physics technique (Stensrud et al., 2000; Du et al., 2003) is employed in the REPS to represent the model uncertainty. The practical application of the multi-physics method in the REPS focuses on the combinations of cumulus convective parameterizations, boundary layer schemes, and land surface schemes, wherein the selection of each physical parameterization is tested by carefully designed experiments (Deng et al., 2010).

3. Verification data

The evaluation methodology here is based on the station-to-station comparison between the interpolated model forecasts and observations. The gridded forecasts of the REPS and GEPS are interpolated onto the station points using bilinear interpolation technique.

The verification period is from 20 July to 24 August 2008, which includes 36-day forecasts initialized at 1200 UTC daily. The observations come from 400 conventional surface stations shown in Fig. 1, and the quality control with the operational standard of data collecting at NMC has been applied to these data. It should be noted that the verified GEPS forecasts are not on their original resolution due to the unavailability of the original data, but at 1.0 degree obtained from the TIGGE archive. This might affect slightly the fairness of comparison due to the additional interpolation process.

In this study, our goal is to evaluate the value of the REPS for precipitation forecasts within the forecast window of 36 h compared to the GEPS forecasts and the forecasts from the REPS control run. To investigate the forecast ability of the REPS for quantitative precipitation forecasts (QPFs), three kinds of evaluation are conducted. First, suitable verification measures will be used to evaluate the performance of each accumulated 6-h precipitation forecast during the

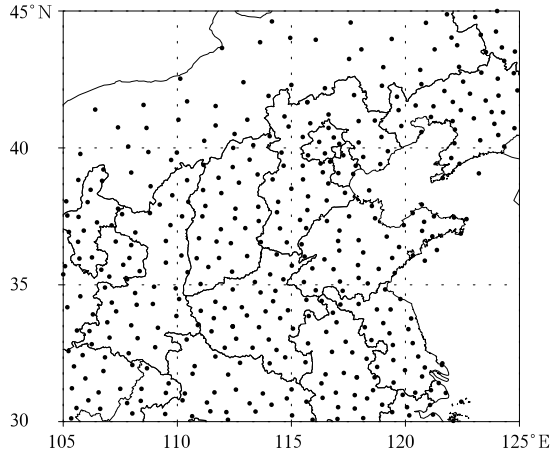

Fig. 1. Surface observation stations (dots) for precipitation forecast verification.

36-h forecast window that indudes a total of 6 verification periods (00–06, 06–12, 12–18, 18–24, 24–30, and 30–36 h; hereafter referred as to verified periods 1, 2, 3, 4, 5, and 6). Then, the specific thresholds for dichotomous predictands for accumulated 6-h precipitation are chosen to be evaluated by various probabilistic scores. For this type of evaluation, unlike the evaluation for each 6-h period listed before, the 6-h precipitation accumulation is examined instead. Finally, the 24-h accumulated precipitation is also measured by using the GEPS forecasts as reference.

4. Comparative verification of the REPS and GEPS

The verification methodology for ensemble probabilistic forecasts is quite different from the traditional verification technique for deterministic forecasts. The specific attributes related to probabilistic forecasts, "reliability" and "resolution" (Jolliffe and Stephenson, 2003), need to be evaluated. The "reliability" indicates the statistical agreement between the predicted probability of an event and the mean observed frequency of the event under consideration. The "resolution" is the ability of the forecasts to resolve the set of sample events into subsets with characteristically different frequencies. Most of the ensemble verification scores are specifically developed or designed to assess one or both of specific attributes with various focuses. To assess the performance of ensemble system comprehensively, various verification measures are needed concurrently.

For the evaluation of two systems, the estimation of statistical significance of performance difference is necessary. In this paper, the statistical bootstrap technique (Efron and Tibshirani, 1993) is adopted to estimate the uncertainty of the verification scores. When applying the bootstrap technique to the verification scores, we recompute the scores a number of times (N_b) with a sample randomly extracted from the sample pool of $n = N_\mathrm{s} \times N_\mathrm{d}$ realizations, with replacement from the original data set. Here, N_s is the number of verification stations, and N_d is the number of the days in the verification period. Following Candille et al. (2007), we resample over N_d days instead of over all n realizations, where each new sample of n realizations is obtained with all the N_s observations of each selected day. Then, the 90% confidence interval (CI) is obtained by upper and lower bounds (95% and 5%) with N_b times (200 times used in this paper).

4.1 The continuous ranked probability score

The continuous ranked probability score (CRPS; Stanski et al., 1989; Hersbach, 2000; Candille et al., 2007) measures the distance between the predicted and the observed cumulative density functions (CDFs) of a scalar variable as follows:

$$\mathrm{CRPS} = \int_{-\infty}^{\infty} \left[P_\mathrm{f}(x) - P_\mathrm{o}(x)\right]^2 \mathrm{d}x, \tag{1}$$

where P_f and P_o are the forecast and observed CDFs for the variable of interest, respectively, and

$$\begin{aligned} P_\mathrm{f}(x) &= \int_{-\infty}^{x} \rho(y)\mathrm{d}y, \\ P_\mathrm{o}(x) &= H(x - x_o), \end{aligned} \tag{2}$$

where $\rho(y)$ is the probability density function (PDF) of the forecast variable x from ensemble system, and x_o is the actually observed value. P_o follows the distribution of the Heaviside function. To apply CRPS to ensemble system with N numbers, at each verification location, the outcomes of N ensemble members are ranked from the lowest to the highest ($x_1, \ldots, x_N$, where $x_i < x_j$ if $i < j$), and $N + 1$ bins are obtained relative to the sorted ensemble outcomes. In the discrete cases, P_f can be expressed as a piecewise constant function with respect to these $N+1$ bins, for instance, in the bin i [x_i, x_{i+1}], P_f is written as $P_\mathrm{f}(x) = p_i = i/N$ for $x_i < x < x_{i+1}$. For P_o, its value is either 0 or 1, or partly 0, partly 1, depending on the location of observed x_o in the bin $[x_i, x_{i+1}]$. In this case, the discrete CRPS can be formulated as:

