AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Atmospheres

RESEARCH ARTICLE
10.1002/2014JD021733

Key Points:
- The model upgrade in EPS cannot always guarantee forecast skill improvements
- The enlarged ensemble spread of CMC after the upgrade increases the QPF errors
- The day +1 QPFs from JMA have unusually large moist biases in the NH tropics

Correspondence to:
H. Yuan,
yuanhl@nju.edu.cn

Citation:
Su, X., H. Yuan, Y. Zhu, Y. Luo, and Y. Wang (2014), Evaluation of TIGGE ensemble predictions of Northern Hemisphere summer precipitation during 2008–2012, *J. Geophys. Res. Atmos.*, *119*, 7292–7310, doi:10.1002/2014JD021733.

Received 10 MAR 2014
Accepted 31 MAY 2014
Accepted article online 6 JUN 2014
Published online 26 JUN 2014

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[3,4], and Yuan Wang[1]

[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, Nanjing, China, [3]Environmental Modeling Center/NCEP/NWS/NOAA, College Park, Maryland, USA, [4]I.M. Systems Group, Inc., 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 Interactive Grand Global Ensemble (TIGGE) data set are evaluated against the Tropical Rainfall Measuring Mission observations using a series of area-weighted verification metrics during June to August 2008–2012 in the Northern Hemisphere (NH) midlatitude and tropics. Results indicate that generally the European Centre for Medium-Range Weather Forecasts 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 midlatitude than 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 Agency have remarkably large moist biases in the NH tropics, which leads to the discontinuity of forecast performance with the lead times. Performance changes due to the major EPS upgrades during the five summers are also examined using the forecasts from CMA as the reference to eliminate the interannual variation. After the EPS upgrade, CMC improves the PQPF skill at light precipitation threshold while its excessively enlarged ensemble spread increases the overall QPF and PQPF errors.

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. Compared to single forecasts, probabilistic forecasts designed to estimate the probability density function of forecast states are more valuable because they can assess the probability of various events and are more consistent than corresponding single forecasts issued on consecutive days [*Buizza*, 2008]. 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 mesoscale 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 mesoscale 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 data [*Hou et al*., 2012] 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 U.S. 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. 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, 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 with better data assimilation, more appropriate initial perturbations, improved model components, and more properly simulated model approximations, it is of interest to quantitatively analyze the improvements of QPFs and PQPFs after the EPS upgrades.

This study focuses on the 24 h accumulated ensemble mean QPFs and PQPFs generated from individual TIGGE centers in the Northern Hemisphere (NH) midlatitude and tropics to obtain a comprehensive understanding and summary of the precipitation forecast properties of six selected operational global EPSs during the recent 5 year (2008–2012) summers (June to August, JJA). The overall five-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 postprocessing and the potential use in economic decision making for the EPSs. In addition, performance changes before and after major EPS upgrades are assessed with reference 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 data sets 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 Estudios Climáticos of Brazil (CPTEC), ECMWF, the Japan Meteorological Agency (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 the ECMWF data portals. 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 included in the NH midlatitude. 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 saved in the TIGGE archive 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.

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 EPS Upgrade Time
CMA (China)	00/12	14 + 1	0.56° × 0.56°	0–10	BVs	-	-
CMC[a] (Canada)	00/12	20 + 1	1.0° × 1.0°	0–16	EnKF	PTP + SKEB multiphysics	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° × 1.25°	0–9	SVs	SPPT	17 Dec 2010
NCEP[d] (USA)	00/06/12/18	20 + 1	1.0° × 1.0°	0–16	BV-ETR	STTP	23 Feb 2010
UKMO[e] (UK)	00/12	23 + 1	0.83° × 0.56°	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 June 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° × 0.83° before 9 March 2010.

CMA uses bred vectors (BVs) [*Toth and Kalnay*, 1997] for the T213 global model (~0.5625°) [*Wang et al*., 2008] as the initial perturbations to construct the EPS, and no model uncertainties have been taken into account. Since no EPS 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 EPS 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, multiphysics 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 with the finer model horizontal grid spacing of 66 km changing from about 100 km. The model grid type has also been changed form a latitude-longitude grid to a Gaussian grid (http://collaboration.cmc.ec.gc.ca/cmc/CMOI/product_guide/docs/changes_e.html). 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, 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, e.g., 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 in 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 (30°N–90°N) while moist SVs are targeted for the tropics (20°S–30°N) [*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 0.56°, while the archived output data are on 1.25° × 1.25° 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*., 2008] 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 horizontal resolution of NCEP 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 exchange (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 (RPs) 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 $1.25° \times 0.83°$ to $0.83° \times 0.56°$.

