Evaluation of TIGGE ensemble predictions of Northern Hemisphere summer precipitation during 2008-2012

Xiang Su[1], Huiling Yuan[1,2], Yuejian Zhu[3], Yan Luo[3], Yuan Wang[1]

Corresponding author:

Prof. Huiling Yuan

Email: yuanhl@nju.edu.cn

(Last update on February 4, 2014)

Submit to *Journal of Geophysical Research: Atmospheres*.

[1] Key Laboratory of Mesoscale Severe Weather/Ministry of Education and School of Atmospheric Sciences, Nanjing University, Nanjing, China

[2] Jiangsu Collaborative Innovation Center for Climate Change, China

[3] Environmental Modeling Center/NCEP/NWS/NOAA, College Park, Maryland, USA

Abstract

The ensemble mean quantitative precipitation forecasts (QPFs) and probabilistic QPFs (PQPFs) from six operational global ensemble prediction systems (EPSs) in The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE) dataset are evaluated against the Tropical Rainfall Measuring Mission (TRMM) observations using a series of area-weighted verification metrics during June to August 2008-2012 in the Northern Hemisphere (NH) midlatitudes and tropics. Results indicate that generally the European Centre for Medium-Range Weather Forecasts (ECMWF) performs best while the Canadian Meteorological Centre (CMC) is relatively good for short-range QPFs and PQPFs at light precipitation thresholds. The overall forecast skill is better in the NH midlatitudes than that in the NH tropics. QPFs and PQPFs from China Meteorological Administration (CMA) have very little discrimination ability of different observed rain events in the NH tropics. The day +1 QPFs from Japan Meteorological Administration (JMA) have remarkably large moist biases in the NH tropics, which leads to the discontinuity of forecast performance with the lead time.

Performance changes due to the major model upgrades during the five summers are also examined using the forecasts from CMA as the reference to eliminate the interannual variation. After the model upgrade, the excessively enlarged ensemble spread of CMC increases the forecast errors, while the QPFs and PQPFs from the US National Centers for Environmental Prediction (NCEP) are significantly improved in various verification metrics.

Keywords: TIGGE, quantitative precipitation forecast (QPF), probabilistic quantitative precipitation forecast (PQPF), Ensemble Prediction System (EPS), verification

1. Introduction

Quantitative precipitation forecasts (QPFs) are of vital importance in preventing and mitigating natural disasters [*Fritsch et al.*, 1998]. Precipitation, a diagnosed variable in numerical weather predictions, is extremely difficult to forecast because the related subgrid physical processes, such as cumulus convective, microphysical, and land surface processes, are hard to be parameterized accurately. Because of the existing large uncertainties in QPFs, it is necessary to employ the ensemble approach to deal with the uncertainty problems. Ensemble prediction systems (EPSs) can give a representation of forecast uncertainties through initial perturbations and model perturbations, and can be used to generate probabilistic QPFs (PQPFs), which are widely used in meteorological and hydrological risk management.

As a major component of The Observing System Research and Predictability Experiment (THORPEX), the THORPEX Interactive Grand Global Ensemble (TIGGE) [*Bougeault et al.*, 2010] makes it possible for research on the operational global ensemble precipitation forecasts. TIGGE started at a workshop in 2005, with the objectives to enhance worldwide collaboration on improving the accuracy of 1-day to 2-week high-impact weather forecasts and advancing the research of ensemble forecasting [*Richardson et al.*, 2005].

Case studies on TIGGE precipitation forecasts have been carried out extensively in heavy rain events and hydrological flood warnings. *Pappenberger et al.* [2008] used TIGGE data as meteorological input to the European Flood Alert System for studying a flood event in Romania in October 2007 and found that awareness of the flood could have been raised as early as 8 days in advance. *He et al.* [2009] applied a coupled atmospheric-hydrologic-hydraulic cascade system driven by the TIGGE data to investigate a flood warning case on a meso-scale catchment in the Midlands regions of England and found that the precipitation uncertainties dominate and propagate through the cascade chain. Similarly, another case study in the Upper Huai catchment during July to September 2008 showed a reliable warning of flood as early as 10 days in advance [*He et al.*, 2010]. *Schumacher and Davis* [2010] examined the skill of the European Centre for Medium-Range Weather Forecasts (ECMWF) EPS in nine heavy rainfall events over 5-day periods in the central and eastern United States during 2007-2008, including three cool-season cases, three warm-season cases, and three tropical cyclone cases. *Wiegand et al.* [2011] studied a heavy precipitation event at the Alpine south side and Saharan Dust over Central Europe through the investigation of the forecast quality and predictability of synoptic and meso-scale aspects and found that ensemble-mean multimodel QPFs can be accurate enough to forecast day 4 for a successful severe-weather warning.

There are several studies of regional cases on TIGGE precipitation forecasts. *Krishnamurti et al.* [2009] concluded that the multimodel superensemble has higher skill than the best single model, by investigating the TIGGE precipitation forecasts over China monsoon region with deterministic verification metrics. *Hamill* [2012] compared the PQPFs from four TIGGE centers with Climatology-Calibrated Precipitation Analysis (CCPA) data over the contiguous United States during July to October 2010, focusing on the TIGGE multimodel and ECMWF reforecast-calibrated PQPFs. His study showed that PQPFs from the Canadian Meteorological Centre (CMC) are most reliable but least sharp, while those from the US National Centers for Environmental Prediction (NCEP) and the United Kingdom Meteorological Office (UKMO) are least reliable but sharper.

However, systematic studies on TIGGE precipitation forecasts are quite few. Thus, a more comprehensive study is needed to reveal detailed properties of QPFs and PQPFs from different centers. For example, the quality of reliability and resolution may provide the useful information about the potential of post-processing to improve precipitation forecasts in the EPS. This study not only uses various verification metrics, but also considers area-weighted forecast scores, aiming to provide overall performance of QPFs and PQPFs. Owing to the availability of global EPSs, the model's ability to simulate heavy rainfall in important areas, such as the Inter Tropical Convergence Zone (ITCZ), can be evaluated with a global view. Fortunately, the global quantitative precipitation estimate (QPE) products, such as the Tropical Rainfall Measuring Mission (TRMM) products [*Huffman et al.*, 2007], make the investigation possible. Since the EPSs have been upgraded from time to time, the benefit of the EPS upgrade is not easily to be assessed by the forecast performance, which is sensitive to the validation period and interannual variation. It is of great interest to quantitatively analyze the improvements of QPFs and PQPFs after the model upgrade.

This study focuses on the 24-h accumulated ensemble mean QPFs and PQPFs generated from individual TIGGE centers in the Northern Hemisphere (NH) midlatitudes and tropics, to obtain a comprehensive understanding and summary of the precipitation forecast properties of six selected operational global EPSs during the recent five-year (2008-2012) summers (June to August, JJA). The overall 5-summer forecast performance of the EPSs is evaluated, including the discrimination ability of rain events, which can indicate the possible improvement of the EPSs through post-processing, and the potential use in economic decision-making for the EPSs. In addition, performance changes before and after major model upgrades are assessed referenced to the China Meteorological Administration (CMA) EPS, which has not been upgraded and can be used to eliminate the impact of the interannual variability on the verification scores.

Section 2 provides an overview of the TIGGE EPSs, while Section 3 describes the datasets and verification methods. Section 4 demonstrates the results with summary and discussions followed in Section 5.

2. Overview of the TIGGE EPSs

Ten operational forecast centers participate in the TIGGE program, including the Bureau of Meteorology of Australia (BoM), CMA, CMC, the Centro de Previsão de Tempo e Estudos Climáticos of Brazil (CPTEC), ECMWF, the Japan Meteorological Administration (JMA), the Korea Meteorological Administration (KMA), the National Meteorological Service of France (Météo-France), NCEP and UKMO. One can access to the TIGGE data about a delay of 48 h through three data portals: the ECMWF portal (http://tigge-portal.ecmwf.int/), the CMA portal (http://bridge.cma.gov.cn:8080/tigge/index.jsp), and the US National Center for Atmospheric Research (NCAR) portal (http://tigge.ucar.edu/).

Six centers are selected in this study: CMA, CMC, ECMWF, UKMO, NCEP and JMA. Four other centers (BoM, CPTEC, Météo-France and KMA) are not included in this investigation for various reasons. BoM stopped providing data to TIGGE on 20 July 2010. CPTEC is a center located in the Southern Hemisphere and its initial perturbations are not performed in the NH midlatitudes. Météo-France only provides short-range ensemble forecasts with 1-3 (1-4.5) day lead times for the 0600 (1800) UTC cycle. For KMA, precipitation fields have not been added to its EPS until 18 December 2009. For the readers' convenience, the main configurations and important upgrades of the six EPSs during 2008-2012 are briefed in Table 1.

CMA uses bred vectors (BVs) [*Toth and Kalnay*, 1997] for the T213 global model (~0.5625º) [*Wang et al.*, 2008] as the initial perturbations to construct the EPS and no model uncertainties have been taken into account. Since no model upgrade has been performed, QPFs and PQPFs from the CMA EPS are chosen to be the benchmark of fluctuated forecast skill due to interannual variability, which makes it possible to investigate the performance changes due to model upgrades in other five EPSs.

The CMC EPS uses Ensemble Kalman Filter (EnKF) [*Houtekamer et al.*, 2009] to generate initial perturbations. To represent model uncertainties, multi-physics schemes (such as different deep convections, surface schemes, mixing lengths, vertical diffusions and gravity wave drags) as well as two stochastic parameterization schemes, *i.e.*, Perturbations of Physics Tendencies (PTP) and Stochastic Kinetic Energy Backscatter (SKEB) [*Gagnon et al.*, 2011] are adopted. On 17 August 2011, the CMC EPS has been upgraded to version 2.0.2 with the finer model horizontal grid spacing of 66 km changing from about 100 km. However, the horizontal resolution of the output data archived in the TIGGE portal remains unchanged.

The ECMWF EPS used the evolved and the initial-time singular vectors (EVO-SVINI) [*Leutbecher*, 2005] as its initial perturbations before 24 June 2010, and since then has been upgraded to the ensemble of data assimilation and the initial-time singular vectors (EDA-SVINI) [*Buizza et al.*, 2008; *Buizza et al.*, 2010]. The Stochastic Perturbation of Physics Tendency (SPPT) [*Buizza et al.*, 1999b] has been applied to account for model uncertainties. The Spectral Stochastic Backscatter Scheme (SPBS) [*Berner et al.*, 2009] was also introduced into the ECMWF EPS to simulate upscale-propagating errors caused by unresolved subgrid-scale processes on 9 November 2010. Actually, the ECMWF EPS has been upgraded frequently, for example, the upgrade on 26 January 2010 (Table 1, more details can refer to http://www.ecmwf.int/products/data/operational_system/evolution/index.html). For simplicity, only the major upgrade time on November 2010 has been assessed.

The JMA EPS uses the singular vectors (SVs) to create initial perturbations. Dry SVs are targeted for the NH extratropics (30ºN-90ºN) while moist SVs are targeted for the tropics (20ºS-30ºN) [*Yamaguchi and Majumdar*, 2010]. Since 17 December 2010, the SPPT method has been applied to account for model uncertainties, with simplified-physics in the NH extratropics and full-physics (also add gravity wave drag, long-wave radiation, clouds and large scale convection and cumulus convection) in the tropics [*Sakai et al.*, 2008]. The model horizontal resolution is about 0.56º, while the archived output data is on 1.25º×1.25º grids (see http://tigge.ecmwf.int/metadata/TIGGE_metadata_v5_JMA.xls).

The NCEP EPS uses the bred vector - ensemble transform with rescaling (BV-ETR) [*Wei et al.*, 2008] to generate initial perturbations. Since 23 February 2010, the Stochastic Total Tendency Perturbation (STTP) scheme [*Hou et al.*, 2006; *Hou et al.*, 2008; *Hou et al.*, 2010] has been introduced into the NCEP EPS to account for model uncertainties, and the model horizontal resolution has been upgraded from T126 (~110 km) to T190 (~70 km) (http://www.emc.ncep.noaa.gov/gmb/ens/ens_imp_news.html). The output data archived in the TIGGE portal remains unchanged. On 14 February 2012, a major upgrade time, the NCEP EPS has been advanced to version 9.0, including the improved BV-ETR initialization and STTP schemes, the upgraded horizontal resolution of T254 (~55 km) for 1-8 day forecasts (9-16 day forecasts remain T190) and the add of sunshine duration for TIGGE data exachange (http://www.emc.ncep.noaa.gov/gmb/yzhu/imp/i201109/GEFS_Science_20120208.pdf).

The UKMO EPS uses the Ensemble Transform Kalman Filter (ETKF) [*Bishop et al.*, 2001; *Bowler et al.*, 2008] as the initial perturbation strategy. Random Parameters (RP) and Stochastic Kinetic Energy Backscatter (SKEB) schemes are used to represent model uncertainties (http://tigge.ecmwf.int/metadata/EGRR_TIGGE_metadata_v14.xls). The version of the UKMO EPS has been changed several times during 2008-2012. On 9 March 2010 (a major upgrade time), the UKMO EPS has been upgraded to version 8 and its horizontal resolution has been improved from 1.25º×0.83º to 0.83º×0.56º.

3. Datasets and verification methods

3.1 Validation dataset

The validation dataset is from the recently created Version 7 TRMM research product 3B42 (ftp://meso-a.gsfc.nasa.gov/pub/trmmdocs/3B42_3B43_doc.pdf). The dataset combines multi-satellite microwave-IR estimates and is adjusted by quality-controlled gauges [*Huffman et al.*, 2007]. The original dataset is 3-hourly and covers 50ºS-50ºN, 180ºW-180ºE, with a horizontal resolution of 0.25º×0.25º. In order to compare with the TIGGE forecast data, it is bilinearly interpolated in space and time to the 1.0º×1.0º daily (1200UTC-1200UTC) precipitation data. The verification region is focused on the NH tropics (0ºN-20ºN) and NH midlatitudes (20ºN-49ºN).