$$\begin{aligned} c_i &= \int_{x_i}^{x^{i+1}} [P_i - H(x - x_\mathrm{o})]\mathrm{d}x \\ &= \alpha_i p_i^2 + \beta_i (1 - p_i)^2, \\ \mathrm{CRPS} &= \sum_{i=0}^{N} c_i, \end{aligned} \tag{3}$$

where α_i and β_i are the parameters that depend on the location of x_o with respect to the i bin, with the

dimension of variable x and the details about specifying the above two parameters can be found in Hersbach (2000). Since the CRPS has the dimension of the predicted variable, it can be used as a global skill to quantitatively evaluate the performance of an ensemble system. Moreover, the CRPS can be decomposed into the reliability and resolution component, and the decomposition of Eq. (3) can be written as:

$$\begin{aligned}\mathrm{CRPS} &= \sum_{i=0}^{N} g_i[(1-o_i)p_i^2 + o_i(1-p_i)^2] \\ &= \mathrm{Reli} + \mathrm{CRPS_{pot}} \\ &= \sum_{i=0}^{N} g_i(o_i - p_i)^2 + \sum_{i=1}^{N} g_i o_i(1-o_i), \end{aligned} \tag{4}$$

where g_i is the average width of the bin i: $g_i = x_{i+1} - x_i$, and o_i is the average frequency that the observation x_o is less than the middle of the bin i, that is, $(x_i + x_{i+1})/2$. The Reli in Eq. (4) is the reliability term of the CRPS, and $\mathrm{CRPS_{pot}}$ measures the difference between resolution term and the uncertainty term associated with the variable considered. Since the uncertainty term is only related to the reference climatology and is independent of the ensemble system, so the $\mathrm{CRPS_{pot}}$ can be used to represent the resolution information. The CRPS and its decomposition components are all negatively oriented, indicating that smaller values have higher forecast skills. The Reli is equal to 0 if the system is perfectly reliable, and $\mathrm{CRPS_{pot}}$ reaches its minimum for a perfect deterministic system.

The overall performance of the REPS and the GEPS for the precipitation forecasts are evaluated by the CRPS. The significance of CRPS difference between the two systems (the score of GEPS minus the score of REPS used in this paper, hereafter GEPS–REPS) is calculated by the bootstrap technique. The negatively oriented features of the CRPS and its decomposition components lead to the following comparison for the CRPS difference: the positive (negative) GEPS–REPS difference means that the REPS has a better (worse) forecast skill than the GEPS. Figure 2 displays the CRPSs of the REPS and GEPS (left panel), and the CRPS difference between the two systems with the CIs (right panel). Compared to the GEPS, salient skill improvements for 6-h precipitation forecasts are observed in the REPS for all verified periods, and the magnitude of improvement is generally more than 0.20 mm $(6\ \mathrm{h})^{-1}$. Also, there is clear skill improvement for REPS after spin-up time (6 h). The improvement of the REPS compared to the GEPS is statistically significant since the 5% and 95% confidence bounds of the CRPS differences (GEPS–REPS) are all greater than zero for most verified periods. Figure 3 shows the decomposition components of CRPS for 6-h precipitation during different verified periods. The results show that the REPS has clear advantages in both the reliability and resolution features compared to the GEPS, and both improvements are

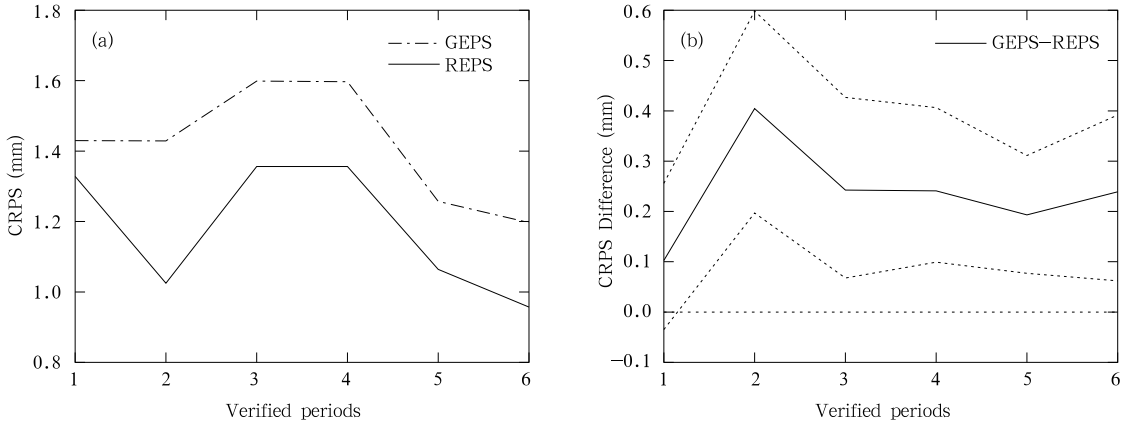

Fig. 2. The CRPSs of accumulated 6-h precipitation from the REPS and GEPS, and the CRPS difference between GEPS and REPS with CIs (5%–95%), as a function of verified periods (1: 00–06 h, 2: 06–12 h, 3: 12–18 h, 4: 18–24 h, 5: 24–30 h, and 6: 30–36 h).