3. Data Sets and Verification Methods

3.1. Validation Data Set

The validation data set is from the recently created Version 7 TRMM research product 3B42 (ftp://meso-a.gsfc.nasa.gov/pub/trmmdocs/3B42_3B43_doc.pdf). The data set combines multisatellite microwave-IR estimates and is adjusted by quality-controlled gauges [*Huffman et al*., 2007]. The original data set is 3-hourly and covers 50°S–50°N, 180°W–180°E, with a horizontal resolution of $0.25° \times 0.25°$. A negligible amount (0.02%) of missing TRMM data has been removed through interpolation. It is bilinearly interpolated in space and time to the $1.0° \times 1.0°$ daily (1200 UTC–1200 UTC) precipitation data, in order to compare with the TIGGE forecast data. The verification region is focused on the NH tropics (0°N–20°N) and NH midlatitude (20°N–49°N).

3.2. Forecast Data Set

The original ensemble precipitation forecast data of CMA, ECMWF, UKMO, and JMA are converted onto the same $1.0° \times 1.0°$ grid as CMC and NCEP before downloading, using the bilinear interpolation software provided by the ECMWF TIGGE data portal. Whole perturbed members (the control forecast is not included because the control forecast has different performance with perturbed members, figures not shown) of each center are used to compute 24 h ensemble mean QPFs and PQPFs. Only the +1 to +9 day forecasts initialized at 1200 UTC are examined due to the limit of the JMA forecast data. The time period of the verification covers JJA 2008–2012 (1 June to 30 August, total $91 \times 5 = 455$ days). Several 1200 UTC cycles of the NCEP forecast data are missing, including the dates of 8, 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 available.

After processing the TIGGE data to the 24 h accumulated precipitation forecasts (usually taking subtraction from the accumulated total precipitation), there are some negative values (https://software.ecmwf.int/wiki/display/EMOS/Precipitation) during the five summers: a small portion (1.0%–3.6%) of negligible values (-1–$0\ \mathrm{mm\ d^{-1}}$) due to the scaling of values during GRIdded Binary (GRIB) packing or interpolation errors, and a very low fraction of large negative values for CMC (0.0034%, $< -1\ \mathrm{mm\ d^{-1}}$) after EPS upgrade, which may be generated during the interpolation from a Gaussian grid to a latitude-longitude grid. All negative values of 24 h precipitation forecasts are set to zeros here.

3.3. Verification Methods

In this study, multiple deterministic and probabilistic verification methods are carried out to demonstrate different aspects of QPFs and PQPFs, including the root-mean-square error (RMSE), spatial correlation, discrimination diagram, bias score (frequency bias, Bias), equitable threat score (ETS), probability of detection (POD), false alarm ratio (FAR), spread-skill relationship (spread versus RMSE), continuous ranked probability skill score (CRPSS), Brier skill score (BSS), reliability diagram, area under the relative operating characteristic curve (ROCA), and potential economic value (PEV). 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 detailed calculation of the area-weighted scores can be found in Appendix A.

Error bars are shown for the RMSE, ensemble spread, and CRPSS, representing the 95% confidence intervals using bootstrapping method by randomly selecting the statistics 10,000 times [*Hamill*, 1999]. The referenced continuous ranked probability score (CRPS) is generated using the cumulative distribution function (CDF) of the observed samples (i.e., sample climatology) on each grid point. Similarly, the referenced Brier Score (BS) is

