Tellus
SERIES A
DYNAMIC
METEOROLOGY
AND OCEANOGRAPHY
PUBLISHED BY THE INTERNATIONAL METEOROLOGICAL INSTITUTE IN STOCKHOLM

Tellus (2011), 63A, 531–549

TELLUS

Verification and intercomparison of mesoscale ensemble prediction systems in the Beijing 2008 Olympics Research and Development Project

By MASARU KUNII[1,*], KAZUO SAITO[1], HIROMU SEKO[1], MASAHIRO HARA[1], TABITO HARA[2], MUNEHIKO YAMAGUCHI[1], JIANDONG GONG[3], MARTIN CHARRON[4], JUN DU[5], YONG WANG[6] and DEHUI CHEN[3], [1]*Meteorological Research Institute, Tsukuba, Ibaraki 305-0052, Japan;* [2]*Numerical Prediction Division, Japan Meteorological Agency, Tokyo, Japan;* [3]*National Meteorological Center, Chinese Meteorological Administration, Beijing, China;* [4]*Environment Canada, Dorval, Québec, Canada;* [5]*National Centers for Environmental Prediction, Washington D.C., USA;* [6]*Zentralanstalt für Meteorologie und Geodynamik, Wien, Austria*

(Manuscript received 20 April 2010; in final form 13 January 2011)

ABSTRACT

During the period around the Beijing 2008 Olympic Games, the Beijing 2008 Olympics Research and Development Project (B08RDP) was conducted as part of the World Weather Research Program short-range weather forecasting research project. Mesoscale ensemble prediction (MEP) experiments were carried out by six organizations in near-real time, in order to share their experiences in the development of MEP systems. The purpose of this study is to objectively verify these experiments and to clarify the problems associated with the current MEP systems through the same experiences.

Verification was performed using the MEP outputs interpolated into a common verification domain with a horizontal resolution of 15 km. For all systems, the ensemble spreads grew as the forecast time increased, and the ensemble mean improved the forecast errors compared with individual control forecasts in the verification against the analysis fields. However, each system exhibited individual characteristics according to the MEP method.

Some participants used physical perturbation methods. The significance of these methods was confirmed by the verification. However, the mean error (ME) of the ensemble forecast in some systems was worse than that of the individual control forecast. This result suggests that it is necessary to pay careful attention to physical perturbations.

1. Introduction

In recent years, the need for probabilistic information to prevent natural hazards such as localized heavy rainfall or wind gust has arisen. With the progress of numerical models and computer capability, it has become possible to simulate mesoscale phenomena with probabilistic information using mesoscale ensemble prediction (MEP) systems. In the process of developing MEP systems, it is difficult to compare systems because many different strategies are used in constructing them; therefore, adequate experiments are necessary to assess the systems' capabilities and forecast skills. As a pioneering project, the Storm and Mesoscale Ensemble Experiment (SAMEX) has been performed to compare MEP systems under the same conditions (Hou et al., 2001). In this experiment, four different numerical models with horizontal resolution of 30 km were used. The initial and boundary perturbations were generated with the breeding method (Toth and Kalnay, 1993; 1997), the random perturbation method (Mullen and Baumhefner, 1989) or the simple scaled lagged average forecasting (SLAF) method (Ebisuzaki and Kalnay, 1991). In addition, the initial and boundary perturbations, as well as several combinations of changes in the model physical parametrizations, were considered.

Since then, more advanced ensemble prediction systems (EPSs) have been developed [e.g. the singular vector method and the Ensemble Kalman filter (EnKF) technique]. Singular vectors calculated using coupled tangent linear and its adjoint models are optimum structures for describing perturbation growth over a finite forecast time interval. This method was developed at the European Centre for Medium-Range Weather Forecasts

*Corresponding author.
e-mail: mkunii@mri-jma.go.jp
DOI: 10.1111/j.1600-0870.2011.00512.x

Table 1. Data transfer list in the B08RDP experiment

Level	Surf	850 hPa	700 hPa	500 hPa	250 hPa
U	○ (10 m)	○	○	-	○
V	○ (10 m)	○	○	-	○
T	○ (2 m)	○	-	○	○
Z	-	-	-	○	-
RH	○ (2 m)	○	-	-	-
P_{SEA}	○	-	-	-	-
RAIN	○	-	-	-	-
CAPE			○		
CIN			○		
SAUI			○		

Table 2. Specifications of control forecasts of each participant in the B08RDP experiment

Participants	Initial condition	Forecast model
MRI/JMA	JMA Meso-4D-Var (+NHM 3-h forecast)	NHM (L40)
MSC	MSC Global EnKF	GEM (L28)
ZAMG & Meteo-Fr.	ECMWF Global 4D-Var	ALADIN (L37)
NCEP	NCEP Global 3D-Var	WRF-ARW, WRF-NMM, GEFS-Downscaled (L60)
NMC/CMA	WRF-3D-Var	WRF-ARW (L31)
CAMS/CMA	GRAPES-3D-Var	GRAPES (L31)

(ECMWF) (Buizza et al., 1993) to create a set of initial perturbations for ensemble predictions as well as for sensitivity analyses of adaptive observations to improve the initial conditions of their global models. The EnKF, to begin with Evensen (1994), has been the most attractive method because both data assimilation and ensemble forecasting can be performed based on the same model. Moreover, it can generate initial perturbations reflecting analysis errors, which are based not only on model constraints but also on to the observational networks through data assimilation.

The World Weather Research Program (WWRP) short-range weather forecasting research project was conducted for the Beijing 2008 Olympic Games. The project was divided into two components: the Forecast Demonstration Project (B08FDP) for the forecast time (FT) of 0–6 h based on nowcasting, and the Research and Development Project (B08RDP) for FT = 6–36 h based on the MEP. The B08FDP experiment was conducted to implement advanced nowcast systems for the Beijing 2008

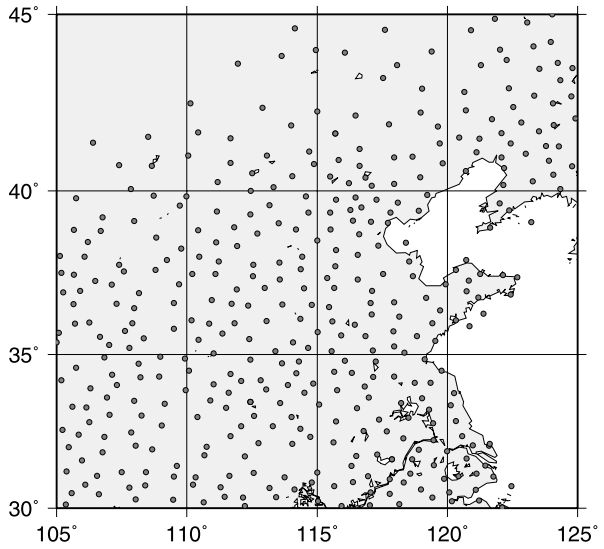

Fig. 1. Surface observation stations used for verification in the B08RDP experiment.

Olympic Games, and to quantify the impact of the nowcast systems on the quality of forecasters and end users. On the other hand, the objectives of the B08RDP were to improve understanding of the high-resolution probabilistic prediction processes through numerical experimentation and to share experiences in developing a real-time MEP system. The B08FDP/RDP project succeeded the Sydney Olympic 2000 Forecast Demonstration Project (Sydney 2000FDP), which aimed to demonstrate the benefits of disseminating real-time forecasts to users in high impact weather situations (Keenan et al., 2003).

In the B08RDP experiment, six organizations operated their MEP systems in near-real time. Perturbations based on bred vectors, singular vectors and the EnKF approach were utilized in their systems. In addition, participants applied their advanced forecast models for this experiment, including state-of-the-art non-hydrostatic models. It is important to understand the characteristics of the MEP system through verification and intercomparison in order to refine each MEP system in the future. Verification is an indispensable part of meteorological research and operational forecasting activities (Casati et al., 2008).

The purpose of this work is verification and intercomparison of EPSs in B08RDP, and clarification of the problems involved in the present MEP systems. Verification was performed for non-perturbed control forecasts as well as ensemble forecasts to assess the characteristics of the individual forecast model and the improvement by ensemble forecasting.

This paper is organized as follows. Section 2 briefly describes the B08RDP and MEP systems of each participant. Section 3 presents the verification results as deterministic forecasts, and Section 4 presents them as probabilistic forecasts. Results are discussed in Section 5. Section 6 provides a summary and conclusions.

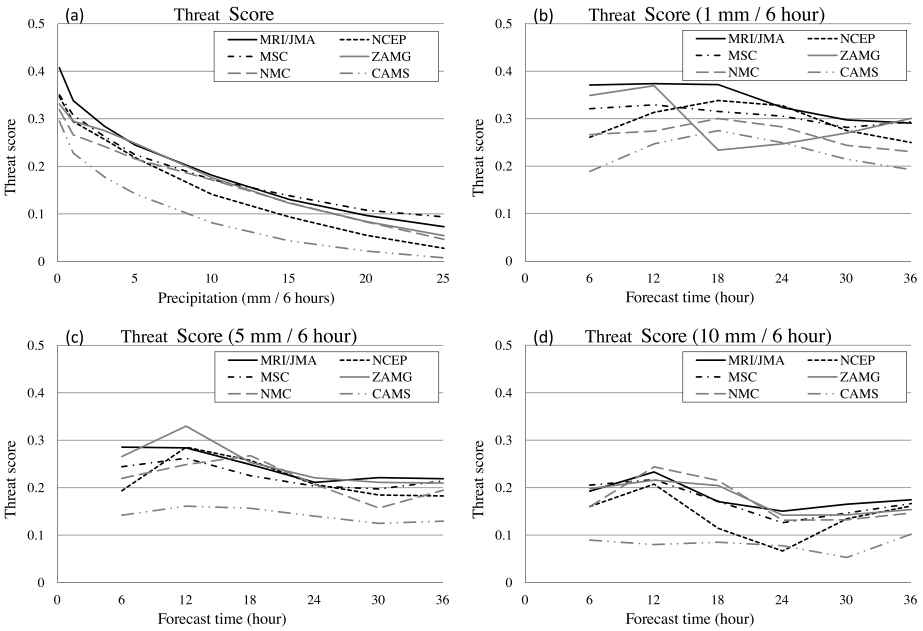

Fig. 2. Threat scores of all participants for 6-h accumulated precipitation averaged between 25 July and 23 August 2008 (a) as a function of threshold, (b) at a threshold of 1 mm, (c) at a threshold of 5 mm and (d) at a threshold of 10 mm.