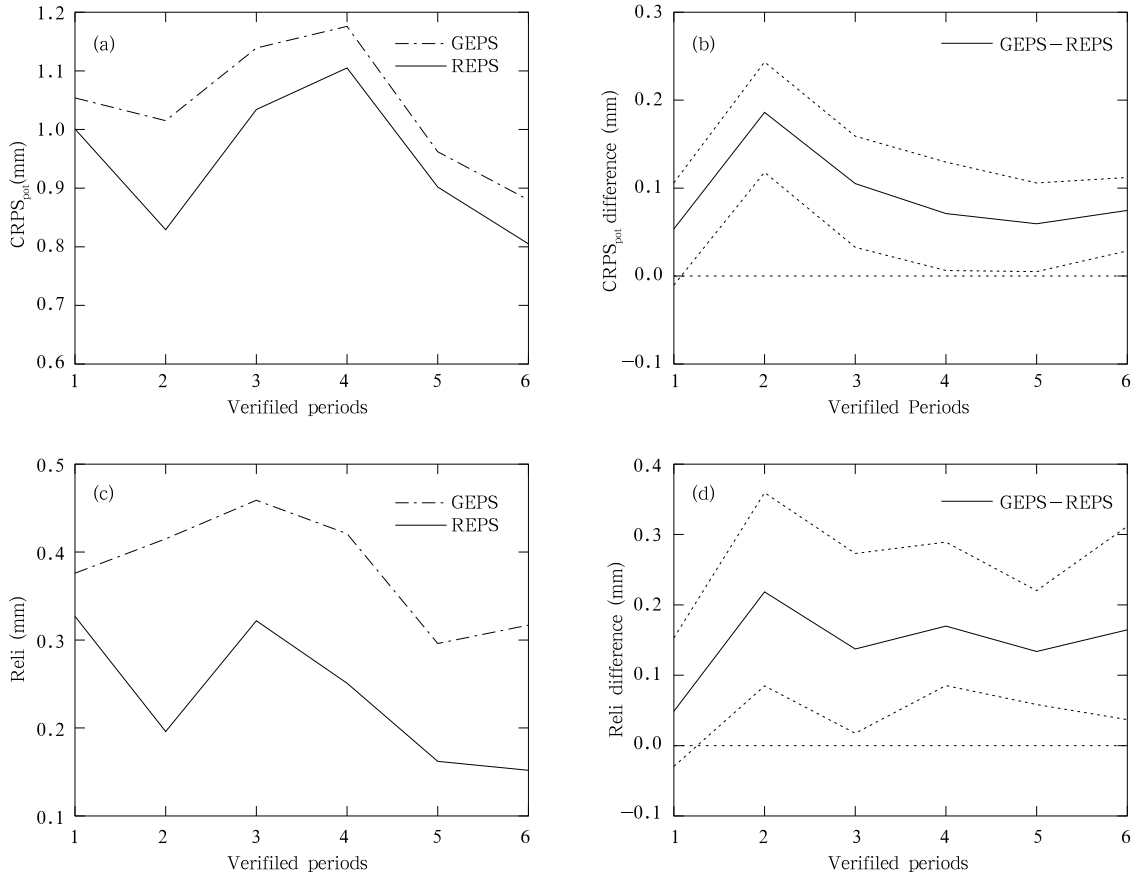

Fig. 3. As in Fig. 2, but for the CRPS resolution (upper panels) and reliability components (lower panels).

statistically significant for verified periods except during the spin-up time. It is noted that for the REPS the improvement in the reliability seems generally larger than in the resolution.

4.2 The reduced centered random variable

The reduced centered random variable (RCRV; Talagrand et al., 1999) is a score to further investigate the reliability property of a system. After Candille et al. (2007), the modified RCRV taking the observation error into account is used in this study:

$$y = \frac{x_{\rm o} - x_{\rm m}}{\sqrt{\sigma_{\rm o}^2 + \sigma^2}}, \tag{5}$$

where $x_{\rm o}$ is the observed value of the verified variable and $\sigma_{\rm o}$ is its observation error, and $x_{\rm m}$ and σ are ensemble mean and standard deviation of the corresponding ensemble prediction. Two statistical parameters can be further derived from Eq. (5). Firstly, the average of y over all the verification realizations,

$$b = E(y), \tag{6}$$

is calculated to measure the bias of the ensemble system (hereafter referred to as the bias term of RCRV); secondly, the standard deviation of y,

$$d = \sqrt{\frac{n}{n-1} E[(y-b)^2]}, \tag{7}$$

is computed to identify the agreement of ensemble spread and the specified observational error. This parameter can provide the dispersion attribute (systematic over- or under-dispersive) of the ensemble system (hereafter referred to as the dispersion term of RCRV). A perfect reliable system will have zero value of bias term b and 1 value of dispersion term d, and the sign of bias value indicates the bias type, and the value of dispersion term d greater/smaller than 1 represents under-dispersion/over-dispersion of a system.

In order to compare the GEPS (A) with REPS

(B) using RCRV score, the following treatments are applied to the bias and dispersion difference: (1) the absolute value of bias difference will be compared, for instance, if $|A| - |B| > (<)0$, which means system B has less (more) bias than system A; (2) function $F(A, B)$ is defined to measure the dispersion difference: if $|\log A| > |\log B|$, then $F(A, B) = e^{|\log A| - |\log B|} - 1$; if $|\log A| < |\log B|$, then $F(A, B) = 1 - e^{|\log A| - |\log B|}$, and the positive (negative) value of $F(A, B)$ indicates that system B has a better (worse) dispersion feature than system A. Figure 4 shows the bias and dispersion term associated with the RCRV from the GEPS and REPS, and bias and dispersion differences between the two systems for 6-h precipitation forecasts. It is noticed that both systems have positive biases, however, the REPS has significantly less bias for all verified periods. With respect to the dispersion attribute, both systems are underdispersive for all listed periods, but the REPS has less dispersion with statistical significance than the GEPS. Moreover, the dispersion of the REPS seems to exhibit small changes for verified periods.