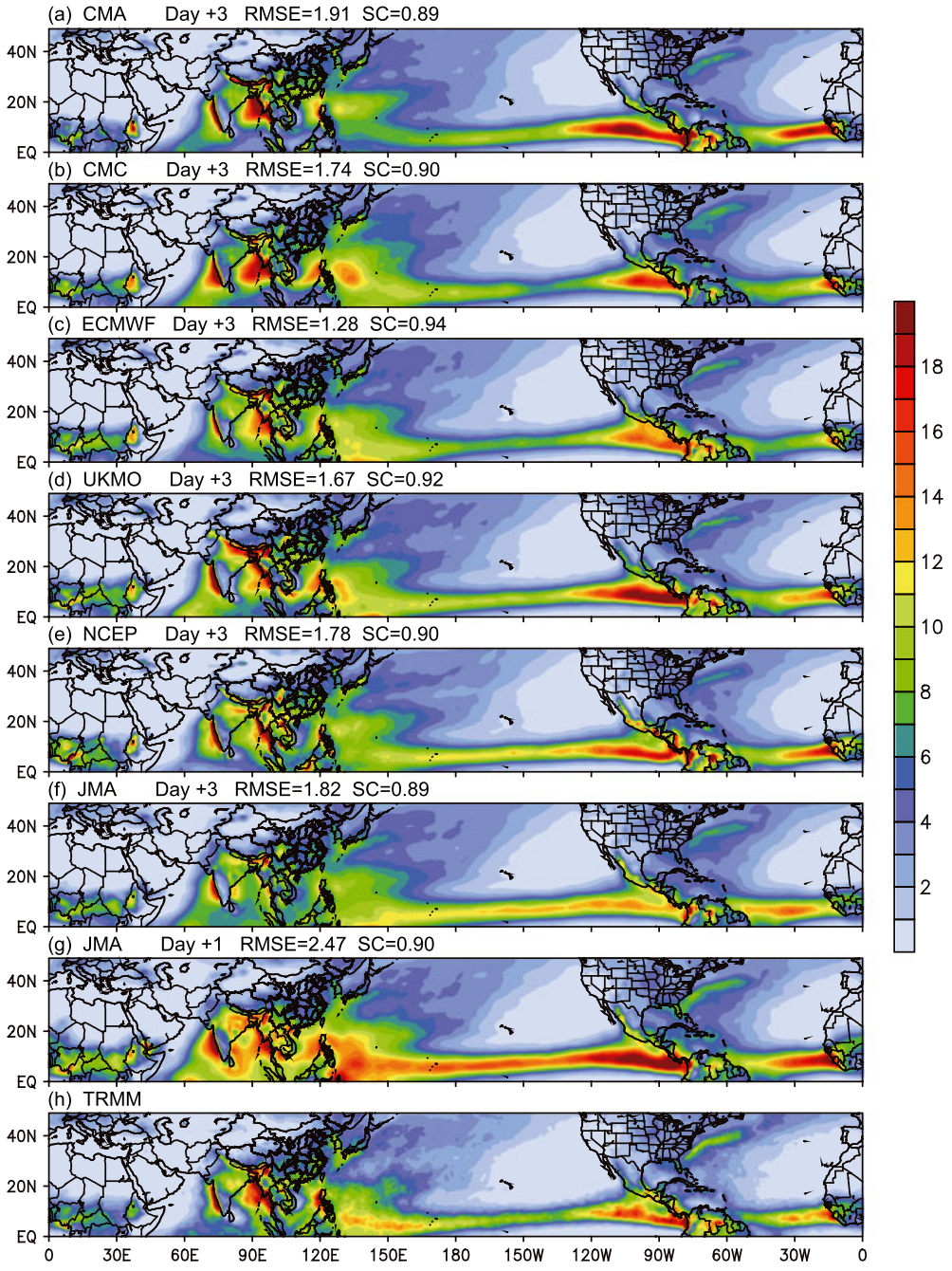

Figure 1. Average precipitation (mm d^{-1}) of ensemble mean forecasts from the six EPSs and TRMM observation during JJA 2008–2012. The RMSE (mm d^{-1}) and spatial correlation (SC) of forecast and observation averages are shown as the numbers in the titles.

generated using the sample climatology frequency on each grid point. Thus, 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, resolution, and uncertainty [*Murphy*, 1973]. The reliability term (i.e., the conditional bias) can be calibrated through statistical postprocessing, while the resolution term can only be improved through upgrading model dynamics, physics, and configurations. One unique method is the analog postprocessing using long-term reforecast data [*Hamill et al.*, 2006], which improves the resolution term through selecting historical observation data as the ensemble forecasts.