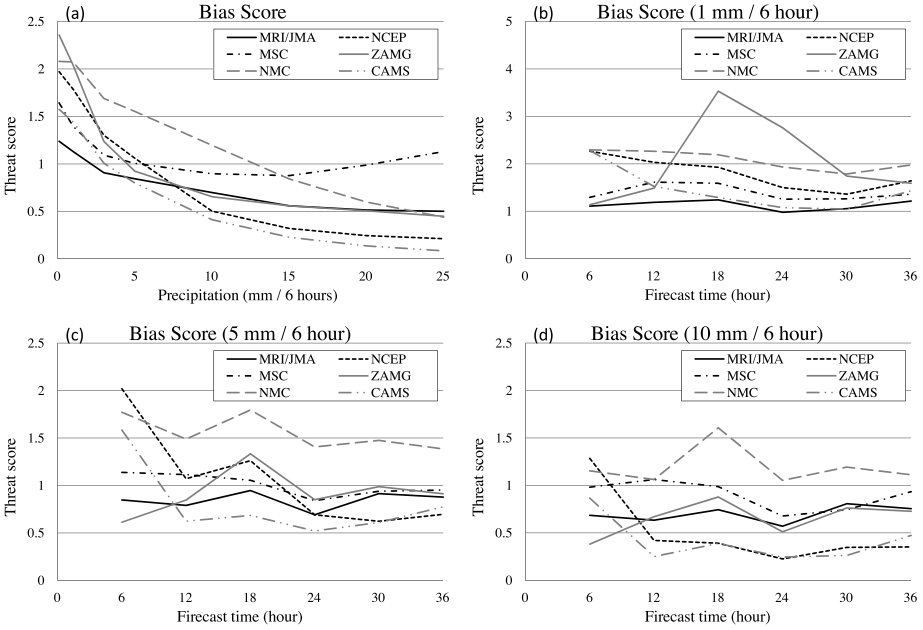

Fig. 3. Same as Fig. 2 but for bias scores.

2. B08RDP experiment system

The Beijing 2008 Olympic Games were held on 8–24 August in Beijing, China. According to statistical data over the past 20 yr, the precipitation frequency in Beijing during July and August is approximately 49.2%, and the occurrence frequency of thunderstorms is 25.9%. Moreover, the average maximum temperature is 29.9 °C, and the maximum temperature is as high as 35.7 °C. Therefore, the appropriate probabilistic forecasts for intense rainfall and extremely hot temperature are key to assessing the MEP system for this project.

2.1. Outline of B08RDP experiment

The WWRP B08RDP experiment in 2008 was conducted for 1 month, from 24 July to 24 August, in conjunction with the

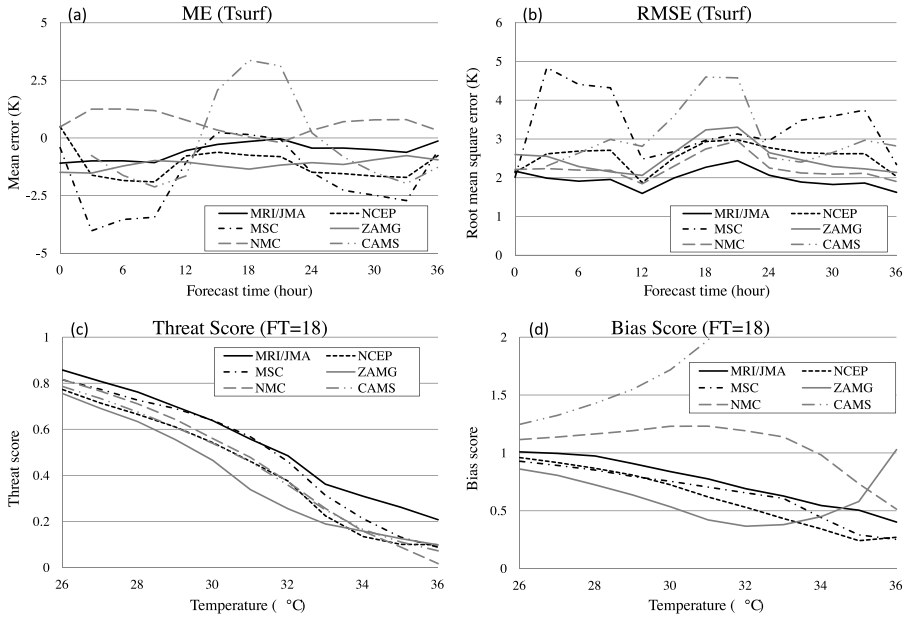

Fig. 4. (a) Mean errors (MEs) of all participants for surface (2 m) temperature averaged between 25 July and 23 August 2008. (b) The same as (a), but for root mean square errors (RMSEs). (c) The same as (a), but for threat scores for FT = 18. (d) The same as (c), but for bias scores.

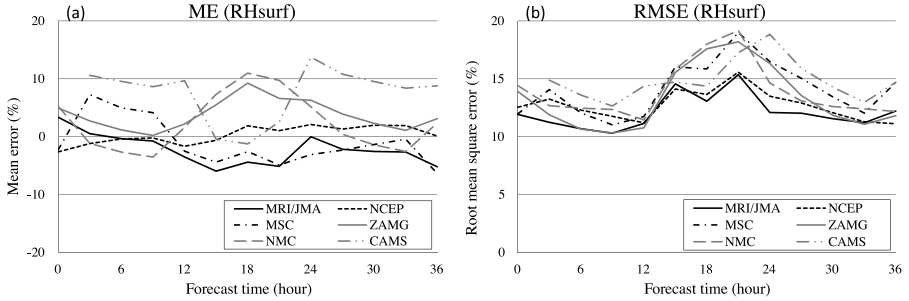

Fig. 5. (a) Mean errors (MEs) of surface (2 m) relative humidity for 36 h forecasts. (b) The same as (a), but for root mean square errors (RMSEs).

Beijing Olympic Games (8–24 August 2008). Collaborating with the Numerical Prediction Division of JMA, the Meteorological Research Institute (MRI/JMA) participated in the B08RDP component. In addition to MRI/JMA, the National Centers for Environmental Prediction (NCEP, USA), the Meteorological Service of Canada (MSC, Canada), the Zentral Anstalt fur Meterologie und Geodynamik (ZAMG, Austria) and the Meteo-France (France), the National Meteorological Center of the China Meteorological Administration (NMC, China) and the Chinese Academy of Meteorological Sciences (CAMS, China) participated in the RDP component. Some papers related to the project have already been published (e.g. Zhou and Du, 2010; Kunii et al., 2010). Moreover, MRI/JMA's activities related to B08FDP/RDP have been discussed by Saito et al. (2010).

2.2. *Participating systems of B08RDP*

In the B08RDP experiment, all participants operated their MEP systems in near-real time and were required to send their ensemble forecast results to the CMA ftp server on time every day. Each forecast output contained specific elements within the common verification region over Beijing. Participants could decide the number of MEP members, depending on their own computer environments, and they were to prepare initial and boundary conditions (ICs and BCs) for MEP. The experiment systems of all participants are given later.

2.2.1. Forecast models. Participants were basically required to operate their MEP systems using regional models with a horizontal resolution of 15 km. In the NCEP system, two versions of the Weather Research and Forecasting (WRF) modelling system were used. One was the Nonhydrostatic Mesoscale Model (WRF-NMM, Janjić et al., 2001), and the other was the Advanced Research version of the WRF (WRF-ARW, Skamarock et al., 2005). In addition to these regional models, the NCEP Global Ensemble Forecast System (GEFS) was also utilized in their MEP system. Specifically, it was a downscaled product. The downscaling method is a dual-resolution-based method called Hybrid Ensembling (Du, 2004). It superimposes forecast

Table 3. Specifications of ensemble forecasts of each participant in the B08RDP experiment

Participants	Model	IC	IC perturbation	LBC	LBC perturbation	Physical perturbation
NCEP	WRF-ARW (R15L60M5) WRF-NMM (R15L60M5) GEFS (T284L60M5)	NCEP global 3D-Var	Breeding	NCEP global EPS	NCEP global EPS	Multimodel
MRI/JMA	JMA-NHM (R15L40M11)	Meso 4D-Var	Targeted global SV	JMA GSM forecast	GSM forecast from targeted SV	None
MSC	GEM (R15L28M20)	MSC global EnKF	MSC global EnKF	MSC global EPS	MSC global EPS	Physical tendency perturbation with Markov chain, surface perturbation
ZAMG & Meteo-Fr.	ALANDIN (R15L37M17)	ECMWF global 4D-Var	Blending ECMWF SV with ALADIN bred mode	ECMWF global EPS	ECMWF global EPS	Multiphysics
NMC/CMA	WRF-ARW (R15L31M15)	WRF 3D-Var	Breeding	CMA global EPS	Global EPS	Multiphysics
CAMS/CMA	GRAPES (R15L31M9)	GRAPES 3D-Var	Breeding	CMA global EPS	Global EPS	Multiphysics

Note: In the 'Model' column, R and T denote the horizontal resolution of the forecast model; L is the vertical level of the model; and M is the number of EPS members.

variances from a lower resolution ensemble onto a higher resolution single run to create a higher resolution ensemble. By taking the advantage of higher resolution model forecast, the resulting new ensemble is generally superior to the original lower resolution ensemble. In this case, the lower resolution ensemble is the NCEP T126L26 GEFS, while the higher resolution single run is the NCEP T284L60 single GFS (Global Forecast System) forecast. For a more detailed description of this dynamical downscaling method, refer to Du (2004).

The WRF-ARW is also used in the NMC system. The MSC system was based on the Global Environmental Multiscale (GEM) model (Côté et al., 1998a, b; Yeh et al., 2002) and its limited area version (GEM-LAM). The ZAMG system was established in the frame of the Aire Limitée Adaptation dynamique Développement InterNational (ALADIN) model. The forecast model of MRI/JMA was the non-hydrostatic model of JMA (NHM, Saito et al., 2006, 2007). Some modifications were made to the physical processes in B08RDP on the Kain-Fritsch convective parametrization and the soil wetness, whose details are described in section E-2 of Saito et al. (2010). The CAMS system was based on the non-hydrostatic model of the Global/Regional Assimilation and Prediction System (GRAPES, Chen et al., 2008). Specifications of the numerical models in B08RDP are listed in Table 2, with the ICs.

2.2.2. Initial and boundary perturbations. In the NCEP, NMC and CAMS systems, the breeding method (Toth and Kalnay, 1993, 1997) was adopted to produce initial perturbations. NMC and CAMS utilized the CMA global EPS based on the breeding method for the lateral BC, whereas the lateral BC for the NCEP system was provided by the ensemble transform based NCEP global EPS (Wei et al., 2008). On the other hand, the targeted global singular vectors (Yamaguchi et al., 2009) were used for both initial and boundary perturbations in the MRI/JMA system. MRI/JMA compared initial perturbation methods prior to the B08RDP experiment (details have been given in Saito et al. 2011). In the ZAMG system, the blending method between bred vectors and singular vectors (Derková and Bellus, 2007) was applied. The breeding method was newly developed for their regional model ALADIN, and the singular vectors were derived from the ECMWF global EPS. The EnKF technique was used in the MSC system for IC perturbations. The lateral boundary perturbations were provided by the MSC GEPS (Charron et al., 2010).