4.3 Brier score and attribute diagram

Besides the above general skill measures, the probabilistic skill for 6-h QPF (6-h accumulated precipitation during 6–36-h forecast lead time) related to certain threshold events also needs to be evaluated.

The Brier score (BS; Brier, 1950; Murphy, 1973) is the most common probabilistic score for dichotomous predictands. It can be estimated from a sample of past forecasts by:

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2, \tag{8}$$

where n is the number of realizations of the forecasts

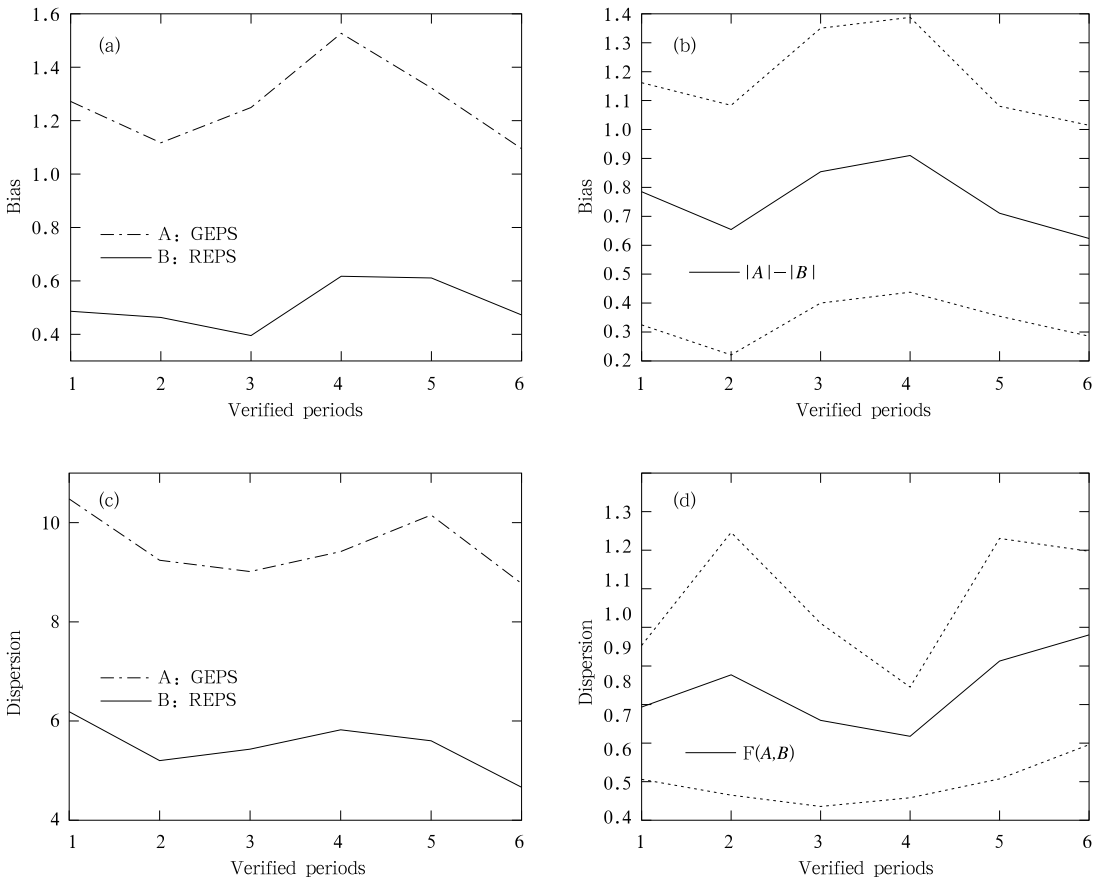

Fig. 4. (a) The biases terms of RCRV from the REPS and GEPS, (b) bias difference between GEPS and REPS with CIs, (c) the dispersion terms of RCRV from the GEPS and REPS, and (d) the dispersion difference between GEPS and REPS with CIs (5%–95%), for accumulated 6-h precipitation as a function of verified periods (1: 00–06 h; 2: 06–12 h; 3: 12–18 h; 4 18–24 h: 5: 24–30 h, and 6: 30–36 h).

issued over which the verification is performed. For each realization i, p_i is the forecast probability of the occurrence of the event, and o_i is actual outcome of the event with value of 1 or 0 depending on whether the event occurs or not. The BS is negatively oriented, with value of zero for the perfect system. The BS can be further decomposed into three components after algebraic transformation (Murphy, 1973):

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{I} N_i(p_i - \overline{o}_i)^2 - \frac{1}{n}\sum_{i=1}^{I} N_i(\overline{o}_i - \overline{o})^2 + \overline{o}(1-\overline{o}), \tag{9}$$

where I is the number of total prescribed forecast probability bins or categories, which is usually determined by the ensemble size (N). For example, with N member, this creates $N+1$ bins with following values: $0/N, 1/N, \ldots, (N-1)/N, N/N$. In this paper, the ensemble sizes of GEPS and REPS are both 15, so the value of I is 16. N_i is the number of times, a forecast p_i is used in the collection of forecast-verification pairs, and $\overline{o}_i = P(o=1|p_i) = \dfrac{1}{N_i}\sum\limits_{k\in N_i} o_k$ is the conditional observed probability, and $\overline{o} = \dfrac{1}{n}\sum\limits_{k=1}^{n} o_k$ is the sample climatology. The three terms on the right-hand side of Eq. (9) are called reliability, resolution, and uncertainty, respectively.

Based on the standards of the Central Meteorological Observatory of CMA, four binary events related to the thresholds 0.1, 1, 2, and 6 mm (6 h)$^{-1}$ for accumulated 6-h precipitation are chosen to evaluate the forecast abilities of the two systems for small to heavy rainfall. Table 1 gives the BSs of the REPS and GEPS, and it is shown that the REPS has better QPF skills for all thresholds compared to the GEPS. Moreover, the significance test of BS difference between the GEPS and REPS reveals that the skill improvement of the REPS is statistically significant for all thresholds (figure omitted).