3.2 Forecast dataset

The original ensemble precipitation forecast data of CMA, CMC, ECMWF, UKMO, NCEP and JMA are all converted onto the same 1.0º×1.0º grid before downloading, using the bilinear interpolation software provided by the ECMWF data portal. Whole perturbed members (without the control forecast) of each center are used to compute 24-h ensemble mean QPFs and PQPFs. Only the +1- to +9-day forecasts initialized at 1200UTC are examined due to the limit of the JMA forecast data. The time period of the verification covers JJA 2008-2012 (1 June – 30 August, total 91×5=455 days). Several 1200 UTC cycles of the NCEP forecast data are missing, including the dates of 08, 13, 16, 18, 20 and 25 August 2008. Considering that replacing this small fraction of data will not influence the final results, the missing NCEP forecast data are substituted with the nearest initial forecast cycles.

After processing the ensemble data (usually taking subtraction from the accumulated total precipitation) to the 24-h accumulated precipitation forecasts, there are some negative values for the five summers: a small portion (0.7%~2.4%) of negligible values (-0.1~0 mm day^{-1}) due to numerical computation errors, and a very rare fraction of large values for the CMA (0.01%, -0.9~-0.1 mm day^{-1}), NCEP (0.01%, -8.9~-0.1 mm day^{-1}) and CMC (0.01%,

-87.2~-0.1 mm day^{-1}) EPSs. For simplicity, all negative values of 24-h precipitation forecasts are set to zeros.

3.3 Verification methods

In this study, multiple deterministic and probabilistic verification methods are carried out to demonstrate different aspects of QPFs and PQPFs. Considering the large meridional span, an area-weighted average method is applied to the common verification scores [*Jolliffe and Stephenson*, 2003; *Wilks*, 2006; and references within].

The area-weighted root mean square error (RMSE) is calculated as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} w_i \cdot (x_i - y_i)^2}{\sum_{i=1}^{N} w_i}} \qquad (1)$$

where x_i and y_i represent the ith forecast and observed values, w_i equals to the cosine latitude of the ith sample and N is the sample size (w has the same definition in other scores). Similarly, the Pearson correlation [*Wilks*, 2006] is modified to the spatial correlation (SC) to measure the similarity of two patterns:

$$SC = \frac{\sum_{i=1}^{N} w_i \cdot (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N} w_i \cdot (x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{N} w_i \cdot (y_i - \overline{y})^2}} \qquad (2)$$

where $\overline{x}$ and $\overline{y}$ are the area-weighted averages of forecast and observed values:

$$\overline{x} = \frac{\sum_{i=1}^{N} w_i \cdot x_i}{\sum_{i=1}^{N} w_i} \qquad (3)$$

$$\overline{y} = \frac{\sum_{i=1}^{N} w_i \cdot y_i}{\sum_{i=1}^{N} w_i} \quad (4)$$

The discrimination diagram can be used to demonstrate the ability of the forecast system to discriminate different rain events. The corresponding forecast and observed rain events are denoted as X_j and Y_k (j,k=1,2,…,M) for M rain events. One rain event Y_k corresponds to one discrimination curve. The forecast relative frequency $f(X_j|Y_k)$ conditioned on the observed kth rain event, is plotted against different forecast categories X_j and is calculated as:

$$f(X_j \mid Y_k) = \frac{\sum_{i=1}^{N} w_i \cdot A_j^i \cdot B_k^i}{\sum_{i=1}^{N} w_i \cdot B_k^i} \quad (5)$$

where $A^i{}_j$=1 if the jth event is forecasted for the ith sample or otherwise $A^i{}_j$=0, and $B^i{}_k$ is similar but for the observed kth event. For a perfect forecast system, $f(X_k|Y_k)$=1 and $f(X_j|Y_k)|_{j\neq k}$=0.

Verification metrics for dichotomous forecasts including the bias score (frequency bias, Bias), the equitable threat score (ETS), the probability of detection (POD) and the false alarm ratio (FAR) are also calculated in the area-weighted form, based on the 2×2 contingency table [*Jolliffe and Stephenson*, 2003]. The contingency table is also area-weighted using all samples constructed by $A^i{}_k$ and $B^i{}_k$ (Equation 5).

For probabilistic forecasts, the forecast scores are calculated in a similar area-weighted form. First, the ensemble spread and the spread-skill relationship (spread vs. RMSE) are evaluated. Usually, the continuous ranked probability score (CRPS) and the continuous ranked probability skill score (CRPSS) are used as the summary scores for probabilistic forecasts, while the Brier Score (BS) [*Brier*, 1950] and the Brier skill score (BSS) are used for dichotomous probabilistic forecasts at a selected precipitation threshold. The CRPSS is calculated based on the area-weighted averages of the CRPS and the referenced CRPS ($CRPS_{ref}$) that is generated using the cumulative distribution function (CDF) of the observed samples (*i.e.* sample climatology) on each grid point. Similarly, the BSS is calculated based on the area-weighted averages of the BS and the referenced BS (BS_{ref}) that is generated using the sample climatology frequency on each grid point. The CRPSSs and BSSs calculated in this study are usually much lower than that using the long-term climatology or the sample-weighted average method for distinct climatological regimes [*Hamill and Juras*, 2006].

The BS can be decomposed into three components: reliability (REL), resolution (RES) and uncertainty (UNC) [*Murphy*, 1973]. As the sample climatology differs on grid points, the decomposition is performed on each grid point: $BS_s = REL_s - RES_s + UNC_s$ (s denotes the sth grid point). Each term is calculated as:

$$BS_s = \frac{\sum_{k=1}^{m}\sum_{j=1}^{n_k^s}(p_k - o_{kj}^s)^2}{N_t} \qquad (6)$$

$$REL_s = \frac{\sum_{k=1}^{m} n_k^s \cdot (p_k - \bar{o}_k^s)^2}{N_t} \qquad (7)$$

$$RES_s = \frac{\sum_{k=1}^{m} n_k^s \cdot (\bar{o}^s - \bar{o}_k^s)^2}{N_t} \qquad (8)$$

$$UNC_s = \bar{o}^s \cdot (1 - \bar{o}^s) \qquad (9)$$

where m denotes the number of forecast categories; when ensemble size is M, $m=M+1$ and the probability of the kth forecast category is $p_k=(k-1)/M$; n^s_k denotes the subsample size for the kth forecast category; N_t is the total sample size on each grid point ($N_t=n^s_1+n^s_2+\ldots+n^s_m$, 455 in this study); on each grid for the jth sample, the observed frequency $o^s_{kj}=1$ if the event occurs or otherwise o^{s}_{kj}=0, the conditional average observed frequency is $\bar{o}_k^s = (o_{k1}^s + o_{k2}^s + ... + o_{kn_k^s}^s)/n_k^s$, and the sample climatology is $\bar{o}^s = (\bar{o}_1^s \cdot n_1^s + \bar{o}_2^s \cdot n_2^s + ... + \bar{o}_m^s \cdot n_m^s)/N_t$. The overall scores: *BS*, *REL*, *RES* and *UNC*, can be derived from the area-weighted averages of all grid points:

$$\underbrace{\frac{\sum_{s=1}^{N_s} w_s \cdot BS_s}{\sum_{s=1}^{N_s} w_s}}_{BS} = \underbrace{\frac{\sum_{s=1}^{N_s} w_s \cdot REL_s}{\sum_{s=1}^{N_s} w_s}}_{REL} - \underbrace{\frac{\sum_{s=1}^{N_s} w_s \cdot RES_s}{\sum_{s=1}^{N_s} w_s}}_{RES} + \underbrace{\frac{\sum_{s=1}^{N_s} w_s \cdot UNC_s}{\sum_{s=1}^{N_s} w_s}}_{UNC} \qquad (10)$$

where N_s denotes the number of grid points. As BS_{ref} equals to *UNC*, the overall BS can be expressed as (*RES*-*REL*)/*UNC*.

To further demonstrate the contribution of various forecast categories to the overall REL and RES, the reliability diagram (RD) is shown with the conditioned observed frequencies plotted against the forecast probabilities. The subsample frequencies shown on the RD, which is also called the sharpness graph, are also area weighted. Sharpness solely depends on the forecast, denoting the ability of the forecast system to predict extreme probabilities (0% and 100%). Forecasts with more subsamples for extreme forecast probabilities are sharper. The forecast only based on climatology of observation is perfectly reliable (overlapping with the diagonal line), but it fails to produce enough extreme probabilities (not sharp) with poor discrimination ability (low RES). The REL term (*i.e.* the conditional bias) can be calibrated, while the RES term is difficult to be improved through post-processing. The forecast system only becomes perfect (BSS=1) when perfect REL (REL=0) and perfect RES (RES=UNC) are obtained at the same time.

Compared with the RD, which is conditioned on the forecasts, the Relative Operating Characteristic (ROC) measures the discrimination ability of probabilistic forecasts conditioned on the observations. First, a set of probability thresholds are used to convert the PQPFs into dichotomous predictands. Then the ROC curve is constructed by plotting the corresponding PODs against false alarm rates (or probability of false detections, POFDs) using the 2×2 area-weighted contingency table. The ROC curve overlapping with the diagonal line indicates no discrimination ability of the occurred and non-occurred events in a forecast system, *i.e.*, PODs are always equal to POFDs. Area under the ROC curve (ROCA) is used as a summary scalar of the discrimination ability, ranging from 0 to 1 (perfect forecast), and a ROCA of 0.5 indicates no skill.

The dichotomous predictands generated in the ROC can also be used in economic decision-making. Based on a simple cost-loss model [*Zhu et al.*, 2002], the economic value (EV) here refers to a relative skill score (not actual economic loss) comparing the economic loss from the decision-making generated using the information of PQPFs to that from a constant decision (always take or not take a precautionary action). An EV above 0 indicates useful information from the PQPFs to the decision-making. For a certain probability threshold, the EVs are plotted against the cost/loss (C/L) ratios. The potential EV (PEV) of the PQPFs is obtained by taking the maximum EV of all probability thresholds for different C/L ratios. The corresponding optimal probability thresholds for different C/L ratios are also plotted as scatters. If the forecast system is perfectly reliable, the scatters should line on the diagonal line of the PEV graph [*Jolliffe and Stephenson*, 2003].

Error bars are shown for the RMSE, ensemble spread and CRPSS, representing the 90% confidence intervals using resampling method by randomly selecting the statistics 10000 times [*Hamill*, 1999]. As for the Bias, ETS, POD, FAR, BSS and ROCA, the error bars are too short and not shown.

Finally, the impacts of major model upgrades on the forecast performance are examined for several scores (ETS, RMSE, CRPSS, BSS, spread, and spread/RMSE ratio). To eliminate the impact of interannual variation, the score changes of other five centers due to the major model upgrade are compared with the corresponding score of CMA (frozen version). Considering 90% confidence intervals, the performance change of the forecast score due to the major model upgrade is thought to be significant when three criteria are satisfied: (a) the score change is significant; (b) the change of the score difference between the center and CMA is significant; (c) the trends of change in (a) and (b) are consistent (same sign).

4. Results

4.1 Verification of ensemble mean QPFs

4.1.1 Precipitation climatology and forecast errors

The precipitation climatology (Figure 1) of the day +3 ensemble mean QPFs from the six EPSs and the TRMM observations during JJA 2008-2012 are compared. All EPSs (Figure 1a-f) can reproduce major observed heavy rain belts globally with high spatial correlation coefficients, but demonstrate different regional forecast errors. The CMC and UKMO EPSs tend to overestimate rain areas in the west coast of India, while the CMA and JMA EPSs have large overall forecast errors (RMSE of 1.8~2.0 mm day^{-1}). The CMA EPS fails to reproduce the heavy rain area in the western Pacific near the equator (120°E-160°E, 0°N-10°N), and the JMA EPS fails to reproduce the heavy rain center in the Bay of Bengal. In general, the ECMWF EPS shows the least RMSE of 1.28 mm day^{-1} for the day +3 QPFs, and the relative performance of precipitation climatology at other lead times is similar (not shown) for all EPSs. In particular, the day +1 JMA EPS (Figure 1g) shows noteworthy moist biases in the NH tropics and causes the discontinuity of forecast scores with the lead time, because JMA employs moist SVs over the entire tropics and perturbs the specific humidity with a large amplitude [*Yamaguchi and Majumdar*, 2010].

Compared to ensemble mean QPFs, the control QPFs from the six EPSs show different overall forecast errors (RMSE, Figure 2). For the control QPFs, the JMA EPS significantly outperforms other EPSs in the NH midlatitudes, especially for longer lead times, while the ECMWF, UKMO and JMA EPSs have less forecast errors than other three EPSs in the NH tropics. For the ensemble mean QPFs, the ECMWF EPS is the best in both regions, while the CMC, UKMO and JMA EPSs are relatively better than the NCEP and CMA EPSs for longer lead times. Although the control QPFs from CMC are inferior to those from JMA and UKMO, the ensemble mean QPFs from the three centers are comparable in both regions. This indicates that the QPFs in the CMC EPS benefit more from the ensemble configuration.