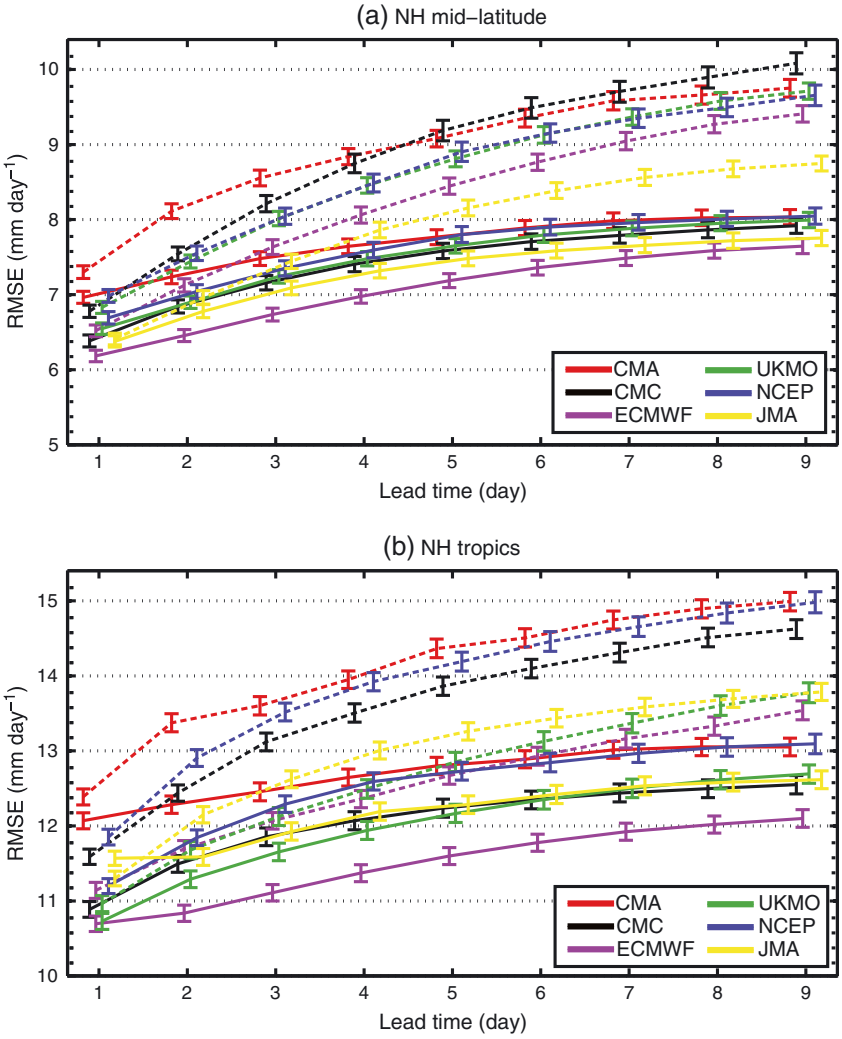

Figure 2. The RMSE of the control forecasts (dotted) and ensemble mean forecasts (solid) (mm d^{-1}) during JJA 2008–2012 in (a) the NH midlatitude and (b) the NH tropics. Error bars represent 95% confidence intervals.

Also, the impacts of major EPS upgrades on the forecast performance are examined. To eliminate the impact of interannual variation, the score changes of other five centers due to the major EPS upgrade are compared with the corresponding score of CMA (frozen version). Considering 95% confidence intervals, the performance change of the forecast score due to the major EPS 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, and (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 (Figures 1a–1f) 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 d^{-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