Table 1. The BSs of accumulated 6-h precipitation from the GEPS and REPS for the different thresholds (0.1, 1, 2 and 6 mm (6 h)$^{-1}$)

	0.1 mm	1 mm	2 mm	6 mm
REPS	0.172	0.114	0.094	0.054
GEPS	0.314	0.182	0.142	0.065

Attribute diagram is the visualized presentation of reliability of system. It compares the prior predicted probability (p_i) against the subsequent observed frequency (o_i) for all probability categories (I). Moreover, the information on resolution of BS and sharpness attribute of the forecast system is also presented in this diagram. The diagonal line in the attribute diagram indicates the perfect reliability (the closer the curve to the diagonal line, the better the reliability is), and the deviation from this line shows the conditional bias. The no resolution line (horizontal and vertical) is featured by the climatological frequency to identify if the forecast probability is higher or lower than the climatological frequency. The flatter the curve is, the lower resolution it has. The sharpness diagram (inset histogram in Fig. 5) shows the occurrence frequency of predicted probability for each forecast probability category. The no skill line is obtained when reliability and resolution components of the BS are equal. When the reliability curve lies between the diagonal and the no skill line, that is, the resolution is greater than reliability, the forecast is skillful.

Figure 5 shows the attribute diagrams of accumulated 6-h precipitation from the GEPS and REPS for thresholds 1, 2, and 6 mm (6 h)$^{-1}$. The sample climatological frequencies associated with the above thresholds are 10.2% (1 mm), 9.2% (2 mm), and 5.2% (6 mm), respectively. It can be clearly found that the REPS is much reliable than the GEPS for all thresholds, though both systems are underdispersive. For the REPS, at 2- and 6-mm thresholds (moderate and high precipitation), the forecast probabilities are underpredicted for the probability categories less than climatological frequency, and are overpredicted for larger values. Moreover, it should be noted that the reliability of the REPS at 2 and 6 mm (6 h)$^{-1}$ is enhanced compared to the one at 1-mm threshold for the low probability categories, and this improvement might indicate the advantage of the REPS forecast for middle to heavy rainfall in the reliability attribute. The relative effect of reliability and resolution component on the BS can be found further through the attribute diagram for the GEPS, which exhibits quite poor resolution (close to the no resolution line) and reliability (below the no skill line). Obviously, these

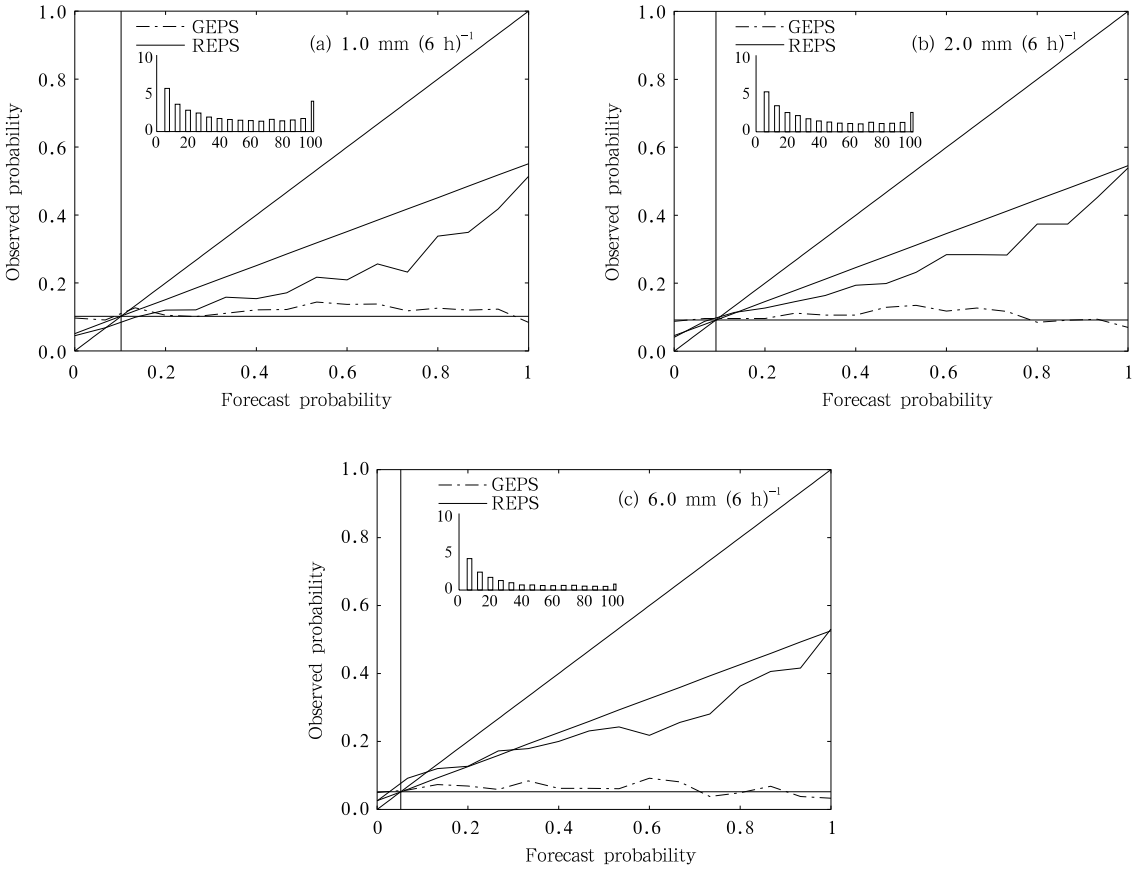

Fig. 5. Attribute diagrams of accumulated 6-h precipitation from the GEPS and REPS for three thresholds: (a) 1, (b) 2, and (c) 6 mm (6 h)$^{-1}$. The slope line below the diagonal line denotes the no skill line, and the horizontal and vertical lines denote the no resolution line (climatological frequency during the verified period). Insert histogram is the sharpness diagram of the REPS for non-zero forecast probability categories. The occurrence frequencies for zero probability categories for different thresholds are 66% (1 mm), 71.9% (2 mm), and 83.7% (6 mm).

have negative contributions to the BS of the GEPS.