2.2.3. Model perturbations. Considering forecast model uncertainty in MEP systems is an essential issue for providing probabilistic information for risk management. In the B08RDP experiment, three strategies were applied: multiphysics, multimodel and stochastic parametrization perturbations on model

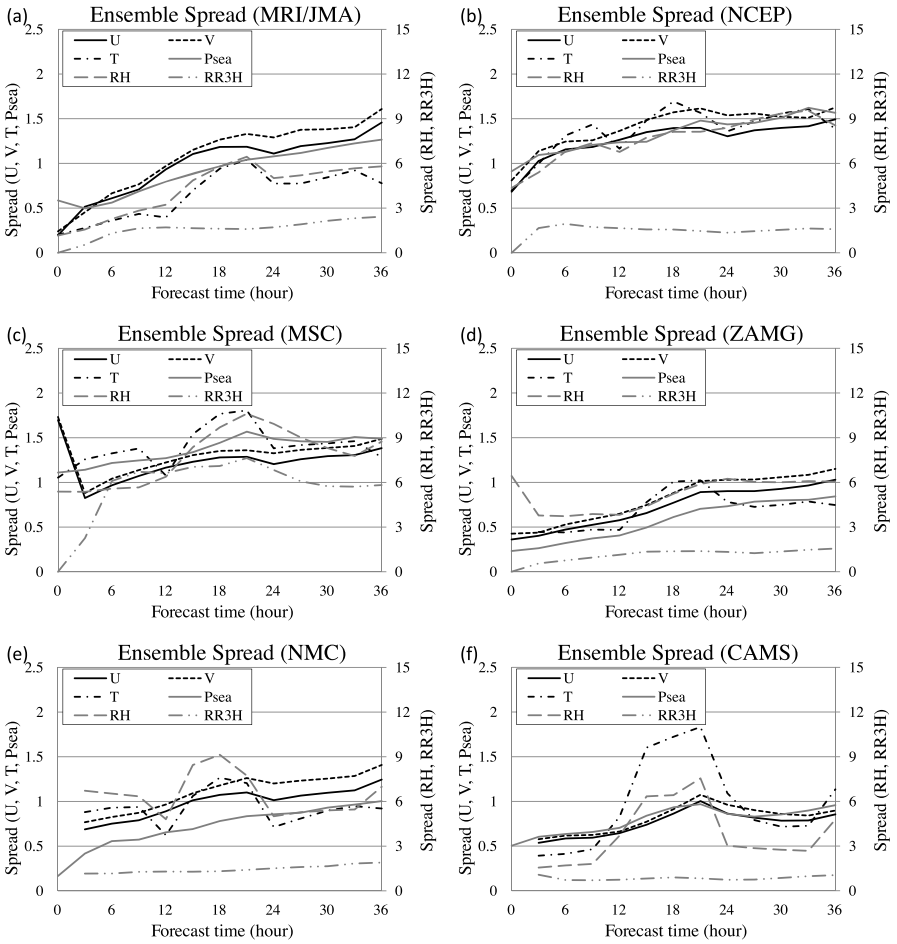

Fig. 6. Ensemble spreads of surface variables of (a) MRI/JMA, (b) NCEP, (c) MSC, (d) ZAMG, (e) NMC and (f) CAMS. The left axis indicates the spreads of zonal wind [U (m s^{-1})], meridional wind [V (m s^{-1})], temperature [T (K)] and sea level pressure [P_{sea} (hPa)]. The right axis indicates the spreads of relative humidity [RH (%)] and 3-h accumulated precipitation [RR3H (mm)].

physics. The multiphysics technique was utilized in the ZAMG, NMC and CAMS systems. The methods were based on the variations of cumulus convective parametrizations, boundary layer schemes, land surface schemes and turbulent transport and diffusion processes. NCEP employed three different forecast models to represent the model diversity in physics and dynamics. The stochastic perturbation method on the model physical tendencies and surface parameters was used in the MSC system. No physical perturbation was implemented in the MRI/JMA system.

2.3. Data management

In the B08RDP experiment, participants were requested to run their ensemble predictions for a forecast time up to 36 h, starting every day at 1200 UTC. The results were interpolated into common verification grids with a resolution of 0.15° over a common verification domain (105 to 125°E, 30 to 45°N). The grid point values were converted into GRIB2 format and transferred to the CMA's data server by 2230 UTC each day. The variables included in the results are listed in Table 1. As an optional diagnostic parameter, the Sauna index (SAUI) was included in the list. SAUI is defined as

$$\mathrm{SAUI}\,(K) = 0.5 \times T_{\mathrm{surf}} + 0.3 \times T_{\mathrm{Dsurf}} + 15, \quad (1)$$

where T_{surf} is temperature and T_{Dsurf} is dewpoint temperature at surface level.

The ensemble products of the participants were displayed by CMA on the B08RDP website in near-real time and were also utilized as reference information for the Beijing Meteorological Bureau's daily forecast for the Olympic Games' venues.

3. Verification of the control forecasts

Motivated by the interest in the performances of non-perturbed control forecasts of each participant, an intercomparison of the control run of each system was performed first for the 2008 B08RDP experiment period over the common verification area. The analysis system used for preparing the IC and the forecast

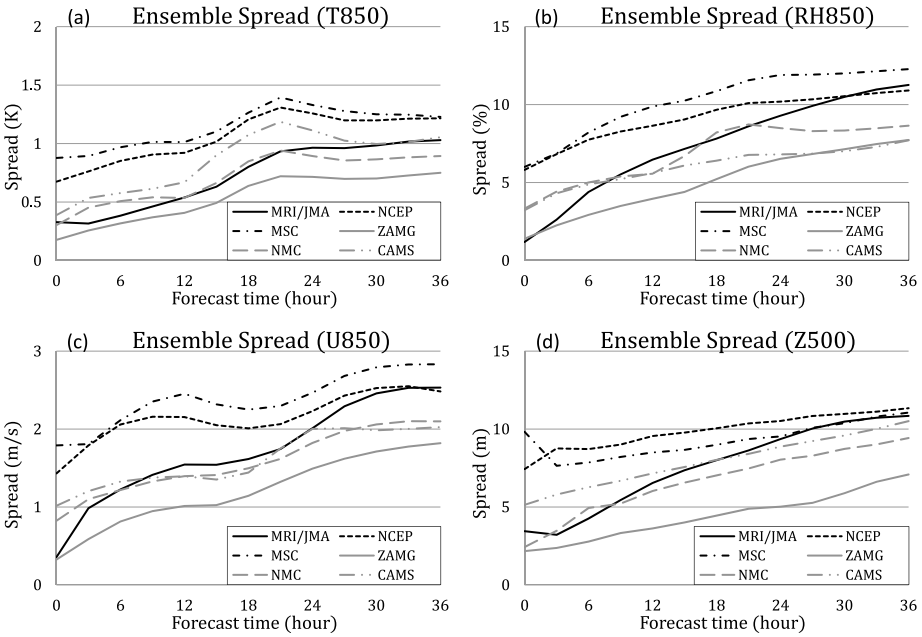

Fig. 7. Ensemble spreads of upper variables: (a) temperature (K) at 850 hPa, (b) relative humidity (%) at 850 hPa, (c) zonal wind (m s^{-1}) at 850 hPa and (d) geopotential height (m) at 500 hPa.

model of each participant is presented in Table 2. MSC applied the EnKF technique for data assimilation by their global model, and the others used a variational assimilation system, 3D-Var or 4D-Var. In the following intercomparison, we treated a forecast initialized by the ensemble mean at an initial time as the control forecast for MSC. NCEP employed three different forecast models with five ensemble members for ensemble prediction. We tentatively regarded a forecast by WRF-NMM from non-perturbed initial fields as the control run.

Forecast results in the common domain were interpolated to verification grids with a resolution of 0.15° and compared with 400 synoptic observation stations (Fig. 1). We verified surface parameters such as 6-h accumulated precipitation, surface temperature (T_{surf}) and relative humidity ($\mathrm{RH}_{\mathrm{surf}}$) by comparing the surface observation obtained from CMA and the mean value of four model grid points surrounding the observation point.

First, the verification results for surface precipitation were determined. Figure 2 indicates the threat scores for 6-h precipitation computed at 6-h intervals between 6 and 36 h. Figure 2a presents the threat scores against precipitation intensity averaged over the forecast period. For weak and moderate rains, the scores of MRI/JMA are substantially better than the others while those of MSC are the best for intense rains. The scores at each threshold are illustrated in Figs. 2b–d. The diurnal changes of the scores are not as obvious, except that of ZAMG for weak rains (Fig. 2b). For intense rains (Fig. 2d), the result of NCEP is considerably more degraded than their results against other thresholds.

Figure 3 presents the bias scores for 6-h precipitation. Figure 3a indicates the bias scores as a function of threshold averaged over the forecast period. The scores of MSC and MRI/JMA are comparatively flat for almost all thresholds. The forecast results of some participants overestimated weak rain frequency and underestimated intense rains. This characteristic is obvious in the results of NCEP and CAMS. The bias scores at each threshold are illustrated in Figs. 3b–d. The diurnal change of bias score is apparent in the control forecast of ZAMG, especially for weak rains (Fig. 3b). For intense rains, the diurnal change of NMC is relatively conspicuous (Fig. 3d). In addition, the forecast models of NCEP and CAMS tend to underestimate intense rainfall, even though their scores are nearly neutral for the first 6 h (Fig. 3d). These results suggest some inconsistencies between the ICs and forecast models in NCEP and CAMS systems. Generally, MRI/JMA's scores are comparatively flat and close to one during the forecast period.

The verification results of surface temperature are indicated in Fig. 4. Figure 4a presents the mean error (ME), and Fig. 4b presents the root mean square error (RMSE). The results of MRI/JMA and NMC are relatively better than those of the others, whereas diurnal changes are obvious in the forecast results of CAMS and MSC. The RMSE of temperature for CAMS is larger in daytime, in contrast with that for MSC (Fig. 4b). With regard to the threat and bias scores[1] at 1400 local time (corresponding to FT = 18, Fig. 4c and d), scores of MRI/JMA are the best of all. The forecast models of CAMS and MSC tend

[1]Frequency bias is used in this paper.