4.1.2 QPFs of categorical and dichotomous events

The discrimination diagram illustrates how different discrimination curves (conditioned on the observed rain events) separate with each other, indicating the ability to discriminate different observed rain events. For the day +1 ensemble mean QPFs (Figure 3), all EPSs are able to discriminate observed light, moderate and heavy rain events to some degree in the NH midlatitude, while the discrimination ability is relatively low in the NH tropics. For example, for the day +1 ensemble mean QPFs in the NH tropics, the poor performance of the CMA EPS causes little discrimination ability among different rain events (Figure 3a3), and the JMA EPS overforecasts more observed light rain events as moderate rain events (Figure 3f3) due to the large moist bias (Figure 1g). The low predictability of QPFs in the NH tropics is perhaps associated with the complex convective processes in this region, which remains a great challenge to the model communities. The discrimination ability decreases with the lead time indicated by the day +1 and day +5 diagrams (Figure 3, other lead times not shown), as the curves representing different observed rain categories gradually become indistinguishable towards light rain events. The day +5 ensemble mean QPFs of most EPSs completely lose discrimination ability, except the marginal discrimination ability in the ECMWF and UKMO EPSs.

Other commonly used dichotomous scores are computed for the ensemble mean QPFs at varied lead times and precipitation thresholds (Figure 4). In both the NH midlatitudes and NH tropics, all EPSs overforecast the light precipitation (>1mm day^{-1}) and underforecast the heavier precipitation (>25 and 50 mm day^{-1}, Figure 4a, b). Generally, ECMWF demonstrates the best forecast quality (ETS, Figure 4c, d), while NCEP has the relatively good bias score (Figure 4a, b). The selected scores are linked, such as the existing relation of Bias=POD/(1-FAR). Accordingly, the relatively lower POD (Figure 4e, f) and lower FAR (Figure 4g, h) of NCEP contribute to the improved bias scores at the light precipitation threshold, and vice versa at the heavier precipitation thresholds. The significantly lower POD and higher FAR of the CMA EPS in the NH tropics are associated with the significantly lower ETS, consistent with the poor discrimination ability (Figure 3). Also, the verification scores reflect different forecast properties, and may not be consistent. For instance, the bias scores of CMA are similar to those of other centers, despite of its other poor scores. This is because a good bias score, independent of location errors, is only a necessary but not sufficient condition of an accurate forecast. Consequently, all scores should be used and interpreted with caution.

4.2 Verification of PQPFs

4.2.1 Spread-skill relationship and CRPSS

A well-constructed EPS should have the fast growing ensemble spread which can capture the growth of forecast error. The spread-skill relationship (Figure 5) is measured by the ensemble spread and ensemble mean error in this study. The CMC EPS uses multi-physics schemes to represent model uncertainties and initiates a large ensemble spread with the fastest growth rate and large day to day variation (long error bars of the spread). With the increasing lead time, the ensemble spread of CMC grows to level with the ensemble mean error in the NH midlatitudes while becomes overdispersive in the NH tropics. Other five EPSs are severely underspersive and suffer from spread deficiencies in both regions. The day +1 ensemble spread of JMA is the largest in the NH tropics due to the use of moist SVs, and drops to the lowest with the slowest growth rate after the day +2 lead times. In addition, an EPS with large ensemble size does not necessarily possess large ensemble spread or improved spread-skill relationship. For example, the ensemble spreads of CMA with 14 ensemble members and ECMWF with 50 ensemble members are very close for longer lead times. Considering larger RMSEs in the CMA EPS, the ECMWF EPS has better spread-skill relationship. Another example is that the JMA EPS (50 members) has worse spread-skill relationship compared to the CMC EPS (20 members), because the former has the similar RMSEs but much smaller ensemble spread.

The overall performance of PQPFs from the six centers is evaluated by the CRPSS (Figure 6) using the CDF of sample climatology on each grid point as the reference forecast. The CRPSS here is conventionally calculated and its value highly depends on the forecast errors of large precipitation amount [*Hamill*, 2012]. Nevertheless, the relative performance of different centers is revealed by the CRPSSs (Figure 6), indicating higher PQPF skills in the NH midlatitudes than that in the NH tropics and the best skill for the ECMWF EPS in both regions. CMC has the second best CRPSS of day +1 PQPFs and the skill rapidly drops from day +2, which may be related to its fast growing of ensemble spread and large forecast errors. For longer lead times (day +3 ~ +9), JMA ranks the second best followed by NCEP and UKMO in the NH midlatitudes. In the NH tropics, UKMO ranks the second best for longer lead times, and CMA has the extremely poor performance as its CRPSS of day +1 PQPFs is even worse than that of day +9 PQPFs from ECMWF.

4.2.2 PQPF skill of dichotomous events

Compared with the CRPSS, the BSS equally weights different grid points irrespective of the distance between the precipitation amount and the precipitation threshold. The BSSs of PQPFs (Figure 7) show that CMC obviously outperforms other centers at the 1 mm day^{-1} threshold, and CMC and ECWMF are more skillful at heavier precipitation thresholds. In addition, the BSS varies with the precipitation threshold, and is sensitive to the conditional bias. ECMWF has the relatively low BSS at 1mm day^{-1} in the NH tropics due to the poor reliability (Figure 8c3). The good reliability of CMC and the good resolution of ECMWF (Figure 8b1-4, c1-4) contribute to higher BSSs in both EPSs. The conditional bias (reliability term) can be calibrated through post-processing while the resolution term is associated with the model itself and difficult to be post-processed. At the 1 mm day^{-1} threshold, the resolution terms (Figures 8d1, 8d3, 8e1, 8e3) of UKMO and NCEP are very close, thus the discrepancy of BSS between these two centers (Figure 7) is mainly caused by the difference of reliability. At the 10 mm day^{-1} threshold, both the reliability and resolution terms of UKMO are better than those of NCEP (Figures 8d2, 8d4, 8e2, 8e4), which leads to better BSSs of UKMO (Figure 7).

The reliability diagrams of day +3 PQPFs at the 1 mm day^{-1} and 10 mm day^{-1} thresholds (Figure 8) show overconfident forecasts with flatter reliability curves by underestimating both ends of extreme probabilities for all EPSs. Though the CMC EPS (Figure 8b1-4) is most reliable (the curves closest to the diagonal line), but is not sharp enough due to the large discrepancy of its ensemble members. The frequencies of CMC forecasts with high probability categories are extremely low (less than one in a thousand) (Figures 8b2, 8b4). In contrast, UKMO and NCEP are sharp, with more forecasts of extreme probabilities (Figures

8d1-4, e1-4). The day +3 PQPFs from CMA have the worst resolution (smallest RES) while those from JMA have the worst reliability (largest REL). This indicates the relatively poorer model quality of CMA and larger conditional biases of JMA. In particular, for the day +1 PQPFs from JMA in the NH tropics (Figure 8g3-4), the observed frequencies of conditional wet biases are increased due to the large moist biases (Figure 1g). For other lead times (not shown), the reliability curves are similar to those of the day +3 PQPFs. At the 25 mm day^{-1} and 50 mm day^{-1} thresholds (not shown), the dry forecast probabilities (zero) are dominated and the frequencies at high probabilities are largely reduced for each center.

4.2.3 Discrimination ability and potential economic value

Figure 9 demonstrates the ROCAs at different precipitation thresholds and lead times. PQPFs from CMC and ECMWF have the strongest ability to discriminate different observed rain events. *Buizza et al.* [1999a] considers an ROCA of 0.7 as the limit of a useful prediction system. Nearly all centers are useful for the day +1 to +5 lead times at the 1 to 25 mm day^{-1} precipitation thresholds in the NH midlatitudes while PQPFs from CMA, NCEP and JMA lack skill at the 50 mm day^{-1} precipitation threshold. ROCAs in the NH tropics are relatively lower for all EPSs, especially for heavier precipitation thresholds. ROCAs of CMA in the NH tropics are very poor and slightly vary with increasing lead times, indicating inferior discrimination ability of PQPFs.

Based on the ROCA, the PEV curves and the optimal probability thresholds (Figure 10) are calculated for taking action as a function of C/L ratios for day +3 PQPFs at different precipitation thresholds. Except the high PEV values of CMC at 1 mm day^{-1} precipitation threshold for high C/L ratio users, ECMWF has the highest PEV values. PQPFs from ECMWF outperform other centers more for heavier precipitation thresholds, indicating large potential use in economic decision making. PQPFs from CMA have the least PEV and smallest range of C/L ratios, showing a large gap compared to other centers (Figures 10a,b,d,e). Among all centers, the optimal probability thresholds against different C/L ratios from CMC are closest to the diagonal lines, especially at the 1 mm day^{-1} precipitation threshold, indicating the best reliability [*Jolliffe and Stephenson*, 2003]. The optimal probability thresholds of ECMWF are close to the diagonal line at heavier precipitation thresholds, but largely deviate from the diagonal line at the 1 mm day^{-1} threshold in the NH tropics, indicating its relatively bad reliability (Figure 8c3). PEV curves of other lead times (not shown) are similar except those from the JMA day +1 PQPFs.

4.3 Performance changes due to model upgrade

One concern in the design of EPS is to gain better spread-skill relationship. Table 2 provides the average ensemble spread of five centers and their spread differences with CMA for the day +3 forecasts before and after the major model upgrade. All the five centers have significant spread changes with 90% confidence interval. Ensemble spread of ECMWF is reduced while other four centers increase their spread. CMC enlarges their ensemble spread remarkably, with an increase of 3.5 and 3.4 mm day^{-1} in the NH midlatitudes and the NH tropics respectively. Figure 11 illustrates the time series of ensemble spread and RMSE of ensemble mean for the day +3 forecasts of each center in the NH midlatitudes. All the five centers have significantly changed the spread/RMSE ratio. Ensemble spread of ECMWF becomes more deficient, while the spread deficiencies of UKMO, NCEP and JMA are mitigated. The changes of ensemble spread and spread-skill relationship at different lead times (Table 3) before and after the major model upgrades are similar with those of the day +3 forecasts, except that the changes of short-range forecasts of JMA are insignificant.

Upgrading the EPS is expected to improve ensemble mean QPFs. RMSEs (Table 4) of the day +3 ensemble mean QPFs from UKMO and NCEP are reduced significantly while there are no significant RMSEs changes for ECMWF and JMA after the major model upgrade. The RMSE of ECMWF QPFs is quite small compared to other centers and is hard to be improved further. Notably, CMC has increased the RMSE after the model upgrade, because oversized ensemble spread (Figure 11b) usually causes large forecast errors. The day +3 10 mm day^{-1} ETSs (Table 5) of ECMWF, UKMO and NCEP are improved in the NH tropics, and little changes exist for JMA and CMC. NCEP has relatively lower ETS in the NH midlatitudes before the model upgrade and achieves the most remarkable improvement in ETS. Table 6 demonstrates the changes of RMSE and ETS of different lead times and precipitation thresholds before and after the major model upgrade for each center in the NH midlatitudes and NH tropics. RMSEs of CMC deteriorate after the model upgrade for most of the lead times while the ETSs do not, because ETS is a dichotomous forecast score associated with the selected precipitation threshold and is insensitive to ensemble spread. The day +1 to +9 RMSEs of UKMO are reduced and the 10 mm day^{-1} ETS in the NH tropics is also improved. However, the 1 mm day^{-1} and 50 mm day^{-1} ETSs are deteriorated. NCEP has not only improved the day +3 to +9 RMSEs, but also the ETSs at heavier thresholds over 10 mm day^{-1} (except 25 mm day^{-1} ETS in the NH tropics).

At the same time, the PQPFs are expected to be improved through the EPS upgrade. All the centers have significantly changed CRPSS for the day +3 PQPFs except JMA (Figure 12). The CRPSSs of ECMWF, UKMO and NCEP are improved significantly after major model upgrades as the gaps between the two time series become larger (Figure 12b-d). However, the CRPSS of CMC becomes even lower than that from the static version of CMA after the model upgrade. The deterioration of CRPSS of CMC is probably due to its remarkably increased ensemble spread. Unlike CRPSS that more depends on precipitation amount, the BSS is sensitive to the selected precipitation threshold. The 10 mm day^{-1} BSSs of ECMWF, UKMO and NCEP are improved (Table 7), while there are no significant changes for CMC and JMA. At different lead times and precipitation thresholds (Table 8), the PQPF skill (CRPSS and BSS) of JMA has not been changed much after the model upgrade; the PQPFs of NCEP generally have been improved in the NH midlatitudes and NH tropics; both ECMWF and UKMO not only have improved the CRPSSs, but also the BSSs of some certain thresholds; though CMC has improved the BSSs at lighter precipitation thresholds, its CRPSS has decreased significantly.

5. Summary and discussions

This study provides a comprehensive verification on ensemble mean QPFs and PQPFs from six operational global EPSs in the NH midlatitudes and NH tropics during the boreal summers of 2008-2012. Taking the latitudinal discrepancies into account, a series of verification metrics are employed using an area-weighted average method to evaluate the performance of different operational centers at different lead times and precipitation thresholds. Performance changes due to the major model upgrade during the five summers are also examined using the forecasts from CMA as the reference to eliminate the interannual variation due to the unavailability of the parallel run results of different model versions.

For the ensemble mean QPFs during the 5-year summers, CMA has relatively large systematic biases in the NH tropics. In fact, different kinds of deterministic and probabilistic verification scores employed here reveal that CMA performs poorly in the NH tropics, with very little discrimination ability of different observed rain events. The day +1 QPFs from JMA has remarkable moist biases in the NH tropics as they employ moist SVs for the entire tropics and perturb the specific humidity with a large amplitude. This causes the discontinuity of QPF performance against lead times and should be treated differently.