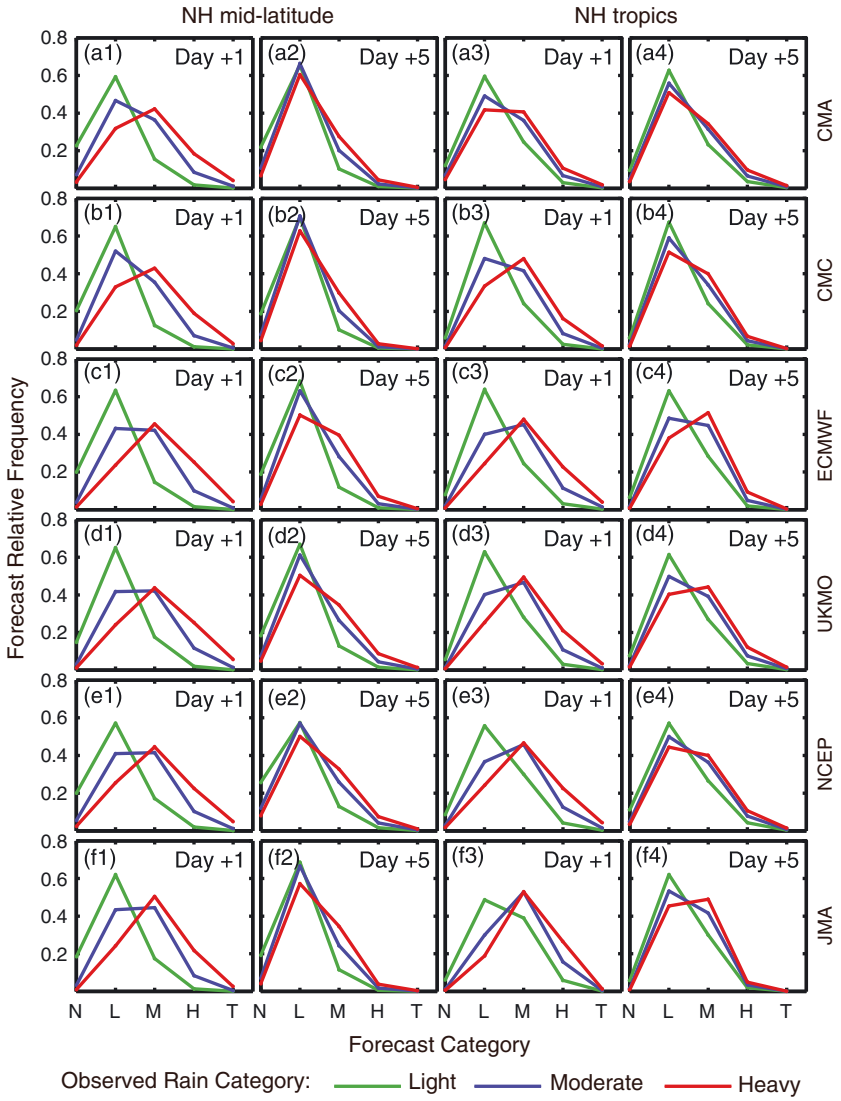

Figure 3. Discrimination diagrams of the ensemble mean QPFs in the NH midlatitude (left two columns) and the NH tropics (right two columns) during JJA 2008–2012. The ordinate shows the forecast relative frequencies of observed light rain (1–10 mm d^{-1}, green), moderate rain (10–25 mm d^{-1}, blue),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

of 1.28 mm d^{-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, JMA significantly outperforms others in the NH midlatitude, especially for longer lead times, while the ECMWF, UKMO, and JMA EPSs have less forecast errors than other three centers in the NH tropics. The probability distribution functions of the control forecasts from each center indicate that JMA underforecasts much more heavy rain amount than other centers, especially for longer lead times (figures not shown). This indicates that only using RMSE will not reflect the true precipitation forecast performance. 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