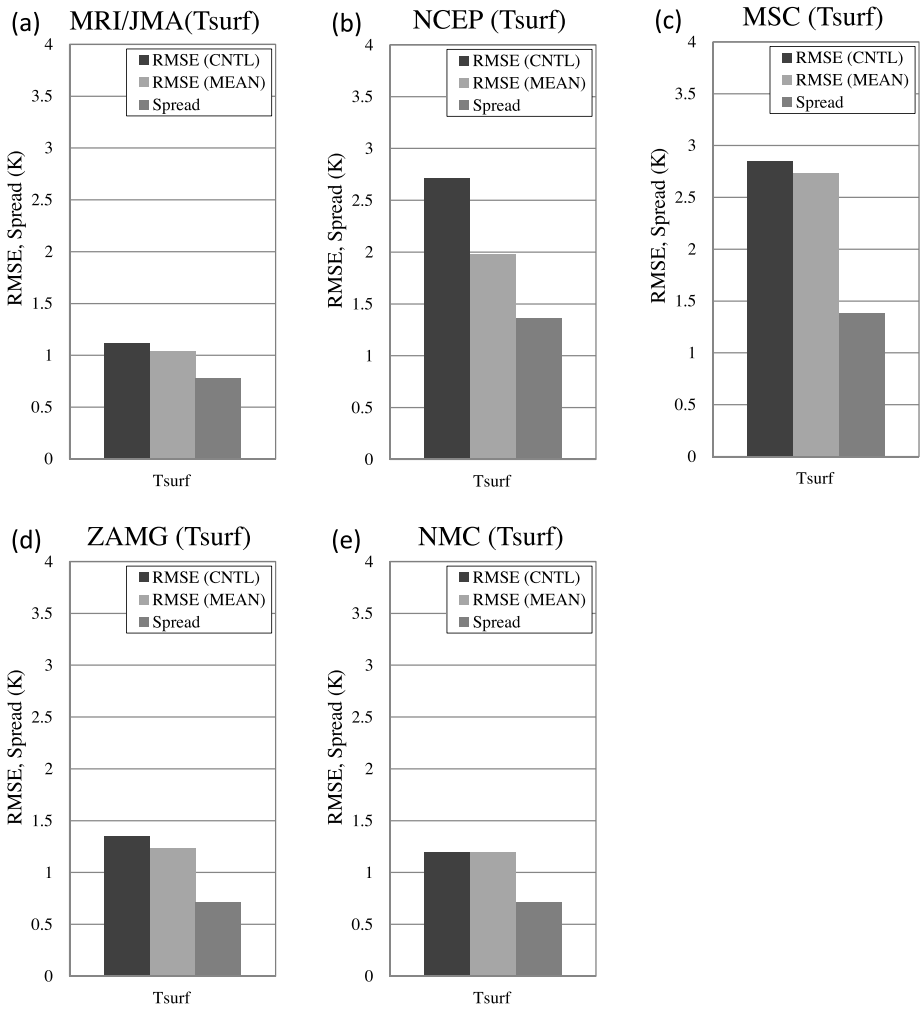

Fig. 8. Comparisons of the RMSEs of control forecast, ensemble mean and ensemble spread for surface temperature (T_{surf}). (a) MRI/JMA, (b) NCEP, (c) MSC, (d) ZAMG and (e) NMC. T2m data were not included in the grib2 data of CAMS. For MSC, member 21 was assumed as the control forecast.

to overestimate the surface temperature at the larger thresholds. These results of MRI/JMA can be related to the effective performance of the tunings of model parameters for physical processes (surface wetness), which were conducted in order to ameliorate the underestimations of convective rains and maximum temperatures on abnormally hot days found in the 2007 preliminary experiment (Saito et al., 2008). Implementation of the Meso 4D-Var analysis (Kunii et al., 2010) was also necessary to improve the performance of the control run.

The verifications for the relative humidity at the surface level are presented in Fig. 5. The results of MRI/JMA exhibit weak dry bias, whereas those of ZAMG and CAMS exhibit wet bias during the forecast period. For the RMSE, the results of MRI/JMA and NCEP are relatively better than those of the others.

4. Verification of ensemble forecast

This section presents the verification results of the ensemble predictions. This verification was performed at MRI/JMA. The specifications of the Tier-1 EPS of each participant are listed in Table 3. The number of ensemble members varies from 9 for CAMS to 20 for MSC. All participants applied lateral boundary perturbations using global EPS. While the physical perturbation method was not implemented in the MRI/JMA system, others adopted the multimodel or multiphysics method to represent the uncertainty in the forecast model.

The verification discussed in this subsection was performed for 25 July to 23 August 2008, and for the surface and upper variables in the common verification region. The verifications for this project including preliminary experiments were also performed in CMA.

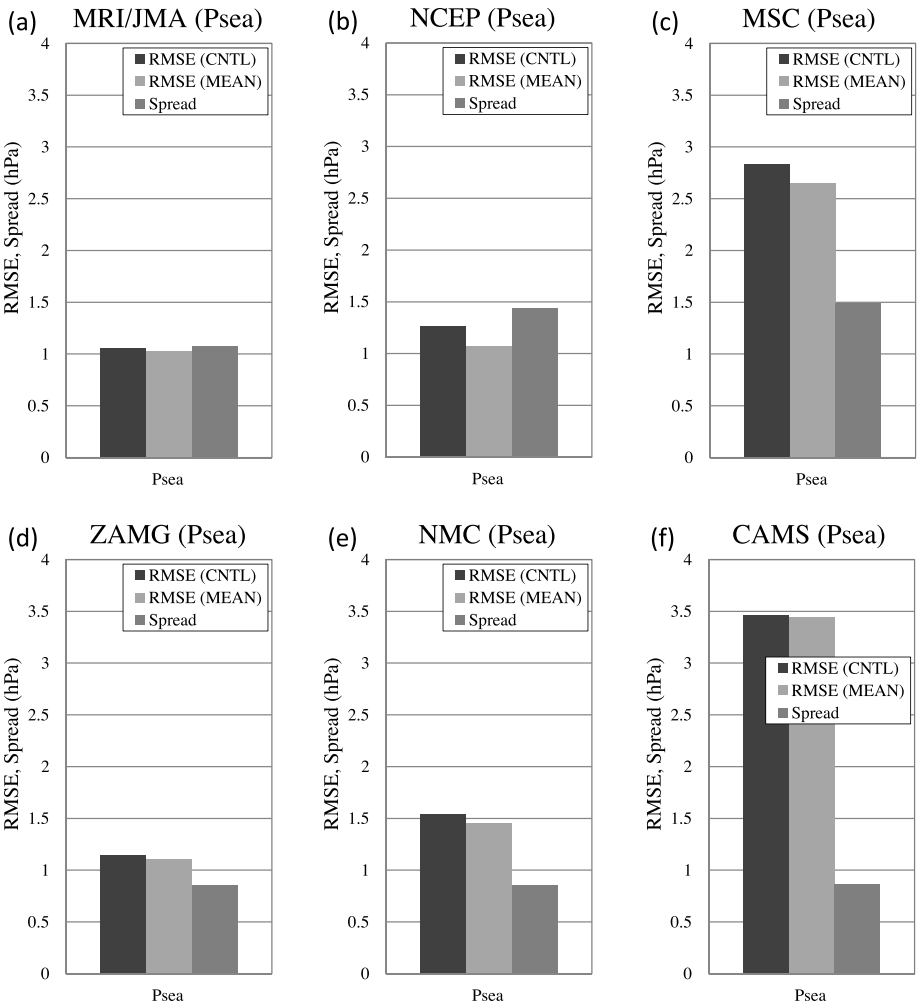

Fig. 9. Same as Fig. 8, but for sea surface pressure. The result of CAMS (f) is included.

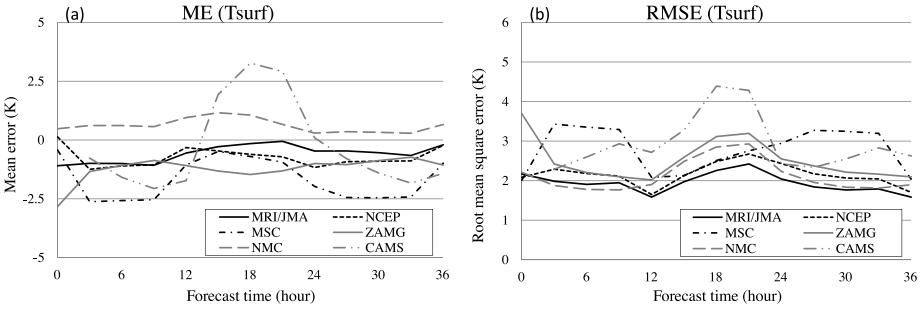

Fig. 10. (a) Same as Fig. 4a, but for ensemble mean. (b) The same as Fig. 4b, but for ensemble mean.

4.1. Evolution of ensemble spreads

Figure 6 depicts the ensemble spreads of wind [U, V (m s^{-1})], temperature [T (K)], sea-level pressure [P_{sea} (hPa)], relative humidity [RH (%)] at surface level and 3-h accumulated precipitation [RR3H (mm)]. For all participants, the steady growth of ensemble spreads through the forecast period. MRI/JMA, MSC, ZAMG and CAMS ensembles have maxima in spreads of temperature and relative humidity at FT = 18 (1400 local time), whereas NCEP and NMC have another peak at FT = 06 to 09 (0200 to 0500 local time). The peak at FT = 18 was caused by the diurnal cycle. The first peak of NCEP may be due to the use of two (global and mesoscale) models in their ensemble system; the perturbation growth patterns varied in the different models.

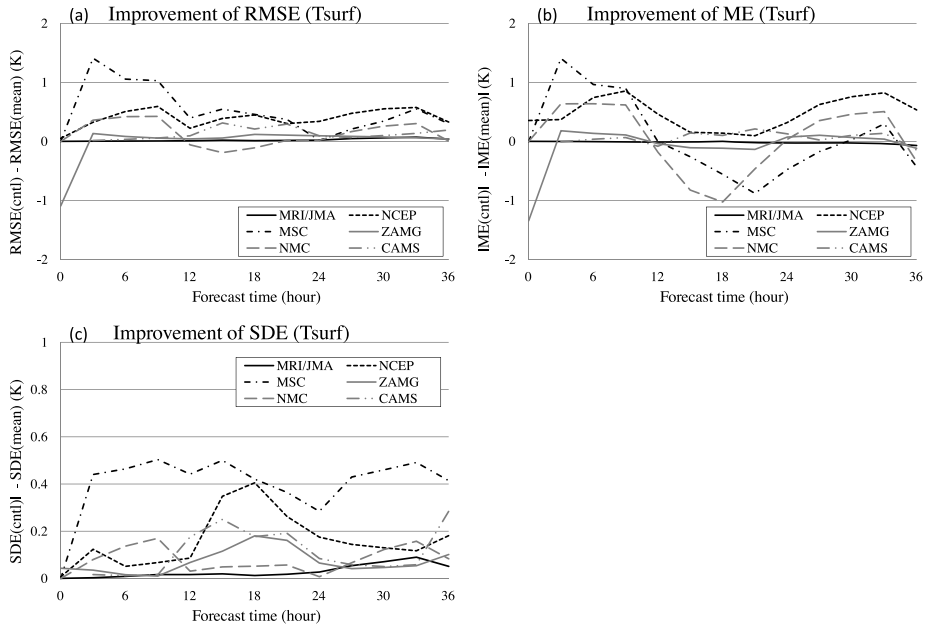

Fig. 11. Improvement of (a) RMSE, (b) ME and (c) SDE due to the usage of ensemble averages for T_{surf}.

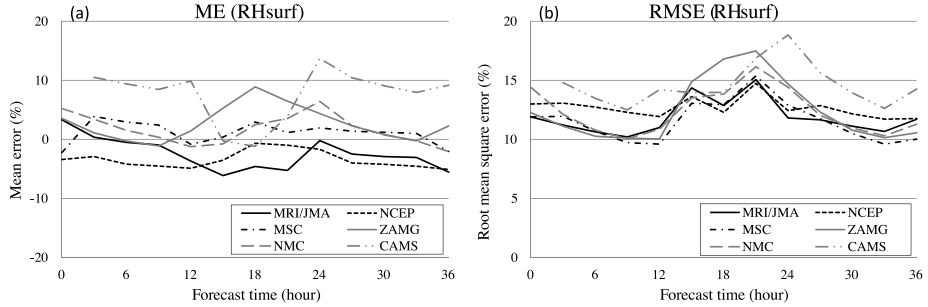

Fig. 12. Same as Fig. 10, but for $\mathrm{RH}_{\mathrm{surf}}$.