The sharpness diagram (shown only for the REPS) reveals the distribution of occurrence frequency for forecast probability categories. The occurrence frequencies for zero probability categories for different thresholds are 66% (1 mm), 71.9% (2 mm), and 83.7% (6 mm), which make positive contributions to the BS for these thresholds. It should be noted that the performance of the GEPS presented here might be deteriorated to some extent by the multi-interpolation process mentioned earlier.

4.4 Area under the relative operating characteristics

The relative operating characteristic (ROC; Mason, 1982) is a verification measure based on the signal detection theory. In the ensemble verification, the ROC diagram is constructed by plotting the hit rate $H = a/(a+c)$ versus the false alarm rate $F = b/(b+d)$ for a range of probability thresholds, where a, b, c, and d are components of the contingency table (Table 2). In this study, 15 values of threshold probabilities are used, ranging from 1/15 to 15/15 with an even interval of 1/15. The area under the ROC curve (AROC) is usually used to indicate the discriminating ability of the event at the selected thresholds, which has the value of 1 for perfect forecast, and the value of 0.5 for

Table 2. The contingency table of precipitation forecasts and observations for a certain event

	Observed	Not observed
Forecasted	a	b
Not forecasted	c	d

An event occurs when the precipitation exceeds a threshold.

no skill forecast.

Figure 6 gives the AROCs of accumulated 6-h precipitation from the GEPS and REPS, and AROC difference between GEPS and REPS. The stable AROC values more than 0.7 are observed in the REPS for all thresholds, and the values of GEPS remain around 0.5 with small changes for verified thresholds, indicating that we can obtain more skillful forecasts from the REPS for small to heavy rainfall. The behavior of the GEPS is consistent with its resolution feature presented in the attribute diagram, showing that it has no forecast skill compared to the climatological frequency of the sample. From the AROC difference between the GEPS and REPS, it can be found that the skill improvement of the REPS is statically significant for all thresholds compared to the GEPS.

4.5 The potential economic value

The potential economic value (PEV; Richardson, 2000; Zhu et al., 2002) is a user-oriented measure based on the cost-loss analysis method for determining the potential economic benefit associated with the use of ensemble forecast relative to the use of climatology information. The specific definition of PEV can be found in related articles. The calculation of PEV is given here only for clarification:

$$\mathrm{PEV} = \frac{\min(r,\overline{o}) - F(1-\overline{o})r + H\overline{o}(1-r) - \overline{o}}{\min(r,\overline{o}) - \overline{o}r}, \quad (10)$$

where $r = C/L$ is the user specified cost-loss ratio, $\overline{o}$ is the climatological probability of the occurring event, and H and F are the hit rate and false alarm rate defined in the description of the ROC. Based on Eq. (10), PEV is the function of the probability thresholds since H and F are the functions of a set of probability thresholds. The PEV is positively oriented with value of 1 for perfect deterministic forecast. The PEV also allows for comparison of the economic value of the ensemble forecast and the one of an equivalent control run. Usually, the envelope of PEV (called optimal PEV) among all probability thresholds is used to represent economic value of the ensemble system, and in this study we utilize the optimal PEV. Figure 7 depicts the optimal PEV from the REPS and its control run for 1, 2, and 6 mm $(6\ \mathrm{h})^{-1}$ thresholds, as a function of cost-loss ratio C/L. The positive optimal PEV of the control forecast for three thresholds are mainly located for C/L ratio ranges of 5%–30%. Compared to the control forecast, not only the range of cost-loss ratios, for which the ensemble forecasts exhibit positive value, is widened up to 3%–40%, but also the corresponding optimal PEV is improved substantially. The largest PEV improvement from the REPS forecast compared to the control forecast is found around 5% for the C/L ratio at 6 mm. Also, it is noticeable that the PEV improvement from the REPS tends to distinctly increase for the C/L ratio range of 20%–40% for all thresholds compared to the control forecast.

Note that for each threshold, the largest economic

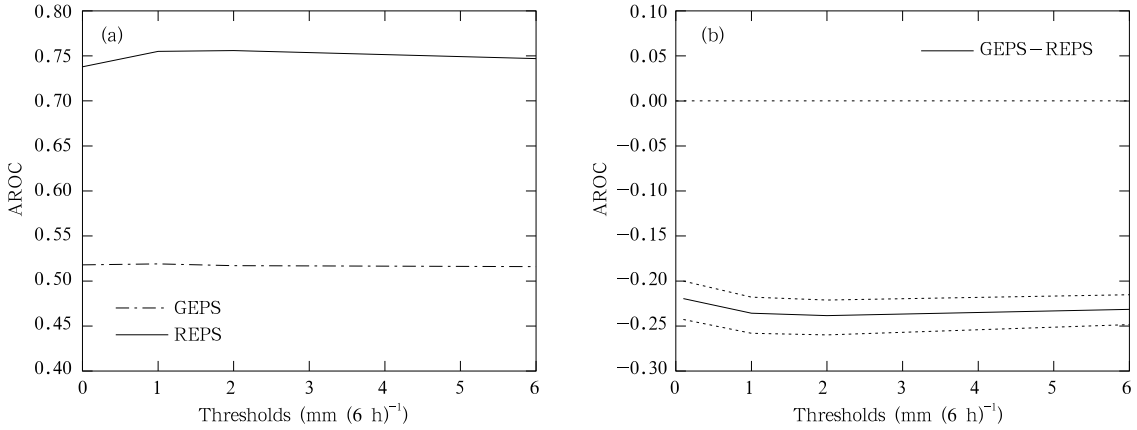

Fig. 6. AROCs of accumulated 6-h precipitation from the GEPS and REPS, and AROC difference between the GEPS and REPS with CIs (5%–95%), as a function of thresholds (0.1, 1, 2, and 6 mm $(6\ \mathrm{h})^{-1}$).