Considering PQPFs during the 5-year summers, ECMWF generally performs best, except at light precipitation thresholds ECMWF and UKMO have lower forecast skill in the NH tropics due to the relatively poor reliability. The PQPF performance of CMC is relatively good for light precipitation thresholds and short-range forecasts. For longer lead times, the ensemble spread of CMC grows excessively large and causes large forecast errors, which mainly results from the use of multi-physics schemes to represent model uncertainties. JMA has the smallest ensemble spread except the day +1 forecasts in the NH tropics. The reliability diagrams reveal that ECMWF has the best discrimination ability (large resolution term); CMC has the least conditional biases (small reliability term), but lacks extremely high probabilities and is the least sharp due to the large discrepancy of its ensemble members. In contrast, PQPFs from UKMO and NCEP are the most sharp.

The verification results are sensitive to the uncertainties and quality of verification data (data quality control, interpolation method, location and so on). *Yuan et al.* [2005] showed that skill scores highly depend on the verification (observation/analysis) data. *Hamill* [2012] investigated PQPFs of TIGGE, and most conclusions about the relative performance of individual centers are consistent with this study. However, some his results are different, for example, the CRPSS from NCEP is superior to that from UKMO, while the CRPSSs of the two centers are of the same level in this study. The difference is that he used a modified version of CRPS to equally weight the dry and wet grid points and verified for different period and geographical location. It is not appropriate to judge which of the two centers has better PQPF skill, but instead to interpret these results with caution.

The ultimate goal of verification study is to improve the performance of QPFs and PQPFs. The post-processing work and the development of the EPSs are two major ways to reach such goal. This study not only evaluates the merits and shortcomings of each EPS for model developers and users, but also provides some useful information about the potential of post-processing to improve precipitation forecasts in the EPS. For example, the ensemble mean QPFs and PQPFs from CMA in the NH tropics have very little discrimination ability of the observed different rain events and thus would be extremely difficult to be improved through calibration. In contrast, though PQPFs from ECMWF are not as reliable as those from CMC, they have enough discrimination ability and the systematic bias can be reduced through calibration. Thus, the centers with less discrimination ability should invest more on the development of the model, while the centers with relatively high model quality can benefit more from the post-processing work to further improve QPFs and PQPFs.

Whether the EPS upgrade may benefit QPFs and PQPFs is of interest to investigate. The EPSs have been upgraded gradually during five years, except for the CMA EPS. Therefore, the performance changes related to the major model upgrades have been evaluated for five operational centers referenced to the CMA EPS. The ensemble spread and spread/RMSE ratio of ECMWF have been significantly reduced while other four centers have significantly increased their spread with inflated spread/RMSE ratios. In particular, after the model upgrade to version 2.0.2 in CMC, remarkably increased ensemble spread leads to increased forecast errors (RMSE) and decreased PQPF skill (CRPSS). After the major upgrade, JMA has not been improved much, while ECMWF, NCEP, and UKMO have reduced forecast errors (RMSEs of ensemble mean QPFs) and increased PQPF skill (CRPSS). The improvements in ETS and BSS vary with selected precipitation thresholds and lead times. The model upgrade cannot always guarantee the skill improvements, and increasing ensemble spread as well as spread/error ratio also may cause negative effect on QPFs and PQPFs.

How to fairly evaluate an EPS is essential for the development and upgrade of the EPSs. A few simple summary scores have limitations and cannot justify whether the old EPS should be upgraded to the new EPS. For example, the bias score denotes the ratio of forecasted events and observed events while cannot express the displacement errors, thus only serves a necessary but not sufficient condition of accurate forecasts. In the NH tropics, bias scores of CMA are close to other centers while the ETSs of CMA have large gaps with other centers. In addition, verification scores or skill scores for dichotomous events (such as ETS and BSS) vary with different precipitation thresholds and lead times, while continuous scores (such as CRPSS) provide an overview of one forecast property. *Gagnon et al.* [2011] examined the PQPFs from two versions of the CMC EPS during 2009 winter and concluded that the new version (2.0.2) outperforms the old version, based on the day +6 and +7 BSs of different precipitation thresholds and the 2.5 and 15 mm day^{-1} precipitation thresholds BSs of different lead times. In this study, though BSSs of PQPFs from CMC are improved at some precipitation thresholds, the CRPSSs are deteriorated as a consequence of the excessively enlarged ensemble spread, because the continuous score CRPSS is sensitive to the precipitation amount. In comparison, NCEP has improved the CRPSSs and BSSs of different thresholds for nearly all lead times. Therefore, both scores for continuous forecasts and dichotomous forecasts at different thresholds for different lead times are suggested to draw a comprehensive conclusion.

Acknowledgements

ECMWF is thanked for providing the TIGGE data portal. This study received support from the National Basic Research Program of China (973 Program) (2013CB430106), the R&D Special Fund for Public Welfare Industry (Meteorology) (GYHY201206005), Natural Science Foundation of China (41175087), and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Berner, J., G. J. Shutts, M. Leutbecher, and T. N. Palmer (2009), A Spectral Stochastic Kinetic Energy Backscatter Scheme and Its Impact on Flow-Dependent Predictability in the ECMWF Ensemble Prediction System, *Journal of the Atmospheric Sciences*, *66*(3), 603-626, doi:10.1175/2008jas2677.1.

Bishop, C. H., B. J. Etherton, and S. J. Majumdar (2001), Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects, *Monthly Weather Review*, *129*(3), 420-436.

Bougeault, P., et al. (2010), THE THORPEX INTERACTIVE GRAND GLOBAL ENSEMBLE, *Bulletin of the American Meteorological Society*, *91*(8), 1059-1072, doi:10.1175/2010bams2853.1.

Bowler, N. E., A. Arribas, K. R. Mylne, K. B. Robertson, and S. E. Beare (2008), The MOGREPS short-range ensemble prediction system, *Quarterly Journal of the Royal Meteorological Society*, *134*(632), 703-722, doi:10.1002/qj.234.

Brier, G. W. (1950), Verification of forecasts expressed in terms of probability, *Monthly Weather Review*, *78*(1), 1-3.

Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli (1999a), Probabilistic predictions of precipitation using the ECMWF Ensemble Prediction System, *Weather and Forecasting*, *14*(2), 168-189.

Buizza, R., M. Leutbecher, and L. Isaksen (2008), Potential use of an ensemble of analyses in the ECMWF Ensemble Prediction System, *Quarterly Journal of the Royal Meteorological Society*, *134*(637), 2051-2066.

Buizza, R., M. Leutbecher, L. Isaksen, and J. Haseler (2010), Combined use of EDA-and SV-based perturbations in the EPS, *ECMWF Newsletter*, *123*, 22-28.

Buizza, R., M. Milleer, and T. Palmer (1999b), Stochastic representation of model uncertainties in the ECMWF ensemble prediction system, *Quarterly Journal of the Royal Meteorological Society*, *125*(560), 2887-2908.

Fritsch, J. M., et al. (1998), Quantitative precipitation forecasting: Report of the eighth prospectus development team, US Weather Research Program, *Bulletin of the American Meteorological Society*, *79*(2), 285-299.

Gagnon, N., G. Pellerin, P. L. Houtekamer, M. Charron, S.-J. Baek, L. Spacek, B. He, and X.-X. Deng (2011), Improvements to the Global Ensemble Prediction System (GEPS 2.0.2).

Hamill, T. M. (1999), Hypothesis tests for evaluating numerical precipitation forecasts, *Weather and Forecasting*, *14*(2), 155-167.

Hamill, T. M. (2012), Verification of TIGGE Multimodel and ECMWF Reforecast-Calibrated Probabilistic Precipitation Forecasts over the Contiguous United States*, *Monthly Weather Review*, *140*(7), 2232-2252.

Hamill, T. M., and J. Juras (2006), Measuring forecast skill: is it real skill or is it the varying climatology?, *Quarterly Journal of the Royal Meteorological Society*, *132*(621C), 2905-2923.

He, Y., F. Wetterhall, H. J. Bao, H. Cloke, Z. J. Li, F. Pappenberger, Y. Z. Hu, D. Manful, and Y. C. Huang (2010), Ensemble forecasting using TIGGE for the July-September 2008 floods in the Upper Huai catchment: a case study, *Atmospheric Science Letters*, *11*(2), 132-138, doi:10.1002/asl.270.

He, Y., F. Wetterhall, H. Cloke, F. Pappenberger, M. Wilson, J. Freer, and G. McGregor (2009), Tracking the uncertainty in flood alerts driven by grand ensemble weather predictions, *Meteorological Applications*, *16*(1), 91-101.

Hou, D., Z. Toth, and Y. Zhu (2006), A stochastic parameterization scheme within NCEP global ensemble forecast system, paper presented at the 18th AMS conference on probability and statistics, Atlanta, Georgia, 29 January - 2 February 2006.

Hou, D., Z. Toth, Y. Zhu, and W. Yang (2008), Impact of a stochastic perturbation scheme on global ensemble forecast, paper presented at the 19th AMS conference on probability and statistics, New Orleans, Louisiana, 21-24 January 2008.

Hou, D., Z. Toth, Y. Zhu, W. Yang, and R. Wobus (2010), A Stochastic Total Tendency Perturbation Scheme Representing Model-Related Uncertainties in the NCEP Global Ensemble Forecast System, *Submitted to Tellus*.

Houtekamer, P., H. L. Mitchell, and X. Deng (2009), Model error representation in an operational ensemble Kalman filter, *Monthly Weather Review*, *137*(7), 2126-2143.

Huffman, G. J., D. T. Bolvin, E. J. Nelkin, D. B. Wolff, R. F. Adler, G. Gu, Y. Hong, K. P. Bowman, and E. F. Stocker (2007), The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales, *Journal of Hydrometeorology*, *8*(1), 38-55.

Jolliffe, I. T., and D. B. Stephenson (2003), *Forecast verification: a practitioner's guide in atmospheric science*, John Wiley & Sons.

Krishnamurti, T., A. D. Sagadevan, A. Chakraborty, A. Mishra, and A. Simon (2009), Improving multimodel weather forecast of monsoon rain over China using FSU superensemble, *Advances in Atmospheric Sciences*, *26*(5), 813-839.

Leutbecher, M. (2005), On ensemble prediction using singular vectors started from forecasts, *Monthly Weather Review*, *133*(10), 3038-3046.

Murphy, A. H. (1973), A new vector partition of the probability score, *Journal of Applied Meteorology*, *12*(4), 595-600.

Pappenberger, F., J. Bartholmes, J. Thielen, H. L. Cloke, R. Buizza, and A. de Roo (2008), New dimensions in early flood warning across the globe using grand-ensemble weather predictions, *Geophysical Research Letters*, *35*(10).

Richardson, D., R. Buizza, and R. Hagedorn (2005), First workshop on the THORPEX interactive grand global ensemble (TIGGE), *WMO, WWRP Document*, 1-39.

Sakai, R., M. Kyouda, M. Yamaguchi, and T. Kadowaki (2008), A new operational one-week Ensemble Prediction System at Japan Meteorological Agency, *CAS/JSC WGNE Research Activities in Atmospheric and Oceanic Modelling*, *38*.

Schumacher, R. S., and C. A. Davis (2010), Ensemble-Based Forecast Uncertainty Analysis of Diverse Heavy Rainfall Events, *Weather and Forecasting*, *25*(4), 1103-1122, doi:10.1175/2010waf2222378.1.

Toth, Z., and E. Kalnay (1997), Ensemble forecasting at NCEP and the breeding method, *Monthly Weather Review*, *125*(12), 3297-3319.

Wang, Y., H. Qian, J.-J. Song, and M.-Y. Jiao (2008), Verification of the T213 global spectral model of China National Meteorology Center over the East-Asia area, *Journal of Geophysical Research*, *113*(D10), D10110.

Wei, M., Z. Toth, R. Wobus, and Y. Zhu (2008), Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system, *Tellus A*, *60*(1), 62-79.

Wiegand, L., A. Twitchett, C. Schwierz, and P. Knippertz (2011), Heavy precipitation at the Alpine south side and Saharan dust over central Europe: A predictability study using TIGGE, *Weather and Forecasting*, *26*(6), 957-974.

Wilks, D. S. (2006), *Statistical methods in the atmospheric sciences*, 2nd ed., Elsevier.

Yamaguchi, M., and S. J. Majumdar (2010), Using TIGGE data to diagnose initial perturbations and their growth for tropical cyclone ensemble forecasts, *Monthly Weather Review*, *138*(9), 3634-3655.

Yuan, H., S. L. Mullen, X. Gao, S. Sorooshian, J. Du, and H.-M. H. Juang (2005), Verification of probabilistic quantitative precipitation forecasts over the southwest United States during winter 2002/03 by the RSM ensemble system, *Monthly Weather Review*, *133*(1), 279-294.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne (2002), The economic value of ensemble-based weather forecasts, *Bulletin of the American Meteorological Society*, *83*(1), 73-83.

Table and figure captions

Table 1. Configurations of six TIGGE EPSs investigated in this study

Table 2. Average ensemble spread (mm day^{-1}) of five centers and their spread differences with CMA for day +3 forecasts before and after the major model upgrade. Boldface represents the significant change with 90% confidence interval.

Table 3. The forecast lead times with significant changes of the ensemble spread and spread/RMSE ratio due to the major model upgrade with 90% confidence interval. The up (down) arrows represents an increase (decrease) change.

Table 4. Same as Table 2, but for the RMSE (mm day^{-1}).

Table 5. Same as Table 2, but for the ETS at the 10 mm day^{-1} threshold.

Table 6. Same as Table 3, but for the RMSE and ETS of ensemble mean QPFs.

Table 7. Same as Table 2, but for the BSS at the 10 mm day-1 threshold.

Table 8. Same as Table 3, but for the CRPSS and BSS of PQPFs.

Figure 1. Average precipitation (mm day-1) of ensemble mean forecasts from the six EPSs and TRMM observation during JJA 2008-2012. The RMSE (mm day^{-1}) and spatial correlation (SC) of forecast and observation averages are shown as the numbers in the titles.