NMC's first peak seemed to be associated with the initial imbalance because of the large amplitude of initial perturbations.[2] This early imbalance can also be seen clearly in the spread of wind components of MSC and relative humidity of ZAMG. The most notable characteristic in Fig. 6 is that the spread of precipitation of the MSC system grows most rapidly in the first 6 h. That spread at FT = 06 is more than three times as large as the second largest one (NCEP). This result suggests that initial perturbations of moist fields had large amplitudes, or some ensemble members predicted intense precipitation because of their physical perturbations.

The ensemble spreads for upper variables are presented in Fig. 7. The spread of these variables evolves differently from the surface variables because they are not directly influenced by the diurnal change. For almost all variables, the spreads of MSC and NCEP are larger than those of the others. As with the growth rate of ensemble spreads, those of MRI/JMA have the steepest gradient, partly because of MRI/JMA's utilization of singular vectors for both initial and boundary perturbations, which have characteristics that maximize the growth rate of forecast perturbations.

4.2. *Forecast errors of ensemble mean*

To investigate the advantage of the ensemble forecast against a single deterministic forecast, we carried out forecast verifications comparing the forecast error of the individual control forecast with that of the ensemble mean. Moreover, forecast errors of the ensemble mean and ensemble spreads were compared to assess the validity of the magnitude of the ensemble spread. In this study, the 24-h forecast by each EPS was verified by evaluating the RMSE against each participant's analysis (the IC of the control run of each system) at the same valid time. Precipita-

[2]Since the transferred data by NMC and CAMS did not include surface variables at initial times, the ensemble spreads of those centers were not drawn at FT = 0.

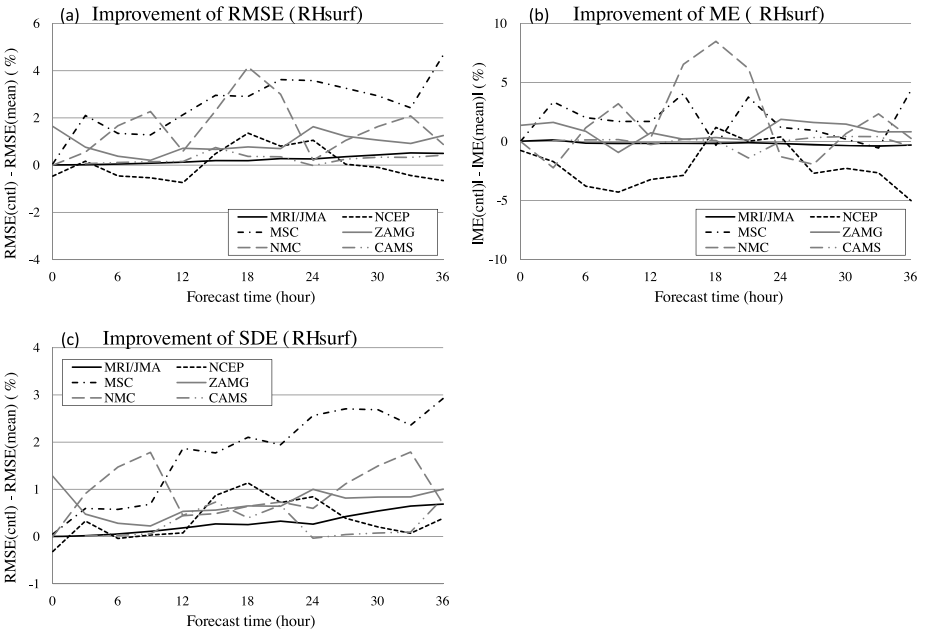

Fig. 13. Same as Fig. 11, but for RH_{surf}.

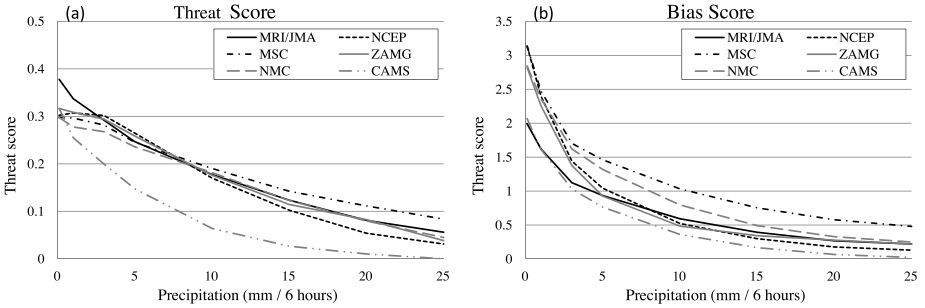

Fig. 14. (a) Same as Fig. 2a, but for ensemble mean. (b) Same as Fig. 3a, but for ensemble mean.

tion forecasts were verified against the rain-gauge observations. Figure 8 presents forecast verifications for T_{surf} at FT = 24.[3] The ensemble mean provides a better forecast than the control one, and the improvement of NCEP is the most obvious. However, for all participants, ensemble spreads are somewhat smaller than corresponding forecast errors of the ensemble mean. The RSMEs of MRI/JMA are relatively small because the incremental forecast after the Meso 4D-Var was employed using NHM, which is the same as the forecast model, and the errors caused from the difference between analysis and forecast model are reduced.

The same comparison for the sea surface pressure is presented in Fig. 9. With the exception of MRI/JMA and NCEP, the ensemble spreads are substantially smaller than the forecast errors. The RMSEs of MSC and CAMS are considerably larger than their ensemble spreads. For MSC, this result is probably due to the difference between the analysis model and the forecast model. These characteristics can be seen in verifications of other surface elements, such as relative humidity and wind components (not shown).

For T_{surf} and RH_{surf}, more detailed verifications against observations are performed. Figure 10 depicts the time series of the RMSE and the ME of the ensemble averages for T_{surf}. The CAMS system has a strong diurnal cycle even when the ensemble average is used (Fig. 10a), apparently due to the tendency of its forecast model to overestimate T_{surf} in the daytime. The most distinctive feature is that, compared with the control forecast (Fig. 4a), the ensemble mean reduces the variation of ME originating from its average, but it does not always reduce the mean bias itself. This trend is relatively obvious in the results of MSC and NMC. As with the RMSE, NCEP and MSC ensemble systems clearly improve the scores; however, with the others we cannot identify any difference.

[3]The ensemble forecast data of CAMS did not include the surface temperature and relative humidity at the initial time.

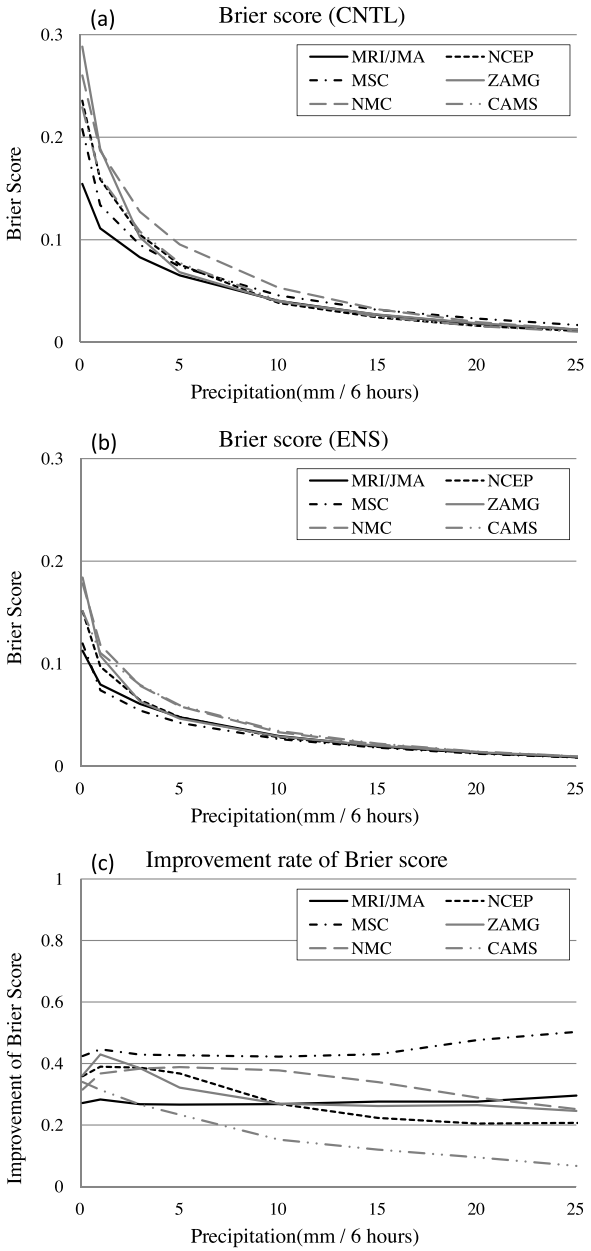

Fig. 15. Brier score of 6-h accumulated precipitation for (a) control forecasts, (b) ensemble forecasts and (c) the improvement rate of the Brier score defined in eq. (4).

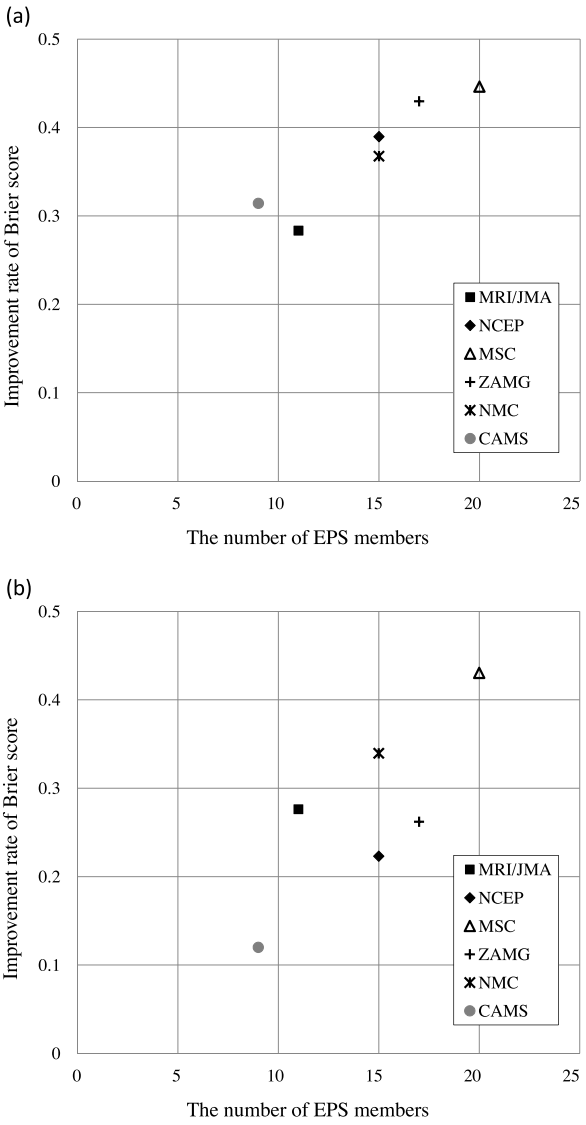

Fig. 16. Relationship between the improvement rate of the Brier score (6-h precipitation) and the number of EPS members. The thresholds are (a) 1 mm per 6 h, and (b) 15 mm per 6 h.