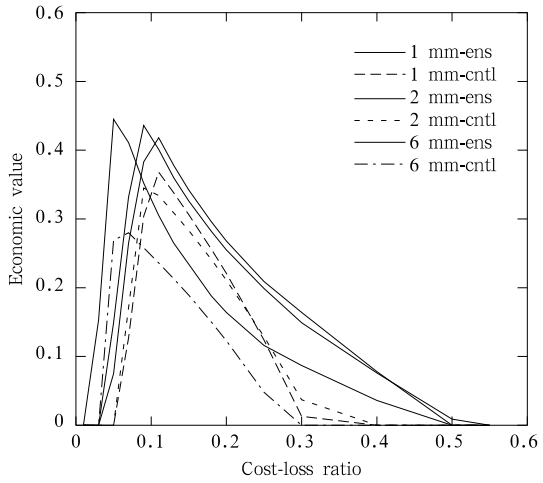

Fig. 7. The optimal REV of the REPS (solid lines) and the control run (dashed lines) for three thresholds (1, 2, and 6 mm $(6\text{h})^{-1}$).

value, as expected from its definition, is obtained when the C/L ratio is approximately equal to climatological frequency, which is 10% for 1 mm, 9% for 2 mm, and 11% for 6 mm. Also, it is interesting to find that with increasing thresholds, this peak value of the REPS is enhanced, showing a maximum value of 0.45 at 6 mm, which is contrary to the reduced peak values with increasing thresholds observed in the control forecast. The above results indicate that a larger group of users can benefit from using the ensemble forecasts compared to using the control forecasts, and encouraging advantage of using the REPS for heavy rainfall forecast is obtained.

4.6 Brier skill score of the REPS for accumulated 24-h precipitation

The Brier skill score (BSS; Jollife and Stephenson, 2003) is a forecast skill score derived from the BS with respect to a reference system:

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}}. \tag{11}$$

The BSS is positively oriented with value of 1 for perfect forecast, and positive value of BSS indicates the skillful forecast compared to the reference system. In this study, the GEPS forecast is used as reference system ($\mathrm{BS}_{\mathrm{ref}}$) to evaluate the performance of the REPS, and the positive (negative) BSS indicates that the REPS forecast is better (worse) than the GEPS forecast.

In previous sections, the significant advantages of the REPS for 6-h QPFs compared to the GEPS are observed. It is now of interest to investigate the forecast skill improvement when using the REPS instead of the GEPS for the accumulated 24-h precipitation (from 12- to 36-h forecast lead time). The evaluation here is conducted for 11 thresholds (1, 2.5, 5, 10, 15, 20, 30, 35, 40, 45, and 50 mm $(24\text{ h})^{-1}$). Figure 8 displays the BSS of the REPS with CIs for accumulated 24-h precipitation, and the positive BSSs are observed for all thresholds, which indicate that we can obtain more skillful 24-h QPFs when using the REPS forecasts instead of using the GEPS. Moreover, based on the 95% confidence bound shown in Fig.8, it can be found that the precipitation forecast skill improvements by the REPS can be significant up to 30 mm.

5. Summary and conclusions

A REPS at NMC was originally established and had been developing as one of the B08RDP participants for providing mesoscale ensemble forecasting to the Beijing 2008 Olympic Games. After the Olympic Games, this system has been running in real time,

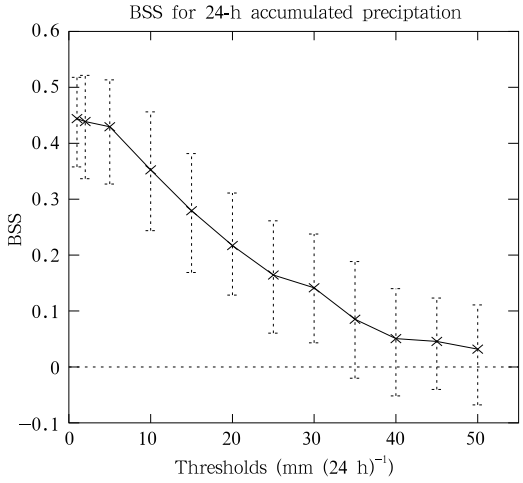

Fig. 8. The BSSs of accumulated 24-h precipitation from the REPS, as a function of precipitation thresholds (1, 2.5, 5, 10, 15, 20, 30, 35, 40, 45, and 50 mm $(24\text{ h})^{-1}$). The error bars represent the 5%–95% confidence interval.

and its products, especially surface parameters, provide extra useful information to forecasters besides the existing GEPS forecasts and other deterministic forecasts. This study focuses on the evaluation of the REPS for short-range precipitation forecasts in comparison with its control run and the GEPS forecasts, and comprehensive verifications are performed over a 36-day period during the Beijing Olympic Games (21 July–24 August 2008). All verification measures used in this study indicate that the REPS performs significantly better than the GEPS for precipitation forecasts, showing that the REPS has prevailing advantages for 6-h precipitation forecasts in the global forecast skill measures. By specifying advantages of the REPS, it is shown that the REPS has better reliability and resolution (discrimination) attributes compared to the GEPS. Moreover, the superior reliability of REPS can be characterized with less bias and better dispersion.