Figure 2. The RMSE of the control forecasts (dotted) and ensemble mean forecasts (solid) (mm day^{-1}) during JJA 2008-2012 in (a) the NH midlatitudes and (b) the NH tropics. Error bars represent 90% confidence intervals.

Figure 3. Discrimination diagrams of the ensemble mean QPFs in the NH midlatitudes (left two columns) and the NH tropics (right two columns) during JJA 2008-2012. The ordinate shows the forecast relative frequencies of observed light rain (1-10 mm day^{-1}, green), moderate rain (10-25 mm day^{-1}, blue), and heavy rain (25-50 mm day^{-1}, red) against five forecast categories: no rain (N, <1 mm day^{-1}), light rain (L, 1-10 mm day^{-1}), moderate rain (M,

10-25 mm day^{-1}), heavy rain (H, 25-50 mm day^{-1}) and torrential rain (T, >50 mm day^{-1}).

Figure 4. The Bias, ETS, POD and FAR of the ensemble mean QPFs against different precipitation thresholds for different forecast lead times (day +1, +3 and +5) during JJA 2008-2012.

Figure 5. The RMSE of the ensemble mean QPFs (dotted) and the ensemble spread (solid) in (a) the NH midlatitudes and (b) the NH tropics during JJA 2008-2012. Error bars represent 90% confidence intervals.

Figure 6. The CRPSS of PQPFs in (a) the NH midlatitudes and (b) the NH tropics during JJA 2008-2012. Error bars represent 90% confidence intervals.

Figure 7. The BSS of PQPFs against different precipitation thresholds for different forecast lead times (day +1, +3 and +5) in (a) the NH midlatitudes and (b) the NH tropics during JJA 2008-2012.

Figure 8. Reliability diagrams for day +3 and +1 PQPFs at the 1 mm day^{-1} and 10 mm day^{-1} thresholds in the NH midlatitudes (left two columns) and the NH tropics (right two columns). The bar graphs show the subsample frequencies at the logarithm scale. The BSS, and the reliability (REL) and resolution (RES) terms of the BS are shown as the numbers. For clearity, the 50 member ECMWF and JMA are converted into 26 probability bins.

Figure 9. The area under the Relative Operating Characteristic (ROC) curve against different precipitation thresholds for different forecast lead times (day +1, +3 and +5) in (a) the NH midlatitudes and (b) the NH tropics during JJA 2008-2012.

Figure 10. Potential economic value (PEV) curves and the optimal probability thresholds for taking action as a function of cost/loss ratio for day +3 PQPFs at different precipitation thresholds.

Figure 11. Time series of the ensemble spread and RMSE for the day +3 of ensemble mean

QPFs in the NH midlatitudes. The dotted vertical line splits the time periods before and after the major model upgrade. The averaged ratios of the ensemble spread and RMSE during the two periods are also shown as the numbers. All changes of the spread/RMSE ratio in the five EPSs (b-f) are significant with 90% confidence interval.

Figure 12. Time series of CRPSS for the day +3 PQPFs in the NH midlatitudes. The dotted vertical line splits the time periods before and after the major model upgrade. The CRPSS differences between each center and CMA during the two periods are also shown as the numbers. Except JMA (e), the CRPSS changes in the four EPSs (a-d) are significant with 90% confidence interval.

Table 1. Configurations of six TIGGE EPSs investigated in this study

Center	Base time (UTC)	No. of ensemble members	Horizontal resolution archived	Forecast length (day)	Initial perturbation method	Model uncertainty	Major model upgrade time
CMA (China)	00/12	14+1	$0.56^{\circ}\times0.56^{\circ}$	0-10	BVs	-	-
CMC[a] (Canada)	00/12	20+1	$1.0^{\circ}\times1.0^{\circ}$	0-16	EnKF	PTP + SKEB multi-physics	17 Aug 2011
ECMWF[b] (Europe)	00/12	50+1	N320(~0.28°) N160(~0.56°)	0-10 10-15	EDA-SVINI	SPPT + SPBS	9 Nov 2010
JMA[c] (Japan)	12	50+1	$1.25^{\circ}\times1.25^{\circ}$	0-9	SVs	SPPT	17 Dec 2010
NCEP[d] (USA)	00/06/12/18	20+1	$1.0^{\circ}\times1.0^{\circ}$	0-16	BV-ETR	STTP	23 Feb 2010
UKMO[e] (UK)	00/12	23+1	$0.83^{\circ}\times0.56^{\circ}$	0-15	ETKF	RP + SKEB	9 Mar 2010

[a]The CMC EPS was upgraded to version 2.0.2 on 17 August 2011.
[b]The ECMWF EPS used a horizontal resolution of N200 (~0.45°) for 0-10 day forecasts and N128 (~0.7°) for 10-15 day forecasts before 26 January 2010. EVO-SVINI was used as the initial perturbation method before 24 Jun 2010. The SPBS method has been added on 9 November 2010.
[c]The JMA EPS began to use the SPPT method on 17 December 2010.
[d]The NCEP EPS was upgraded to version 8.0 and began to use the STTP method on 23 February 2010. In 14 February 2012, the NCEP EPS was upgraded to version 9.0.
[e]The UKMO EPS used a horizontal resolution of $1.25^{\circ}\times0.83^{\circ}$ before 9 March 2010.

Table 2. Average ensemble spread (mm day^{-1}) of five centers and their spread differences with CMA for day +3 forecasts before and after the major model upgrade. Boldface represents the significant change with 90% confidence interval.

	NH midlatitudes			NH tropics		
Center	Before	After	Change	Before	After	Change
CMC	5.8	9.3	**3.5**	11.2	14.6	**3.4**
ECMWF	4.7	4.1	**-0.5**	6.9	5.4	**-1.5**
UKMO	4.3	4.5	**0.2**	4.9	5.2	**0.4**
NCEP	3.1	4.0	**0.9**	4.7	6.1	**1.3**
JMA	3.1	3.5	**0.4**	4.9	5.2	**0.3**
CMC-CMA	1.1	4.4	**3.3**	5.4	8.9	**3.5**
ECMWF-CMA	-0.1	-0.7	**-0.6**	1.0	-0.3	**-1.3**
UKMO-CMA	-0.4	-0.2	**0.2**	-1.1	-0.4	**0.6**
NCEP-CMA	-1.7	-0.8	**0.9**	-1.2	0.4	**1.6**
JMA-CMA	-1.6	-1.3	**0.3**	-1.0	-0.5	**0.5**

Table 3. The forecast lead times with significant changes of the ensemble spread and spread/RMSE ratio due to the major model upgrade with 90% confidence interval. The up (down) arrows represents an increase (decrease) change.

Score	NH Region	CMC	ECMWF	UKMO	NCEP	JMA
SPREAD	midlatitudes	1-9 ↑	1 ↑ 2-9 ↓	1-7 ↑	1-9 ↑	2-9 ↑
	tropics	1-9 ↑	2-9 ↓	1-6 ↑	1-9 ↑	3-9 ↑
SPREAD/RMSE	midlatitudes	1-9 ↑	1 ↑ 2-9 ↓	1-9 ↑	1-9 ↑	2-9 ↑
	tropics	1-9 ↑	2-9 ↓	1-9 ↑	1-9 ↑	5,6,8,9 ↑

Table 4. Same as Table 2, but for the RMSE (mm day^{-1}).

	NH midlatitudes			NH tropics		
Center	Before	After	Change	Before	After	Change
CMC	7.0	7.5	**0.5**	11.6	12.3	**0.7**
ECMWF	6.7	6.6	-0.1	11.1	11.1	-0.0
UKMO	7.4	7.0	**-0.4**	11.9	11.4	**-0.5**
NCEP	7.4	7.2	**-0.3**	12.5	12.1	**-0.4**
JMA	7.0	7.1	0.2	11.7	12.1	**0.4**
CMC-CMA	-0.4	-0.1	**0.3**	-0.6	-0.5	**0.2**
ECMWF-CMA	-0.7	-0.8	**-0.1**	-1.2	-1.6	**-0.4**
UKMO-CMA	-0.1	-0.3	**-0.3**	-0.4	-1.1	**-0.8**
NCEP-CMA	-0.0	-0.2	**-0.2**	0.2	-0.4	**-0.6**
JMA-CMA	-0.5	-0.3	**0.2**	-0.5	-0.5	0.0

Table 5. Same as Table 2, but for the ETS at the 10 mm day^{-1} threshold.

	NH midlatitudes			NH tropics		
Center	Before	After	Change	Before	After	Change
CMC	0.224	0.224	0	0.2	0.2	0
ECMWF	0.290	0.303	**0.012**	0.261	0.281	**0.020**
UKMO	0.252	0.264	**0.012**	0.228	0.241	**0.013**
NCEP	0.227	0.261	**0.034**	0.204	0.215	**0.011**
JMA	0.245	0.249	0.003	0.199	0.201	0.002
CMC-CMA	0.019	0.005	-0.014	0.025	0.03	0.005
ECMWF-CMA	0.085	0.091	0.006	0.080	0.112	**0.032**
UKMO-CMA	0.047	0.053	0.006	0.047	0.072	**0.025**
NCEP-CMA	0.022	0.05	**0.028**	0.023	0.046	**0.023**
JMA-CMA	0.04	0.035	-0.005	0.025	0.028	0.003

Table 6. Same as Table 3, but for the RMSE and ETS of ensemble mean QPFs.

Center	NH Region	CMC	ECMWF	UKMO	NCEP	JMA
RMSE	midlatitudes	2-9 ↑	1 ↓	1-9 ↓	3-9 ↓	-
	tropics	3-9 ↑	1 ↓	1-9 ↓	3-9 ↓	-
ETS	midlatitudes	1,6 ↓	-	1, 7, 8 ↓	-	-
(1 mm day^{-1})	tropics	1 ↓ 2-9 ↑	4-9 ↓	3-9 ↓	2-9 ↓	-
ETS	midlatitudes	-	1, 5 ↑	8 ↑	1-9 ↑	5-9 ↑
(10mm day^{-1})	tropics	6-8 ↑	1-9 ↑	1-9 ↑	1-3,5,6,8,9 ↑	1 ↑
ETS	midlatitudes	7-9 ↑	-	-	1-9 ↑	2-9 ↑
(25 mm day^{-1})	tropics	1,2 ↓ 6-9 ↑	1 ↑ 8 ↓	-	-	1, 2 ↑
ETS	midlatitudes	-	-	-	1-7 ↑	1 ↑
(50 mm day^{-1})	tropics	6-9 ↑	-	1-4 ↓	1-6 ↑	1 ↑

Table 7. Same as Table 2, but for the BSS at the 10 mm day^{-1} threshold.

	NH midlatitudes			NH tropics		
Center	Before	After	Change	Before	After	Change
CMC	0.118	0.139	**0.021**	0.03	0.04	0.011
ECMWF	0.160	0.209	**0.049**	0.036	0.085	**0.049**
UKMO	0.018	0.067	**0.049**	-0.182	-0.107	**0.075**
NCEP	-0.103	0.032	**0.134**	-0.317	-0.15	**0.167**
JMA	0.025	0.014	-0.011	-0.14	-0.162	-0.022
CMC-CMA	0.123	0.117	-0.007	0.245	0.293	**0.047**
ECMWF-CMA	0.165	0.199	**0.034**	0.241	0.338	**0.096**
UKMO-CMA	0.015	0.066	**0.051**	-0.007	0.15	**0.157**
NCEP-CMA	-0.105	0.032	**0.136**	-0.142	0.107	**0.249**
JMA-CMA	0.03	0.004	-0.026	0.065	0.091	0.025

Table 8. Same as Table 3, but for the CRPSS and BSS of PQPFs.