For the quantitative evaluation, the improvement (ensemble average minus individual control forecast) of the RMSE, the ME and the square of the standard deviation of the error (SDE) due to the usage of ensemble averages is plotted in Fig. 11. The RMSE can be decomposed into the ME and the SDE (Murphy, 1988; Hou et al., 2001).

$$\mathrm{RMSE}^2 = \mathrm{ME}^2 + \mathrm{SDE}^2, \tag{2}$$

$$\mathrm{SDE}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - a_i - \mathrm{ME})^2\,, \tag{3}$$

where N is the number of samples, x is a model forecast and a is an observation. These figures indicate that the ensemble systems of NCEP and MSC improve the RMSE in this case, and for MSC this improvement is attributed to the large improvement of SDE. The SDE can be interpreted as the standard deviation of random errors. The usage of multimodels or multiphysics influenced the surface perturbation and thus effectively contributed to improve the score. However, the ensemble systems of MRI/JMA and CAMS improved little in the RMSEs reduction, probably due to

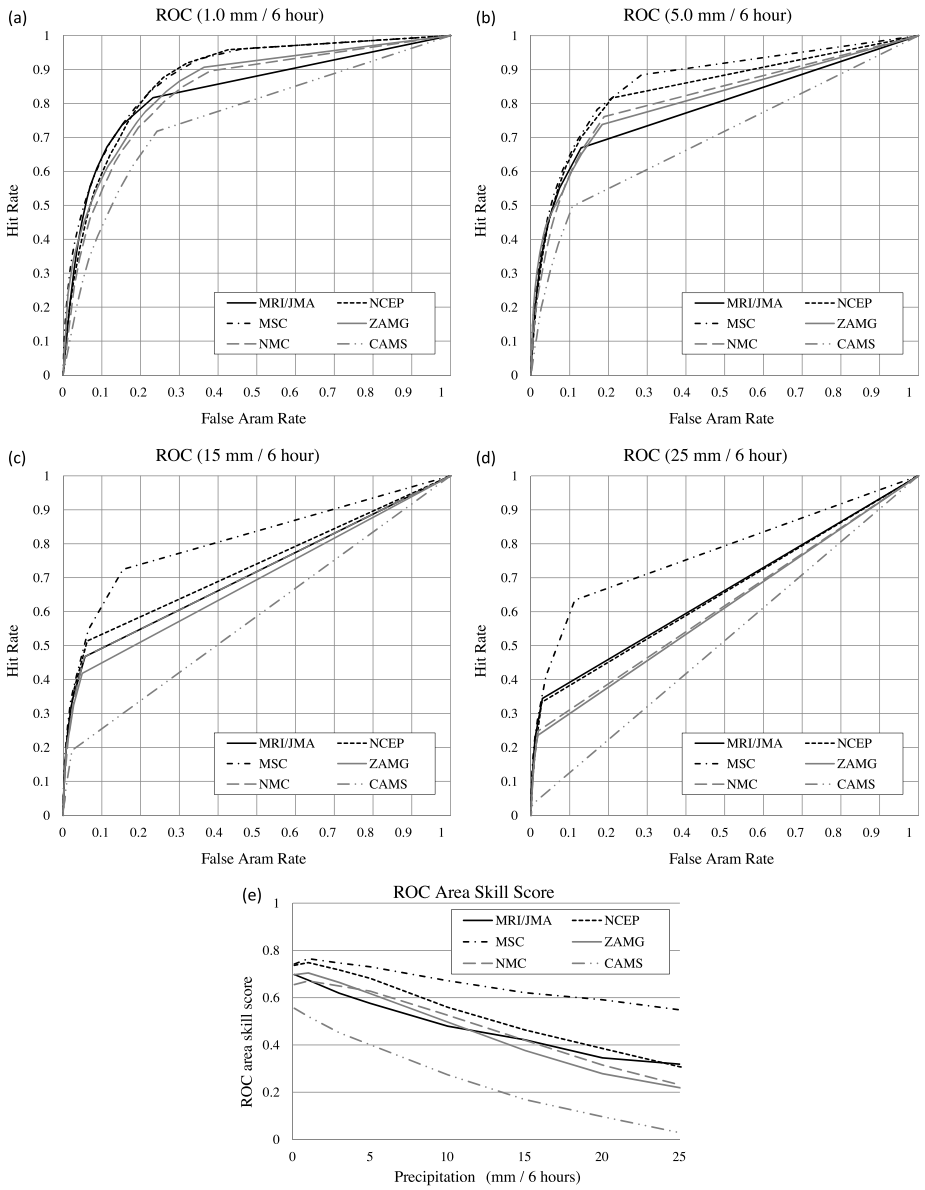

Fig. 17. ROC diagrams for 6-h accumulated precipitation, the thresholds of which are (a) 1 mm, (b) 5 mm, (c) 10 mm and (d) 25 mm. (e) ROCASS.

the lack of physical perturbation and the insufficiency of initial perturbations in the surface processes. In Fig. 11b, the ME of the ensemble average was not always reduced, compared with the individual control forecast. This result suggests the difficulty of introducing and tuning physical perturbation methods.

The same figures are presented for RH_{surf} (Figs 12 and 13). Although the bias of NCEP in the ME is nearly neutral in the control verification (Fig. 5a), the bias becomes negative in the ensemble verification. Since NCEP used three different models for ensemble forecasting, this negative bias was probably due to the use of a different model than the one used for the control forecast. In contrast, the ensemble average of MSC and NMC improved the negative bias observed in daytime for the control forecast, subsequently leading to further improvement of the RMSE for NMC, whereas the contribution of SDE is larger for MSC.

4.3. Verification scores for the precipitation

Next we focus on comparing between the control forecast and ensemble mean with the verification scores for 6-h accumulated precipitation. Figure 14 depicts the threat and bias scores for the precipitation forecasts of ensemble mean against observations averaged over the forecast period. For weak and moderate rains, ensemble means of some participants improved the threat scores (Fig. 14a), while the scores of ensemble mean become worse for

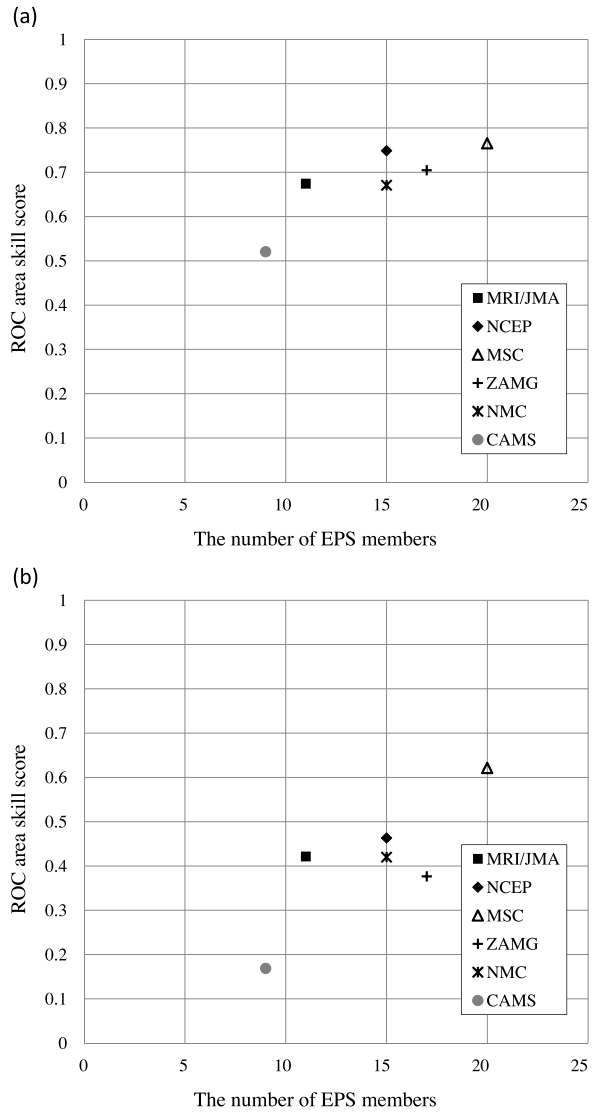

Fig. 18. Relationship between the ROCASS (6-h precipitation) and the number of EPS members. The thresholds are (a) 1 mm per 6 h and (b) 15 mm per 6 h.

intense rains, compared with those of control forecasts (Fig. 2a). This result is not surprising since the forecast fields are smoothed by averaging, and therefore the appearance of heavy rain in those fields decreases. As with the bias score (Fig. 14b), weak rains are overestimated, and intense rains (the border of which is 5 mm per 6 h) are underestimated.

We now introduce the probabilistic verification scores, such as the Brier score (Brier, 1950) and the relative operating characteristic (ROC) diagram (Harvey et al., 1992). The Brier score is defined as the mean squared difference between forecast probabilities and corresponding observational data. If the accuracy of the probabilistic forecast increases, the score approaches zero. Figure 15 presents the Brier scores for individual and ensemble forecasts averaged over all the forecast times. Furthermore, the improvement rate of the Brier score defined by

$$\text{improvement rate} = \frac{\text{Brier score}_{(\text{CNTL})} - \text{Brier score}_{(\text{ENSEMBLE})}}{\text{Brier score}_{(\text{CNTL})}}, \quad (4)$$

is illustrated. This definition is similar to that of the Brier skill score, except that it is based on the Brier score of a control forecast, not a climatological forecast. Ensemble forecasts clearly improve Brier scores, compared with the individual control forecast. The improvement rates of MSC and NMC are relatively higher than those of other systems. For intense rains, the improvement rate of the CAMS system is rather small, perhaps because the control forecast of CAMS tends to underestimate intense rainfall (Fig. 3a). Although CAMS used multiphysics in their EPS system, it did not seem able to improve the bias of their forecast model. The considerably small ensemble spread of precipitation amount (Fig. 6f) also caused insignificant improvement of the score even by using ensemble forecasting. The rate of MRI/JMA is almost constant even for intense rains, which is similar to the result of MSC, whereas the improvement for weak rains is relatively poor. One possible reason is that MRI/JMA's system had the best threat and bias score the individual control forecast, especially for weak to moderate rains (Figs 2 and 3); therefore, improving the score further using ensemble forecasting might be comparatively difficult.

To examine the relationship between the improvement rate of the Brier score and the number of EPS members, the improvement rates of individual systems are plotted as a function of the number of EPS members with different thresholds (Fig. 16). A strong positive relationship exists between improvement due to ensemble forecasting and the number of members. For weak rains, the improvement of MRI/JMA is slightly less than those of the others for the number of members (Fig. 16a). This result is probably due to the fact that the Brier score of MRI/JMA's control forecast is relatively better than those of the others; therefore, it is not easy to improve the score by using ensemble forecasting. The MSC ensemble indicates the superiority of its probabilistic forecast for intense rains (Fig. 16b).