The PEV comparison between the REPS and its control run reveals that the ensemble-based probabilistic forecasts exhibit much higher PEV than the control forecast. Moreover, this PEV improvement tends to increase with increasing of rainfall amount. For the forecasts of accumulated 24-h precipitation, the forecast skill improvement of the REPS against the GEPS as the reference system can be found up to 50 mm, in which the statistic significance is up to 30 mm. Generally, the verification results in this study support the expectation that regional ensemble forecasts are capable of providing more useful information for the short-range forecasts.

Note that the unavoidable multi-interpolation process for the GEPS data might be unfavorable for the evaluation of the GEPS, and the lack of model perturbation might play another important role for the poor performance of the GEPS. In addition, the conclusion drawn here might be affected by undersampling effect (only 36-day cases), since some probabilistic scores are sensitive to the sample climatology. Fortunately, the pre-operationally running of the REPS at NMC (starting from November 2010) is still ongoing, which allows us to conduct extended studies for larger samples.

REFERENCES

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3.

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

Buizza, R., P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu, 2005: A comparison of the ECMWF, MSC and NCEP global ensemble. *Mon. Wea. Rev.*, **133**, 1076–1097.

Candille, G., J. Cote, P. L. Houtekamer, and G. Pellerin, 2007: Verification of an ensemble prediction system against observations. *Mon. Wea. Rev.*, **135**, 2688–2699.

Chessa, P. A., G. Ficca, M. Marrocu, and R. Buizza, 2004: Application of a limited-area short-range ensemble forecast system to a case of heavy rainfall in the Mediterranean region. *Wea. Forecasting*, **19**, 566–581.

Deng Guo, Gong Jiandong, Deng Liantang, et al., 2010: The development of mesoscale ensemble prediction system at National Meteorological Center. *J. Appl. Meteor. Sci.*, **21**(5), 513–523. (in Chinese)

Du, J., and M. S. Tracton, 2001: Implementation of a real-time short-range ensemble forecasting system at NCEP: An update. Preprint, Ninth Conf. on Mesoscale Processes, Ft. Lauderdale, Florida, Amer. Meteor. Soc., 355–356.

——, G. DiMego, M. S. Tracton, and B. Zhou, 2003: NCEP short-range ensemble forecasting (SREF) system: Multi-IC, multi-model and multi-physics approach. Research Activities in Atmospheric and Oceanic Modelling (edited by J. Cote), Report 33, CAS/JSC Working Group Numerical Experimentation (WGNE), WMO/TD-No. 1161, 5.09–5.10.

Duan, Y., J. Gong, J. Du, M. Charron, J. Chen, G. Deng, G.. DiMego, M. Hara, M. Kunni, X. Li, Y. Li, K. Saito, H. Seko, Y. Wang, and C. Wittmann, 2011: An overview of the Beijing 2008 Olympics Research and Development Project (B08RDP). *Bull. Amer. Meteor. Soc.*, doi: 10.1175/BAMS-D-11-00115.1.

Efron, B., and R. Tibshirani, 1993: *An Introduction to the Bootstrap*. Chapman and Hall, 436 pp.

Heizenreder, D., S. Trepte, and M. Denhard, 2006: SRNWP-PEPS: A regional multi-model ensemble in Europe. *The European Forecaster, Newsletter of the WGCEF*, **11**, 29–35.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570.

Janjié, Z. I., J. P. Gerrity Jr., and S. Nickovic, 2001: An alternative approach to nonhydrostatic modeling. *Mon. Wea. Rev.*, **129**, 1164–1178.

Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley and Sons, 240 pp.

Li Zechun and Chen Dehui, 2002: The development and application of the operational ensemble prediction system at National Meteorological Center. *J. Appl. Meteor. Sci.*, **13**(1), 1–15. (in Chinese)

Marsigli, C., A. Montani, F. Nerozzi, and T. Paccangnella, 2005: The COSMO-LEPS mesoscale ensemble system: Validation of the methodology and verification. *Nonlinear Processes Geophys.*, **12**, 527–536.

Mason, I., 1982: A model for assessment of weather forecasts. *Aus. Meteor. Mag.*, **30**, 291–303.

Montani, A., and Coauthors, 2003: Operational limited-area ensemble forecasts based on the Lokal Modell. *ECMWF Newsletter*, No. 98, ECMWF, Reading, United Kingdom, 2–7.

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

Park, Y. -Y., R. Buizza, and M. Leutbecher, 2008: TIGGE: Preliminary results on comparing and combining ensembles. *Quart. J. Roy. Meteor. Soc.*, **134**, 2029–2050.

Richardson, D. S., 2000: Skill and economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–668.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in meteorology. World Meteorological Organization, World Weather Watch Rep. 8, Tech. Doc., 358, 114 pp.

Stensrud, D. J., H. E. Brooke, J. Du, M. S. Tracton, and E. Rogers, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev.*, **127**, 433–446.

——, J. Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev.*, **128**, 2077–2107.

Talagrand, O., R. Vautard, and B. Strauss, 1999: Evaluation of probabilistic prediction systems. Proc. Workshop on Predictability, Reading, United Kingdom, ECMWF, 1–25.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330.

——, and ——, 1997: Ensemble forecasting at NCEP: The breeding method. *Mon. Wea. Rev.*, **125**, 3297–3319.

Wang, Y., M. Bellus, C. Wittmann, M. Steinheimer, F. Weidle, S. Ivatek-Sahdan, A. Kann, W. Tian, X. Ma, S. Tascu, and E. Bazile, 2011: The Central European limited area ensemble forecasting system: ALADIN-LAEF. *Quart. J. Roy. Meteor. Soc.*, **137**, 483–502, doi: 10.1002/qj.751.

Zhu, Y., Z. Toth, R. Wobus, D. Richarson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.*, **83**, 74–83.