Score	NH Region	CMC	ECMWF	UKMO	NCEP	JMA
CRPSS	midlatitudes	1-9 ↓	1-9 ↑	2-8 ↑	1-9 ↑	-
	tropics	1-9 ↓	1-9 ↑	1-9 ↑	1-9 ↑	-
BSS	midlatitudes	1-9 ↑	1-4 ↑	-	3-9 ↑	-
(1 mm day^{-1})	tropics	1-9 ↑	1-5 ↑	-	1-8 ↑	-
BSS	midlatitudes	1-2 ↑	1-8 ↑	2-9 ↑	1-9 ↑	-
(10mm day^{-1})	tropics	1,3-9 ↑	1-9 ↑	1-9 ↑	1-9 ↑	-
BSS	midlatitudes	-	1-9 ↑	1-9 ↑	1-9 ↑	8 ↑
(25 mm day^{-1})	tropics	-	1-7 ↑	1-9 ↑	1-9 ↑	-
BSS	midlatitudes	-	-	-	1-7 ↑	1 ↑
(50 mm day^{-1})	tropics	6-9 ↑	-	1-4 ↓	1-6 ↑	1 ↑

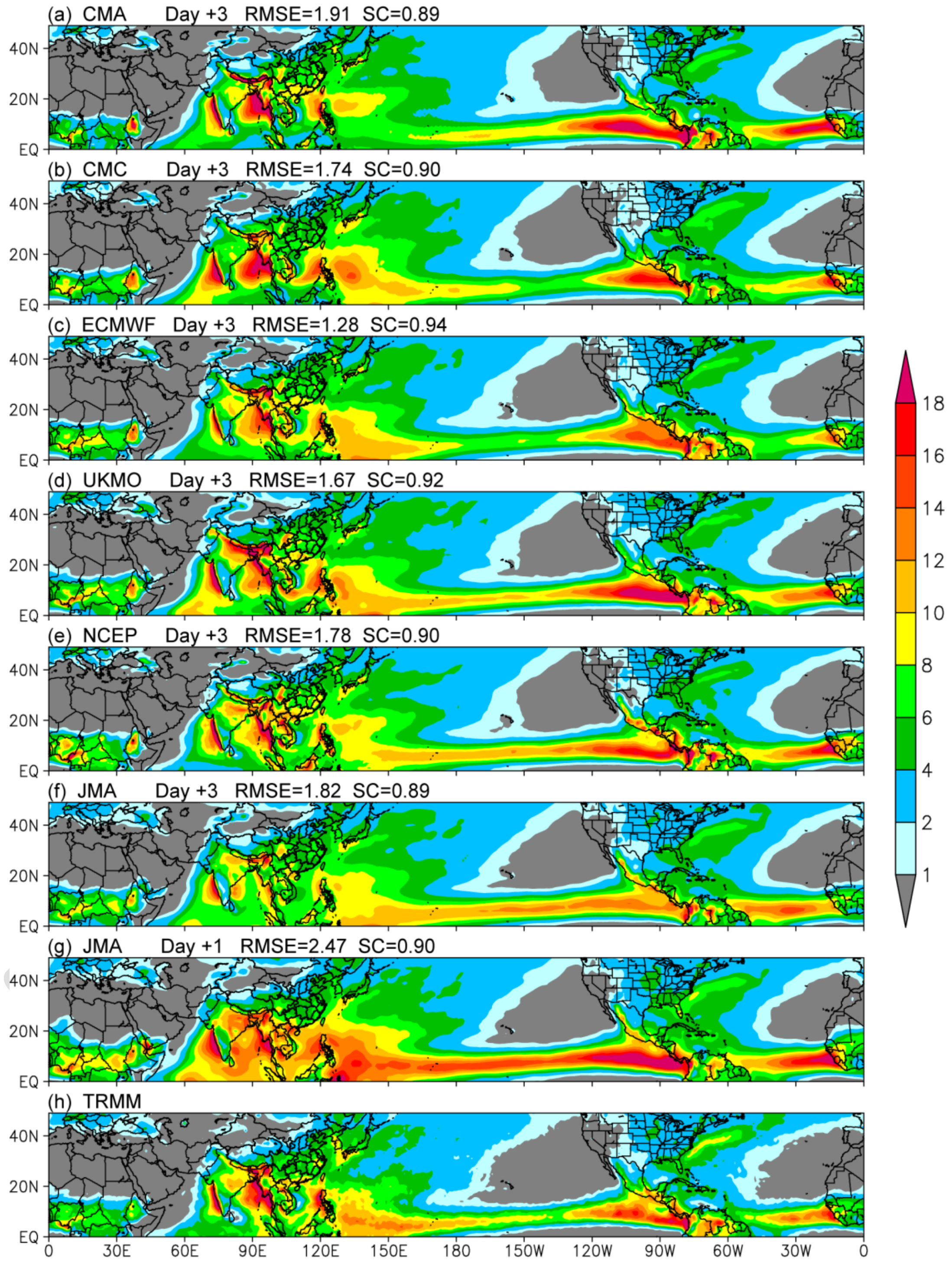

Figure 1. Average precipitation (mm day^{-1}) of ensemble mean forecasts from the six EPSs and TRMM observation during JJA 2008-2012. The RMSE (mm day^{-1}) and spatial correlation (SC) of forecast and observation averages are shown as the numbers in the titles.

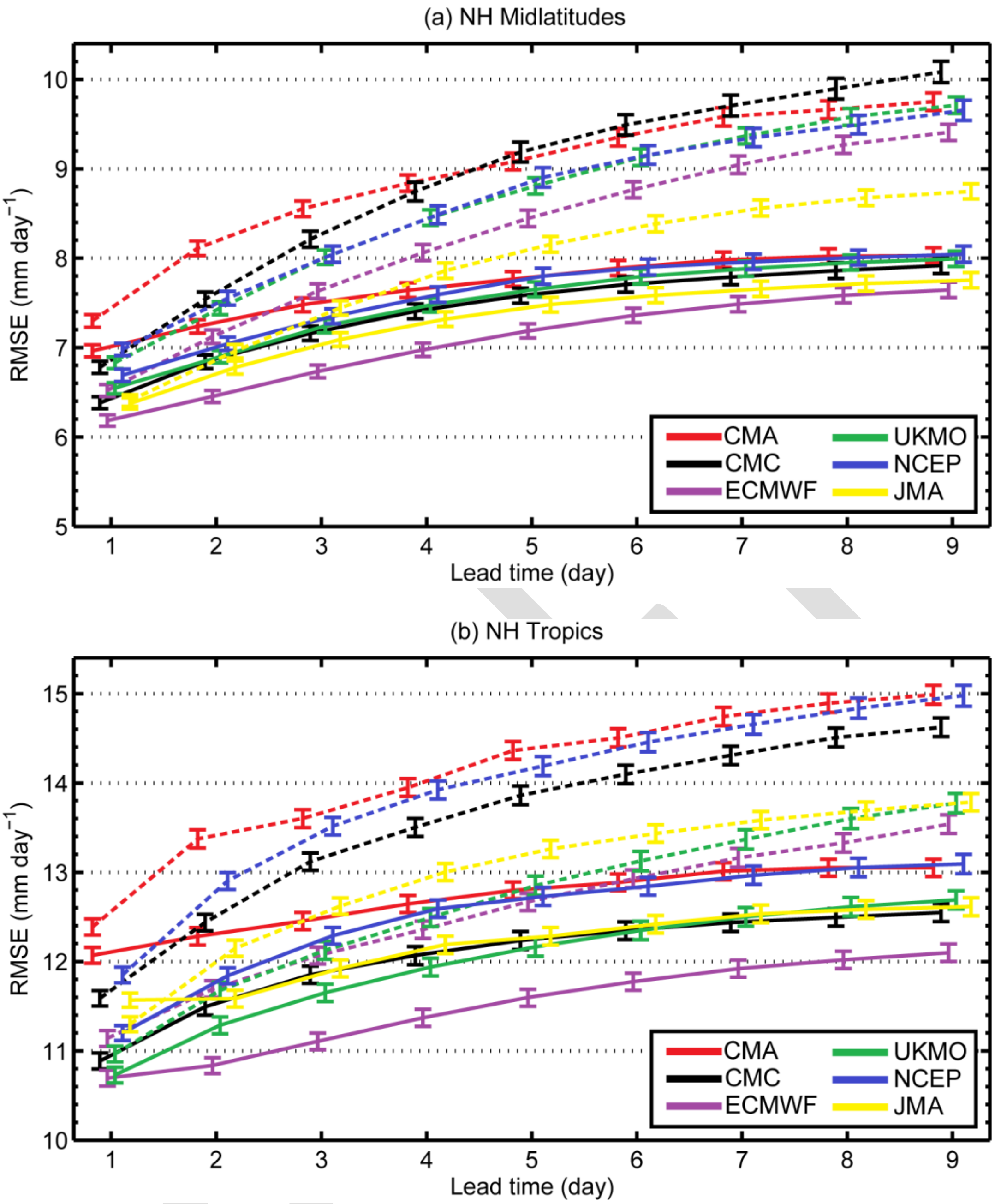

Figure 2. The RMSE of the control forecasts (dotted) and ensemble mean forecasts (solid) (mm day^{-1}) during JJA 2008-2012 in (a) the NH midlatitudes and (b) the NH tropics. Error bars represent 90% confidence intervals.

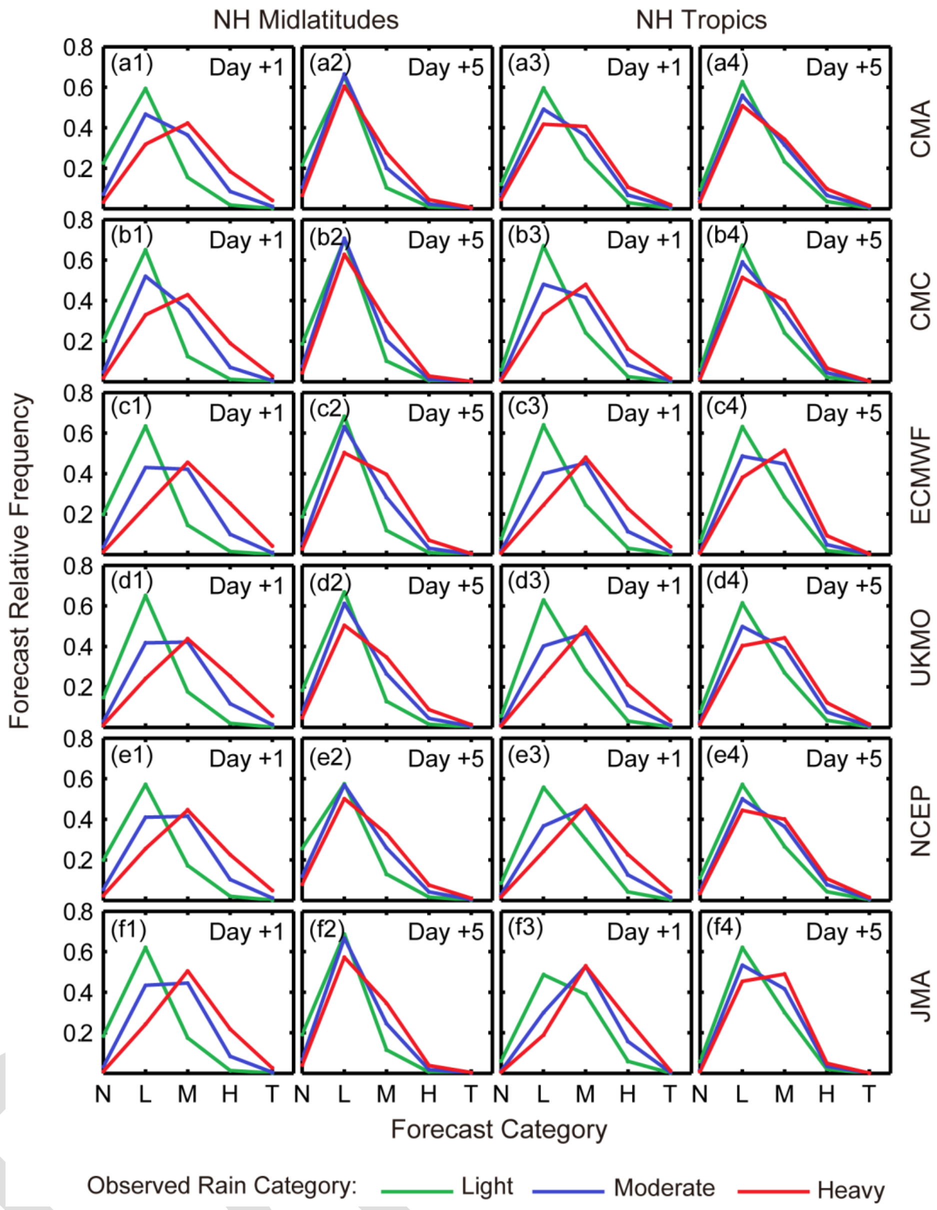

Figure 3. Discrimination diagrams of the ensemble mean QPFs in the NH midlatitudes (left two columns) and the NH tropics (right two columns) during JJA 2008-2012. The ordinate shows the forecast relative frequencies of observed light rain (1-10 mm day^{-1}, green), moderate rain (10-25 mm day^{-1}, blue), and heavy rain (25-50 mm day^{-1}, red) against five forecast categories: no rain (N, <1 mm day^{-1}), light rain (L, 1-10 mm day^{-1}), moderate rain (M, 10-25 mm day^{-1}), heavy rain (H, 25-50 mm day^{-1}) and torrential rain (T, >50 mm day^{-1}).

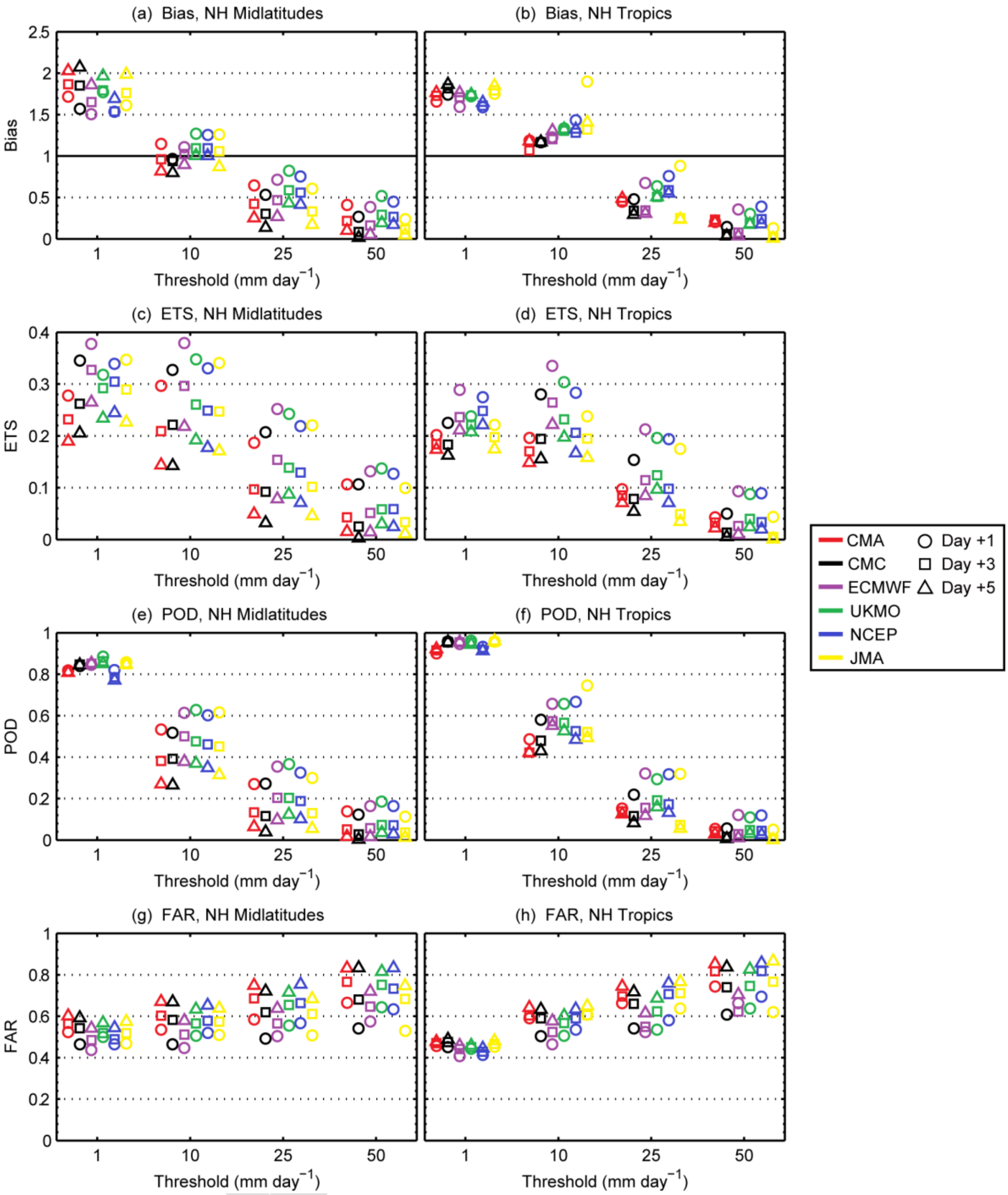

Figure 4. The Bias, ETS, POD and FAR of the ensemble mean QPFs against different precipitation thresholds for different forecast lead times (day +1, +3 and +5) during JJA 2008-2012.