ROC diagrams and ROC area skill score (ROCASS) averaged over all the forecast times are presented in Fig. 17. As in Richardson (2000), ROCASS is defined by the area enclosed by the curve and x–y axis (ROC Area, 'ROCA'):

$$\text{ROCASS} \equiv 2\,(\text{ROCA} - 0.5)\,(-1 \leq \text{ROCASS} \leq 1), \quad (5)$$

where ROCASS equals 1 if the probabilistic forecast is perfect. For weak and moderate rains, the MRI/JMA system has a poor rate of detecting precipitation (Figs. 17a–c), suggesting the importance of the number of EPS members and physical perturbations. However, MRI/JMA's detection rate is not degraded as much as that of the others, even for intense rains (Fig. 17d). The NCEP system has a higher detection rate, especially for weak rains, perhaps due to the inclusion of global model forecasts in its ensemble members. The MSC system

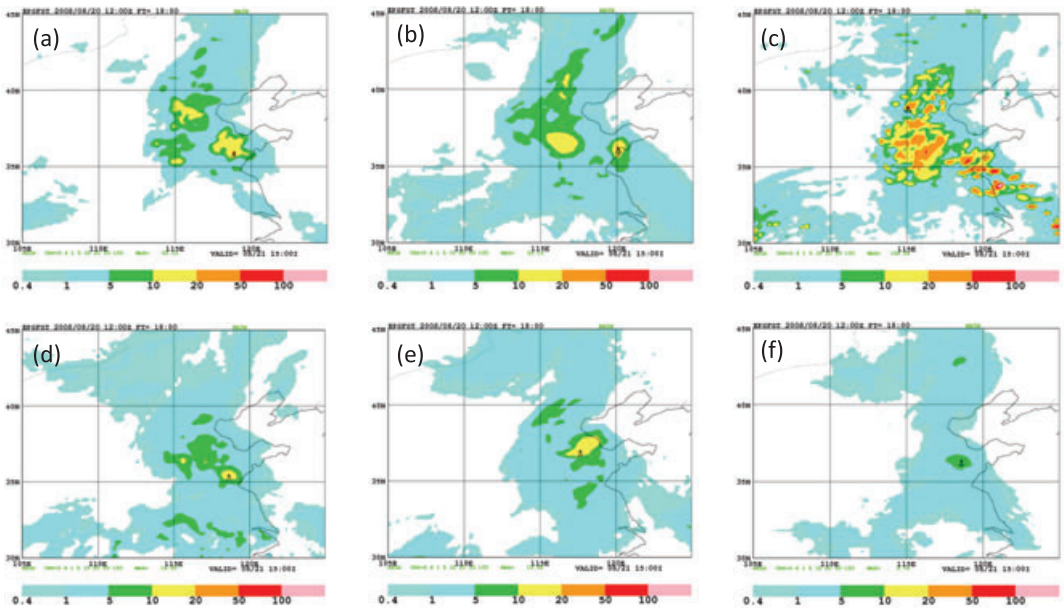

Fig. 19. Ensemble spread of 3-h accumulated precipitation. Initial time is 1200 UTC on 20 August 2008 (FT = 18). (a) MRI/JMA, (b) NCEP, (c) MSC, (d) ZAMG, (e) NMC and (f) CAMS.

exhibits higher performance of the probabilistic forecast of precipitation for all thresholds (Fig. 17e). The MSC's MEP system is the only system based on the use of Ensemble Data Assimilation (EDA). The EDA could have a positive impact on probabilistic scores because it allows better account analysis (or initial) uncertainty. As mentioned later, physical perturbation methods also seem to be useful for improving ensemble forecasting.

These features can also be seen in Fig. 18. The MRI/JMA and NCEP systems have higher detection rates for their ensemble members. However, the ROCASS of CAMS and that of ZAMG seem to be smaller than expected in regard to the number of ensemble members, especially for intense rains (Fig. 18b). As for CAMS, its control forecast has a tendency to underestimate intense precipitation (Fig. 3a), and this problem is not resolved even for ensemble prediction.

Not only verification scores but also actual fields must be investigated. Figure 19 illustrates an example of distributions of the ensemble spread of 3-h precipitation for a heavy rainfall that occurred during the experiment period. A glance at these figures will reveal that the spread of MSC is by far the largest of all, and its horizontal scale is unnaturally small for the resolution of the forecast model. On the other hand, the CAMS system has considerably smaller spreads than the others. This can be seen from the statistical point of view (Fig. 6f). The area with a precipitation spread of over 0.4 mm per 3 h is somewhat small for MRI/JMA, which may be a manifestation of the insufficient detection rate for weak rains (Fig. 17a).

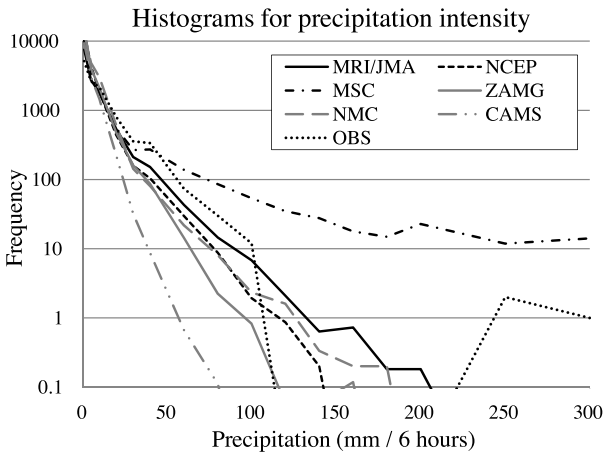

Fig. 20. Histograms for 6-h precipitation intensity. The scale of the vertical axis is logarithmic.

In order to examine the characteristics of individual precipitation forecast, we plotted the frequency distribution of 6-h accumulated precipitation averaged by the number of EPS members for all participants and observations in Fig. 20. Note that the scale of the vertical axis is logarithmic. The frequencies of MRI/JMA, NCEP and NMC are relatively close to the observations, while the frequency of very intense rains predicted by the MSC system is higher than the observations. This result indicates that some of the MSC members tend to overestimate intense rainfalls even when such severe rainfalls are not observed, which is consistent with MSC's precipitation spread in Fig. 6.

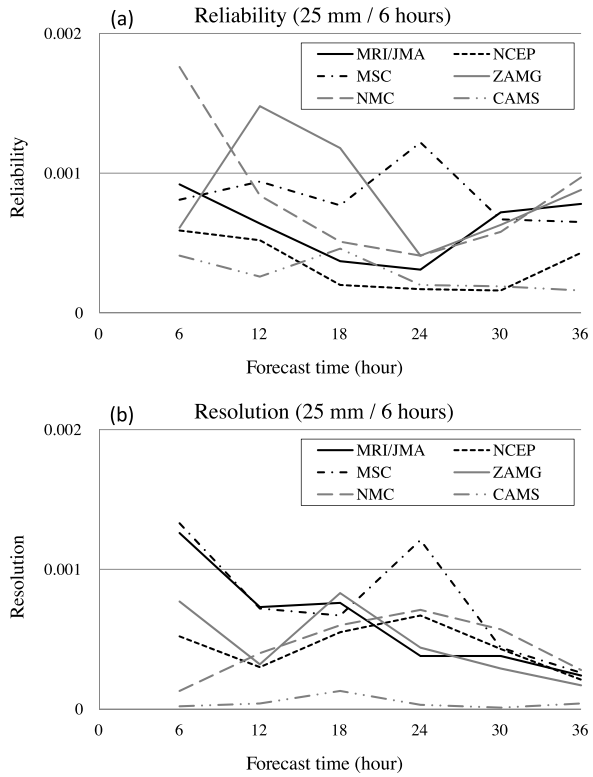

Fig. 21. Time series of (a) reliability and (b) resolution term of the Brier score for the precipitation of 25 mm per 6 h.

5. Discussion

A major concern regarding the verification results discussed earlier is that the MEP system of MSC exhibits the best probabilistic precipitation forecast (Figs. 15–18) although the frequency of very intense precipitation is much higher than the observations (Figs. 19 and 20). For further clarification of the statistical performance of the probabilistic precipitation forecasts, we first investigated the reliability and the resolution of the Brier score. Murphy (1973) demonstrated that the Brier score (BS) can be decomposed into the sum of three components:

$$\begin{aligned} \mathrm{BS} &= \mathrm{BS_{rel}} - \mathrm{BS_{res}} + \mathrm{BS_{unc}} \\ \mathrm{BS_{rel}} &= \sum_{i=1}^{N_p} \left(p_i - \frac{M_i}{N_i} \right)^2 \frac{N_i}{N} \\ \mathrm{BS_{res}} &= \sum_{\mathrm{i=1}}^{\mathrm{N_p}} \left(\frac{M}{N} - \frac{M_i}{N_i} \right)^2 \frac{N_i}{N} \\ \mathrm{BS_{unc}} &= \frac{M}{N} \left(1 - \frac{M}{N} \right), \end{aligned} \tag{6}$$

where N_p is the number of divided bins depending on forecast probability (p_i), N_i and M_i are the numbers of samples of predicted or observed frequency within the ith bin, and N and M are the summation of N_i and M_i. The reliability term ($\mathrm{BS_{rel}}$) represents the difference between forecast probability and observed frequency in each category. A lower score indicates a better probabilistic forecast. The resolution term ($\mathrm{BS_{res}}$) measures the difference between the mean frequency observed in all samples and the observed frequency in specific bins. A minimum resolution term (0) indicates that a probabilistic forecast can be regarded as a climatological one. The third term is the uncertainty ($\mathrm{BS_{unc}}$), which implies the inherent uncertainty of the event. Uncertainty is independent of the forecast system because it is a function of only the observations.

Figure 21 presents the time series of the reliability and resolution of the Brier score for each of the ensemble systems for intense precipitation. NCEP and CAMS had good reliability, and MRI/JMA and MSC had relatively good resolution. The most notable characteristic in the figure is that MSC had relatively good resolution, whereas its reliability was not as good. This suggests that for intense rainfall amounts, the MEP system of MSC was far from a climatological forecast, but its probabilistic forecast was not consistent with the observation frequency. The MSC system's improved Brier score (Fig. 15c) was attributable to the poor reliability of its control forecast (not shown).

The ROC score is calculated using the hit rate (Hr) and the false alarm rate (Fr) defined as:

$$\mathrm{Hr} = \frac{\mathrm{FO}}{\mathrm{FO} + \mathrm{XO}} \tag{7}$$

$$\mathrm{Fr} = \frac{\mathrm{FX}}{\mathrm{FX} + \mathrm{XX}}, \tag{8}$$

where F and O indicate that predicted or observed precipitation exceeding a certain threshold, and X indicates that no precipitation is predicted or observed at the threshold.

Figure 17 indicates that the ROCASS for rare events is mostly dominated by the point that has the highest hit rate and false alarm rate in the ROC diagram. Table 4 indicates the frequency of the forecast and the observation for the thresholds of the reciprocal of the number of EPS members. For intense precipitation, the MSC's MEP system has the highest ROCASS value, although the false alarm (FX) is considerably higher than those of the others. This is because XX is large enough for rare cases so that the influence of FX on the false alarm rate could be small. For example, since FX + XX is 340 on average in this verification of precipitation over 25 mm per 6 h, the false alarm rate is no more than 0.2 even if 20% of XX is shifted to FX. On the other hand, FO + XO is just 3.5 for the threshold of 25 mm per 6 h; therefore, the transition of XO to FX considerably affects the hit rate under these circumstances. Thus, we can infer that the ROCASS for intense precipitation might be improved if one member of the EPS tends to overestimate the precipitation, which is expected to increase FO in spite of an increased FX.

In order to verify the influence of FX and FO on ROCASS, we performed a sensitivity experiment using MRI/JMA's ensemble forecast results. In this experiment (MRI/JMA_C10), the precipitation of the control forecast was uniformly increased to 10

Table 4. Frequency of the prediction and observation for precipitation

	Precipitation (5 mm per 6 h)					Precipitation (25 mm per 6 h)				
	FO	FX	XO	XX	ROCASS	FO	FX	XO	XX	ROCASS
MRI/JMA	13.6	42.5	6.76	286.	0.58	1.17	10.1	2.27	336.	0.32
NCEP	16.7	69.1	3.71	260.	0.68	1.13	9.44	2.29	336.	0.31
MSC	18.0	93.9	2.34	235.	0.73	2.18	39.2	1.24	306.	0.55
ZAMG	15.2	60.7	5.18	268.	0.62	0.82	6.11	2.61	340.	0.22
NMC	15.6	62.2	4.81	267.	0.63	0.84	6.54	2.58	339.	0.23
CAMS	10.1	35.1	10.3	294.	0.40	0.11	1.47	3.31	344.	0.03
MRI/JMA_C10	15.6	56.2	4.78	273.	0.65	2.07	29.6	1.36	316.	0.54

Note: The threshold probability is the reciprocal of the number of EPS members.

times the original value. FO became twice as large as that of the original experiment and FX became four times as large as that of the original experiment, thus helping to improve the ROCASS of MRI/JMA. This result suggests that the overestimation of local rare events by including a small number of EPS members can contribute to improving ROCASS, which might not be suitable for verifying local severe weather.

6. Summary and conclusion

The WWRP B08RDP, an international research project of the WWRP of the World Meteorological Organization for short range ensemble forecasting, was conducted in conjunction with the Beijing 2008 Olympic Games. The objectives of the project were to improve understanding of the high-resolution and short-range probabilistic prediction processes through numerical experimentation, and to share experiences in the development of the real-time MEP system. Six international participants (MRI/JMA, NCEP, MSC, ZAMG, NMC and CAMS) conducted the intercomparison for 1 month, from 24 July to 24 August 2008. Ensemble forecasts were carried out using advanced regional forecast models with a horizontal resolution of 15 km. Various initial perturbation methods (e.g. the singular vector method, the breeding method and the EnKF technique) were utilized in the MEP systems. In addition to the initial perturbations, lateral boundary perturbations for regional forecast models and physical perturbations accounting for model uncertainty were considered. Experience gained in this project will contribute greatly to the development of MEP systems in the future.

It is crucial to understand the characteristics of the MEP systems through objective verification. In this study, we performed verification and intercomparison of EPSs in the B08RDP experiment. First, we investigated the participants' performances of non-perturbed control forecasts. Second, we used verification measures to assess the accuracy of probabilistic forecasts. The results can be summarized as follows:

(1) For all systems, the ensemble spreads grew as the forecast time increased, and the ensemble mean reduced the forecast errors compared with individual control forecast in the verification against the analysis fields.

(2) For almost all systems, the ensemble spreads were somewhat smaller than forecast errors of the ensemble mean (weak underdispersive).

(3) Ensemble forecasts clearly improved Brier scores, compared with the individual control forecast. For intense rains, the improvements of the NCEP and CAMS systems were rather small, as a result of the underestimation of intense rains.

(4) The following problems with individual ensemble systems were identified:

(i) In the ROC curves, the hit rates of precipitation for moderate rains of the MRI system were insufficient, probably due to the small number of ensemble members and the lack of physical perturbations.

(ii) The frequency of very intense rains predicted by the MSC system was higher than the observations.

(iii) For surface conditions (T and RH at 2 m), the MRI (and CAMS) ensemble system exhibited little improvement in RMSE reduction, probably due to the lack of physical perturbation and insufficiency of initial perturbations in the surface processes.

(iv) The ME of the ensemble forecast was sometimes worse than that of the individual control forecast, indicating that careful attention should be paid to physical perturbations.

(5) The overestimation of localized rare events by the small number of EPS members can contribute to improving ROCASS. This result suggests that the ROCASS is not always suitable for local severe weather.

(6) As a whole, concerning the control forecast, the performance of the MRI system was the best, whereas remarkable improvement using ensemble forecast was seen in the MSC system. The improvement of the MSC ensemble system seems to be due to the fact that the MSC system had a large number of EPS members and large ensemble spreads, and that the initial perturbation and physics perturbation functioned effectively in the ensemble forecast.

From the B08RDP experiment, we conclude that the MEP systems constructed here have the potential to provide probabilistic information to prevent natural hazards occurring as a result of mesoscale weather events. The improvement of objective verification methods for the MEP system is an issue for the future.

7. Acknowledgments

The authors are grateful to two anonymous referees for their valuable comments, which significantly improved the manuscript. We also thank the staffs of CMA and the Beijing Meteorological Bureau (BMB) for their efforts in conducting the WWRP Beijing 2008 FDP/RDP project. Numerical computations in this study were performed using an NEC SX-6 super-computer system at MRI.

References

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

Buizza, R., Tribbia, J., Molteni, F. and Palmer, T. N., 1993. Computation of optimal unstable structures for a numerical weather prediction model. *Tellus*, **45A**, 388–407.

Casati, B., Wilson, L. J., Stephenson, D. B., Nurmi, P., Ghelli, A. and co-authors, 2008. Forecast verification: current status and future directions. *Meteor. Appl.*, **15**, 3–18.

Charron, M., Pellerin, G., Spacek, L., Houtekamer, P. L., Gagnon, N. and co-authors. 2010. Toward random sampling of model error in the Canadian ensemble prediction system. *Mon. Wea. Rev.*, **138**, 1877–1901.

Chen D H, Xue, J. S. and Yang, X. S., 2008. New generation of multiscale NWP system (GRAPES): general scientific design. *China Sci. Bull.*, **53**(22): 3433–3445

Côté, J., Gravel, S., Méthot, A., Patoine, A. Roch, M. and co-authors. 1998a. The operational CMC-MRB Global Environmental Multiscale (GEM) model, Part I: design considerations and formulation. *Mon. Wea. Rev.*, **126**, 1373–1395.

Côté, J., Desmarais, J.-G. Gravel, S. Méthot, A. Patoine, A. and co-authors. 1998b. The operational CMC-MRB Global Environmental Multiscale (GEM) model, part II: results. *Mon. Wea. Rev.*, **126**, 1397–1418.

Derková, M. and Bellus, M. 2007. Various applications of the blending by digital filter technique in the ALADIN numerical weather prediction system. *Meteorologický casopis*, **10**, 27–36. Available online at http://www.rc-lace.eu.

Du, J. 2004. Hybrid ensemble prediction system: a new ensembling approach. *Preprints, Symposium on the 50th Anniversary of Operational Numerical Weather Prediction*, University of Maryland, College Park, Maryland, June 14–17, 2004, Amer. Meteor. Soc., CD-ROM (paper p4.2, 5pp). Available online at http://www.emc.ncep.noaa.gov/mmb/SREF/reference.html.

Ebisuzaki, W. and Kalnay, E. 1991. Ensemble experiments with a new lagged average forecasting scheme. *WMO Research Activities in Atmospheric and Oceanic Modeling Rep.* **15**, Geneva, Switzerland, 6.31–6.32.

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99**(C5), 10 143–10 162.

Harvey, L. O., Hammond, K. R., Lusk, C. M. and Mross, E. F. 1992. The application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev.*, **120**, 863–883.

Hou, D., Kalnay, E. and Drogemeier, K. 2001. Objective verification of the SAMEX98 ensemble forecasts. *Mon. Wea. Rev.*, **129**, 73–91.

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

Keenan, T., Joe, P., Wilson, J., Collier, C., Golding, B. and co-authors. 2003. The Sydney 2000 World Weather Research Programme Forecast Demonstration Project. *Bull. A. Met. Soc.*, **84**, 1041–1054.

Kunii, M., Saito, K. and Seko, H. 2010. Mesoscale data assimilation experiment in the WWRP B08RDP. *SOLA*, **6**, 33–36.

Mullen, S. L., and Baumhefner, D. P. 1989. The impact of initial condition uncertainty on numerical simulations of large-scale explosive cyclogenesis. *Mon. Wea. Rev.*, **117**, 2800–2821.

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

Murphy, A. H., 1988. Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.*, **116**, 2417–2424.

Richardson, D. S., 2000. Skill and relative economic value of the ECMWF ensemble prediction system. *Q. J. R. Meteor. Soc.*, **126**, 649–667.

Saito, K., Fujita, T., Yamada, Y., Ishida, J., Kumagai, Y. and co-authors. 2006. The operational JMA nonhydrostatic mesoscale model. *Mon. Wea. Rev.*, **134**, 1266–1298.

Saito, K., Ishida, J., Aranami, K., Hara, T., Segawa, T. and co-authors. 2007. Nonhydrostatic atmospheric models and operational development at JMA. *J. Meteor. Soc. Japan*, **85B**, 271–304.

Saito, K., Seko, H., Kunii, M. and Hara, M. 2008. Mesoscale ensemble prediction experiment for WWRP Beijing Olympic 2008 RDP: 2007 preliminary experiment. *CAS/JSC WGNE Research Activities in Atmospheric and Oceanic Modelling*, **37**, 1.23–1.24.

Saito, K., Kunii, M., Hara, M., Seko, H., Hara, T. and co-authors. 2010. WWRP Beijing Olympics 2008 Forecast Demonstration/Research and Development Project (B08FDP/RDP). *Tech. Rep. MRI*, **62**, 1–146. Available online at http://www.mri-jma.go.jp/Publish/Technical/DATA/VOL_62/62_en.html.

Saito, K., Hara, M., Kunii, M., Seko, H. and Yamaguchi, M. 2011. Comparison of initial perturbation methods for the mesoscale ensemble prediction system of the Meteorological Research Institute for the WWRP Beijing 2008 Olympics Research and Development Project (B08RDP). *Tellus*, **63A**.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M. and co-authors. 2005. A description of the Advanced Research WRF, Version 2. NCAR Technical Note.

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

Toth, Z., and Kalnay, E. 1997. Ensemble forecasting at NCEP: the breeding method. *Mon. Wea. Rev.*, **125**, 3297–3318.

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

Yamaguchi, M., Sakai, R., Kyoda, M. Komori, T. and Kadowaki, T. 2009. Typhoon ensemble prediction system developed at the Japan Meteorological Agency. *Mon. Wea. Rev.*, **137**, 2592–2604.

Yeh, K.-S., Côté, J., Gravel, S., Méthot, A., Patoine, A. and co-authors. 2002. The CMC-MRB global environmental multiscale (GEM) model, Part III: nonhydrostatic formulation. *Mon. Wea. Rev.*, **130**, 339–356.

Zhou, B. and Du, J. 2010. Fog prediction from a multimodel mesoscale ensemble prediction system. *Weather and Forecasting*, **25**, 302–322.