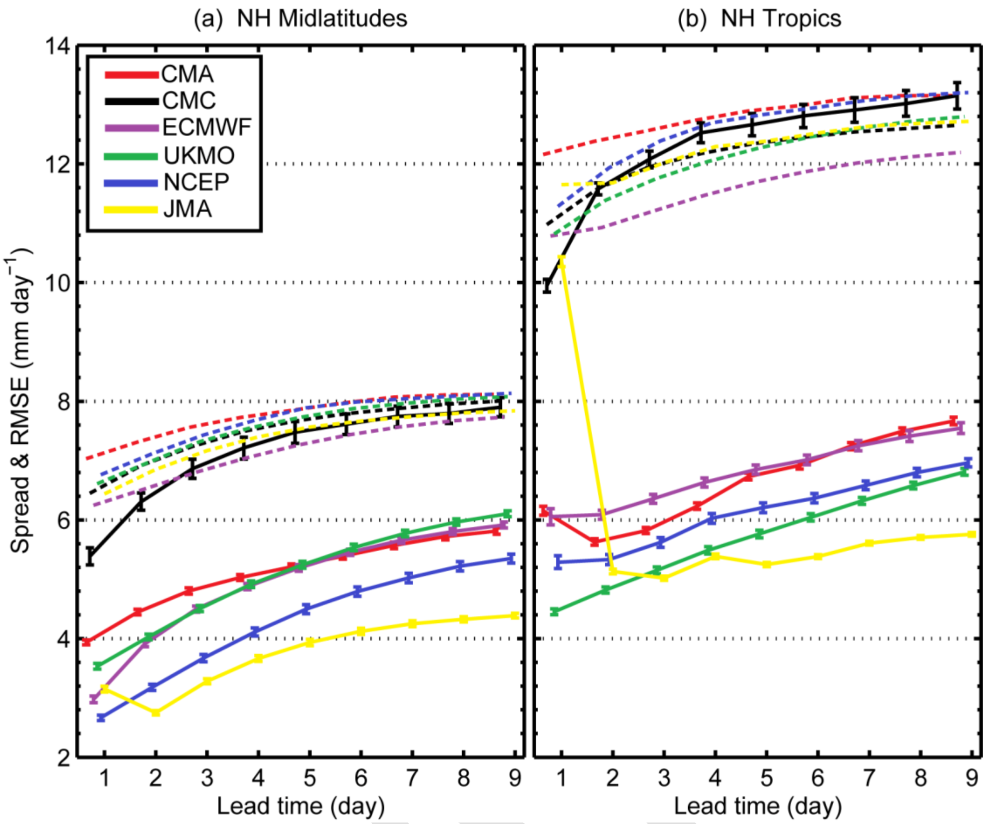

Figure 5. The RMSE of the ensemble mean QPFs (dotted) and the ensemble spread (solid) in (a) the NH midlatitudes and (b) the NH tropics during JJA 2008-2012. Error bars represent 90% confidence intervals.

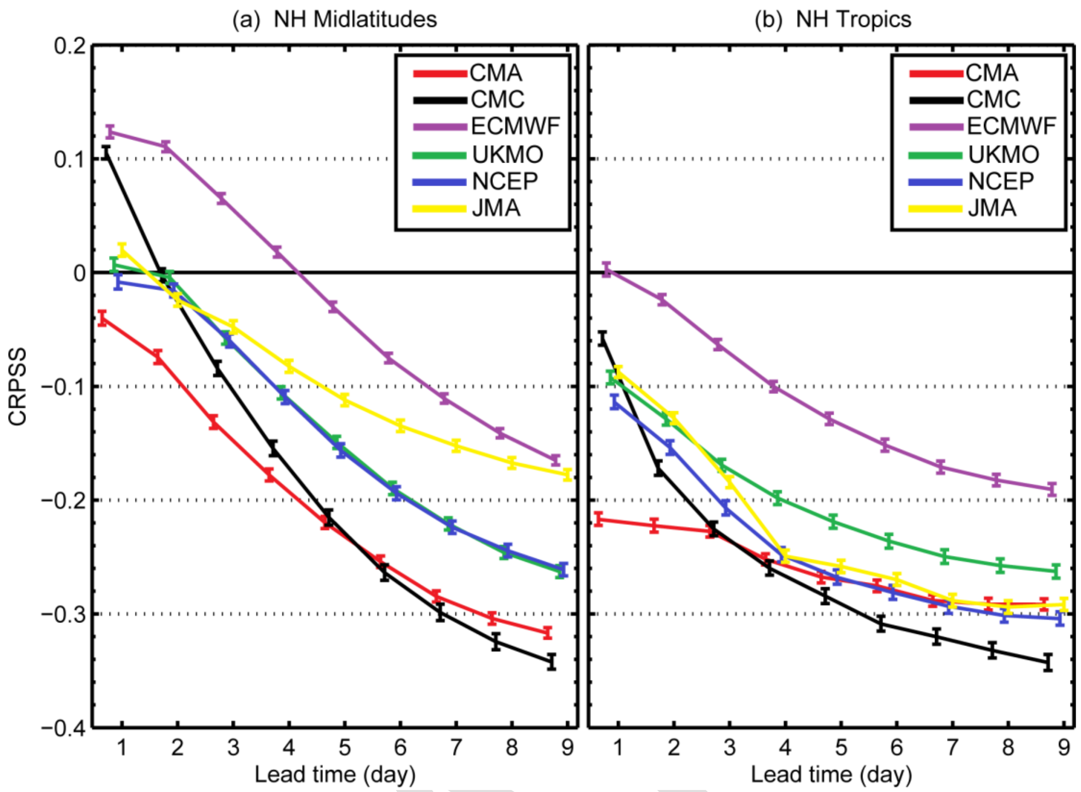

Figure 6. The CRPSS of PQPFs in (a) the NH midlatitudes and (b) the NH tropics during JJA 2008-2012. Error bars represent 90% confidence intervals.

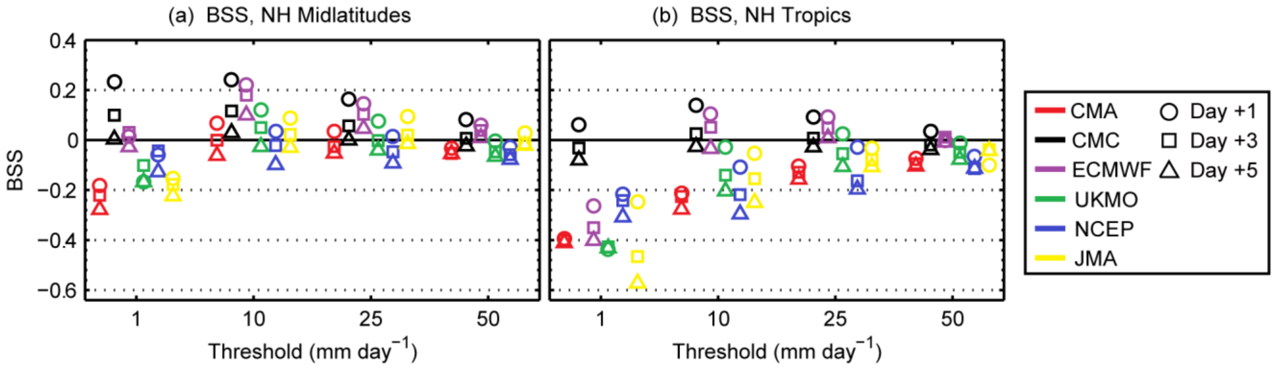

Figure 7. The BSS of PQPFs against different precipitation thresholds for different forecast lead times (day +1, +3 and +5) in (a) the NH midlatitudes and (b) the NH tropics during JJA 2008-2012.

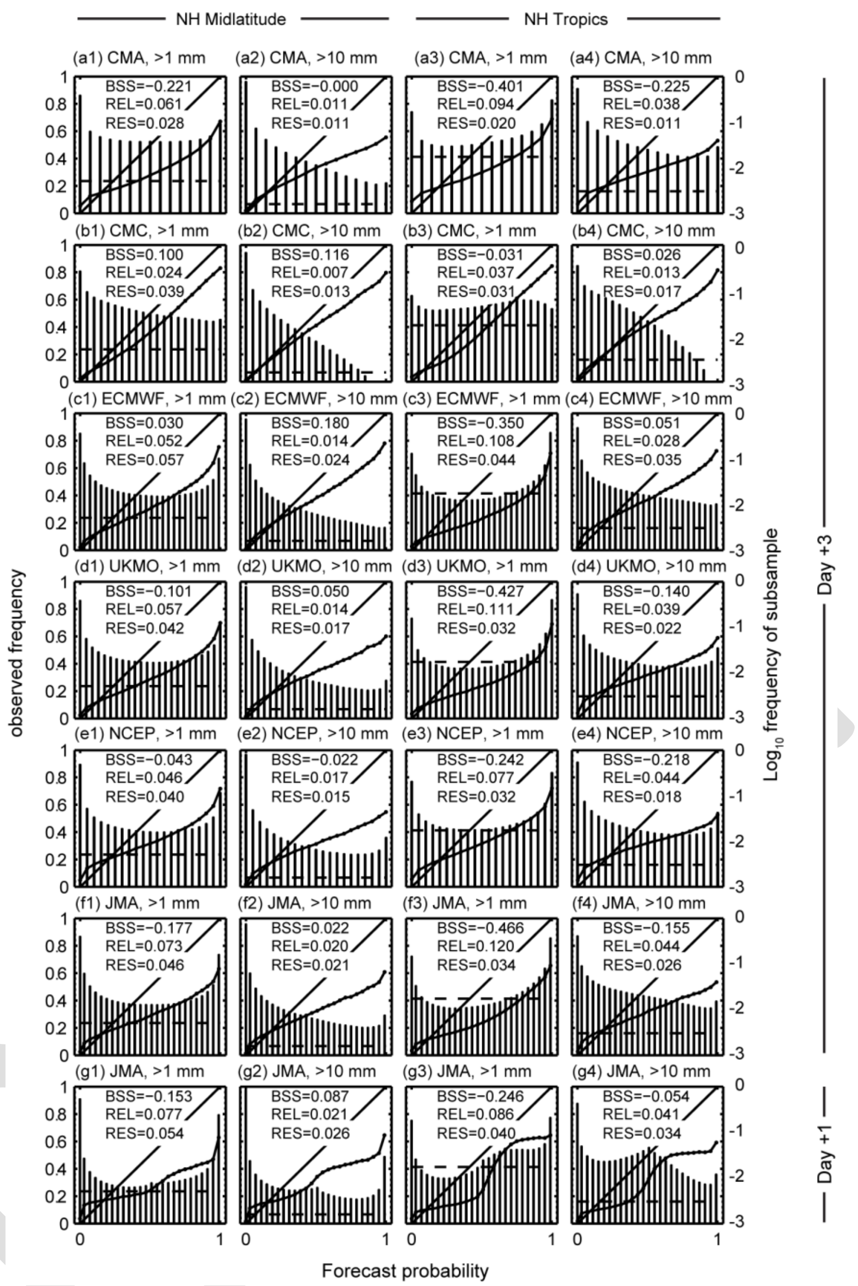

Figure 8. Reliability diagrams for day +3 and +1 PQPFs at the 1 mm day^{-1} and 10 mm day^{-1} thresholds in the NH midlatitudes (left two columns) and the NH tropics (right two columns). The bar graphs show the subsample frequencies at the logarithm scale. The BSS, and the reliability (REL) and resolution (RES) terms of the BS are shown as the numbers. For clearity, the 50 member ECMWF and JMA are converted into 26 probability bins.

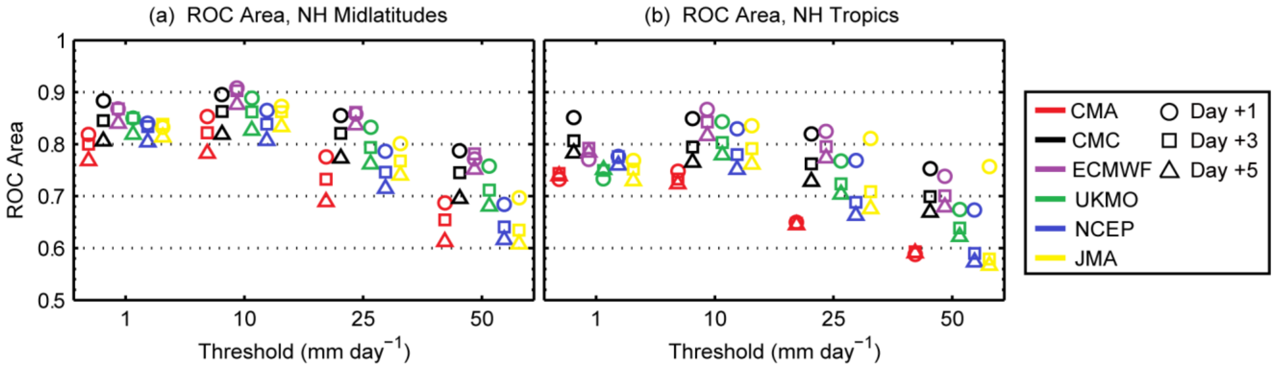

Figure 9. The area under the Relative Operating Characteristic (ROC) curve against different precipitation thresholds for different forecast lead times (day +1, +3 and +5) in (a) the NH midlatitudes and (b) the NH tropics during JJA 2008-2012.

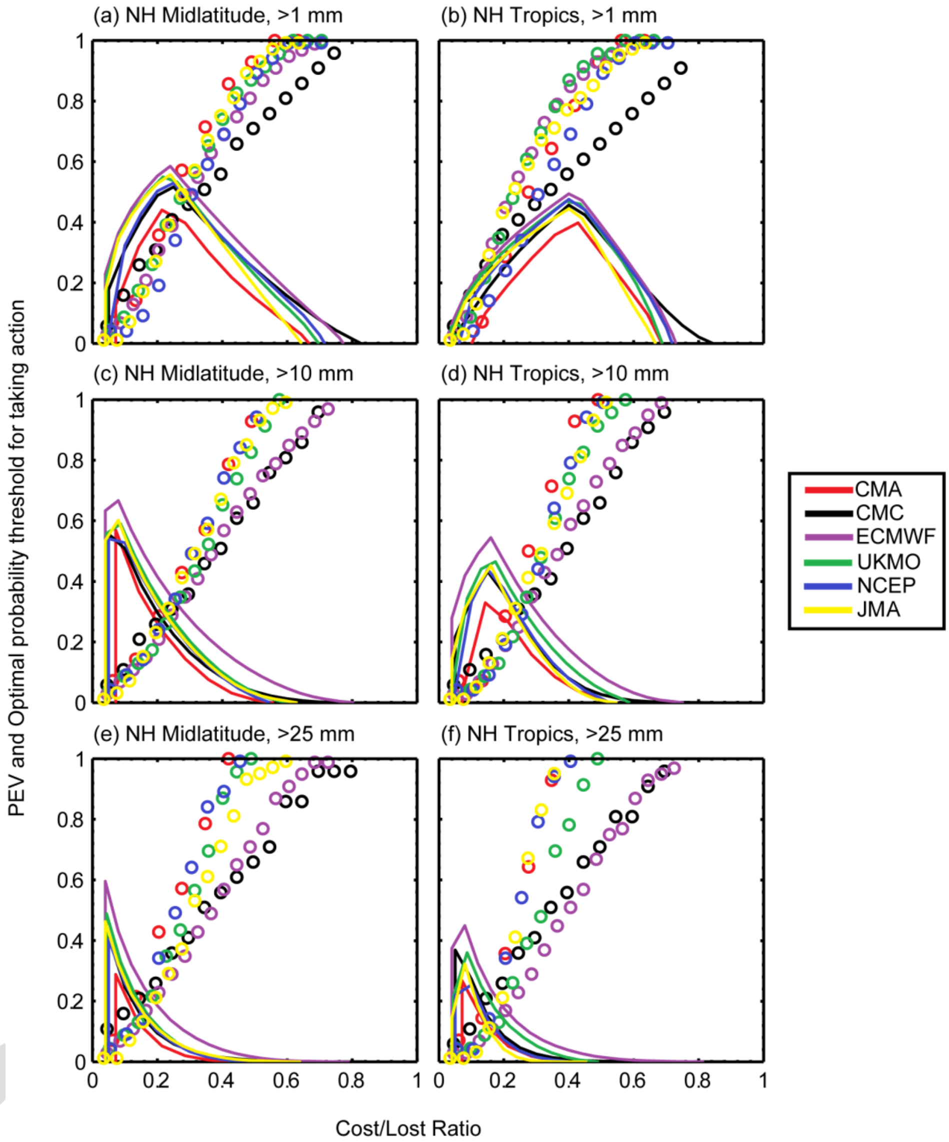

Figure 10. Potential economic value (PEV) curves and the optimal probability thresholds for taking action as a function of cost/loss ratio for day +3 PQPFs at different precipitation thresholds.

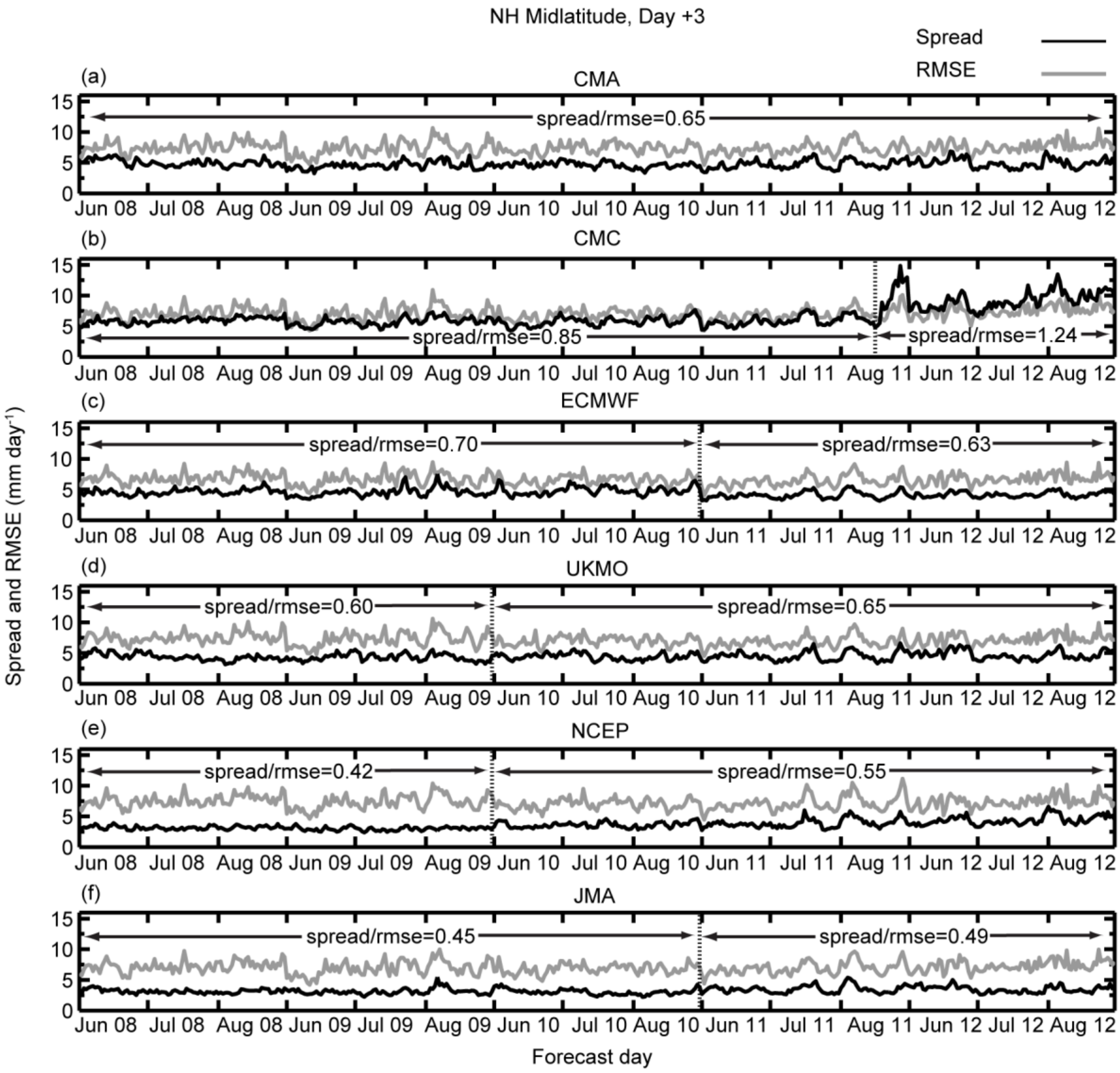

Figure 11. Time series of the ensemble spread and RMSE for the day +3 of ensemble mean QPFs in the NH midlatitudes. The dotted vertical line splits the time periods before and after the major model upgrade. The averaged ratios of the ensemble spread and RMSE during the two periods are also shown as the numbers. All changes of the spread/RMSE ratio in the five EPSs (b-f) are significant with 90% confidence interval.

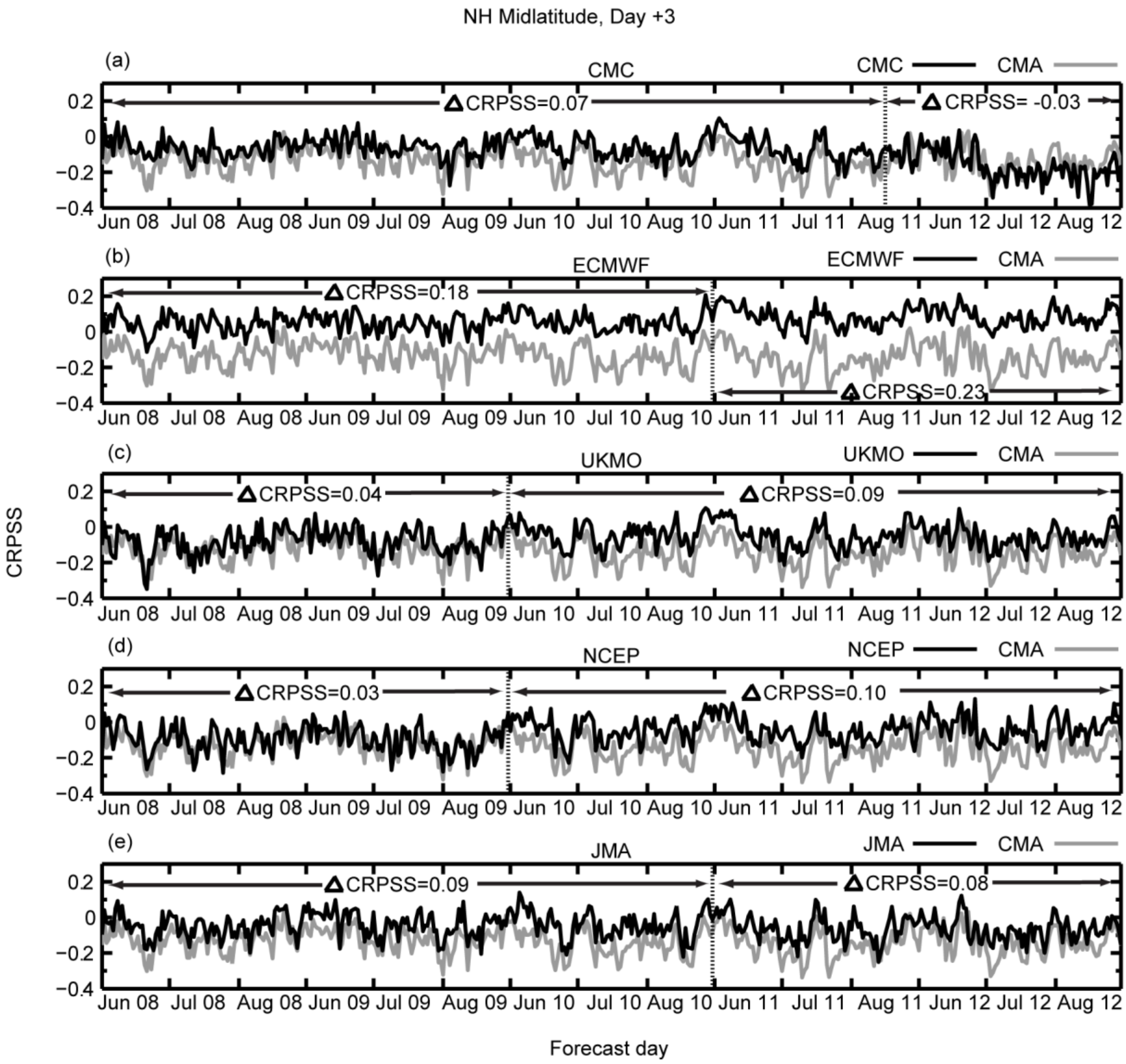

Figure 12. Time series of CRPSS for the day +3 PQPFs in the NH midlatitudes. The dotted vertical line splits the time periods before and after the major model upgrade. The CRPSS differences between each center and CMA during the two periods are also shown as the numbers. Except JMA (e), the CRPSS changes in the four EPSs (a-d) are significant with 90% confidence interval.