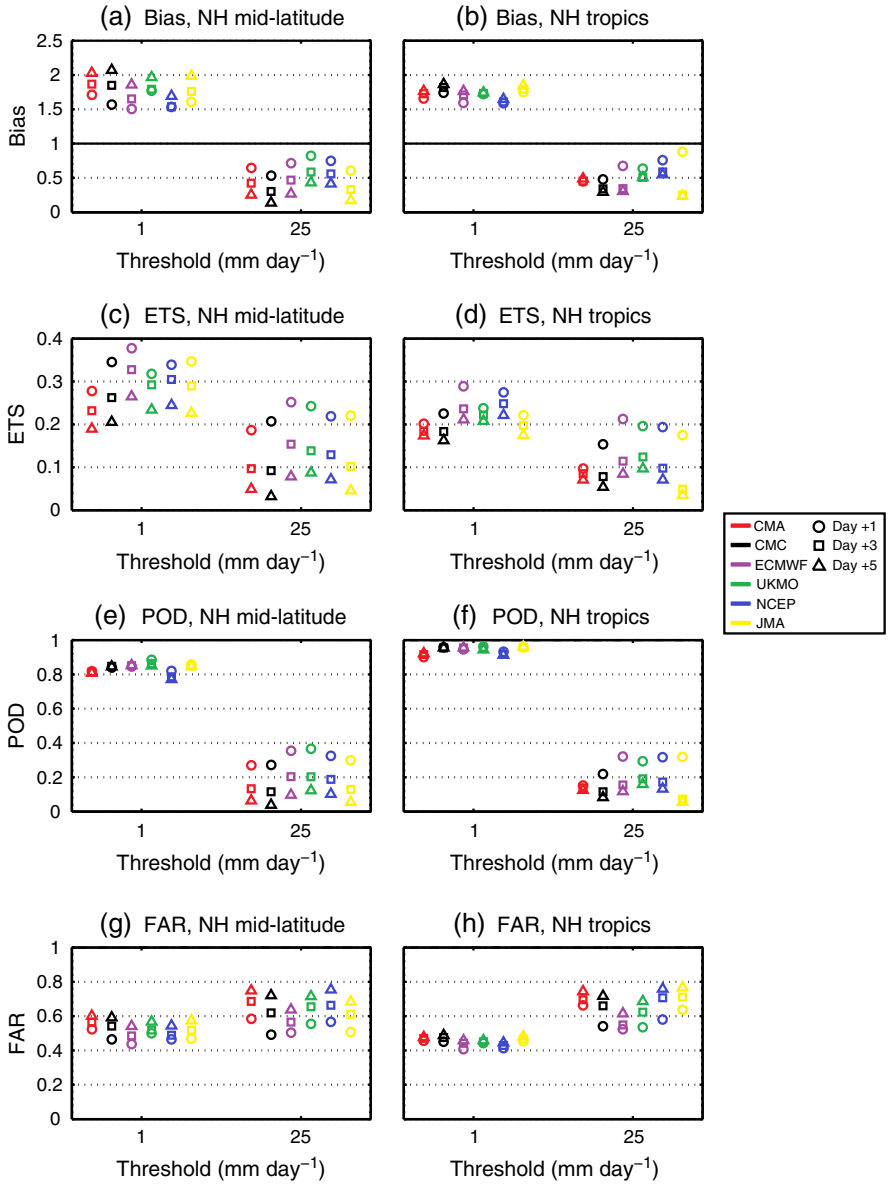

Figure 4. The (a) Bias, (b) ETS, (c) POD, and (d) FAR of the ensemble mean QPFs against different precipitation thresholds for different forecast lead times (day +1, +3, and +5) during JJA 2008–2012.

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 modelers. The discrimination ability decreases with the lead time

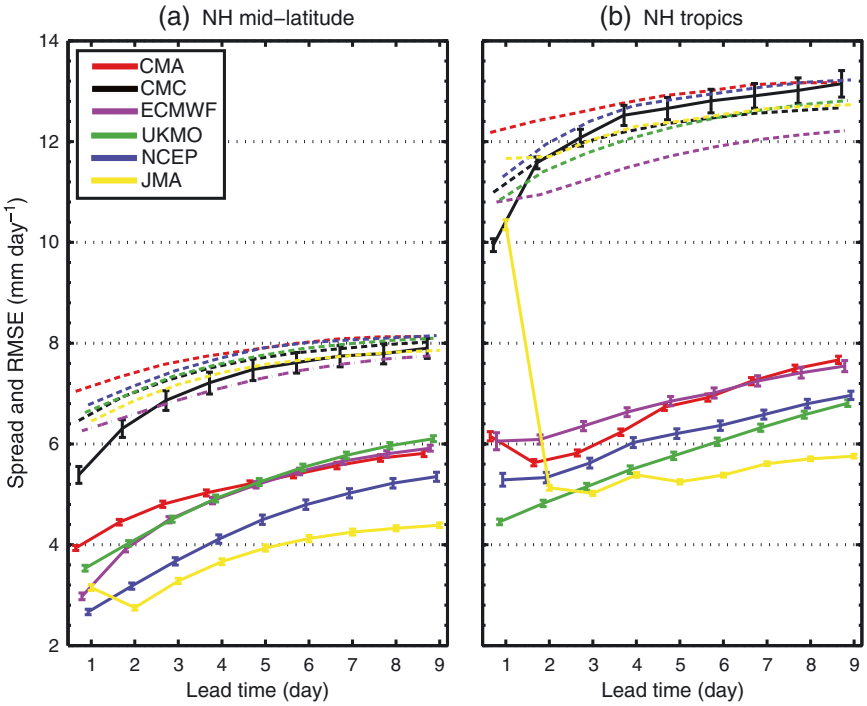

Figure 5. The RMSE of the ensemble mean QPFs (dotted) and the ensemble spread (solid) in (a) the NH midlatitude and (b) the NH tropics during JJA 2008–2012. Error bars represent 95% confidence intervals.

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 toward 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 different lead times and precipitation thresholds (Figure 4). In both the NH midlatitude and the NH tropics, all EPSs overforecast the light precipitation ($>1\,\mathrm{mm\,d^{-1}}$) and underforecast the heavier precipitation ($>25\,\mathrm{mm\,d^{-1}}$, Figures 4a and 4b). Generally, ECMWF demonstrates the best forecast quality (ETS, Figures 4c and 4d), while NCEP has the relatively good bias score (Figures 4a and 4b). The selected scores are linked, such as the existing relation of Bias = POD/(1-FAR). Accordingly, the relatively lower POD (Figures 4e and 4f) and lower FAR (Figures 4g 4h) 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 multiphysics 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 midlatitude while becoming slightly overdispersive in the NH tropics. Other five EPSs are severely underdispersive 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

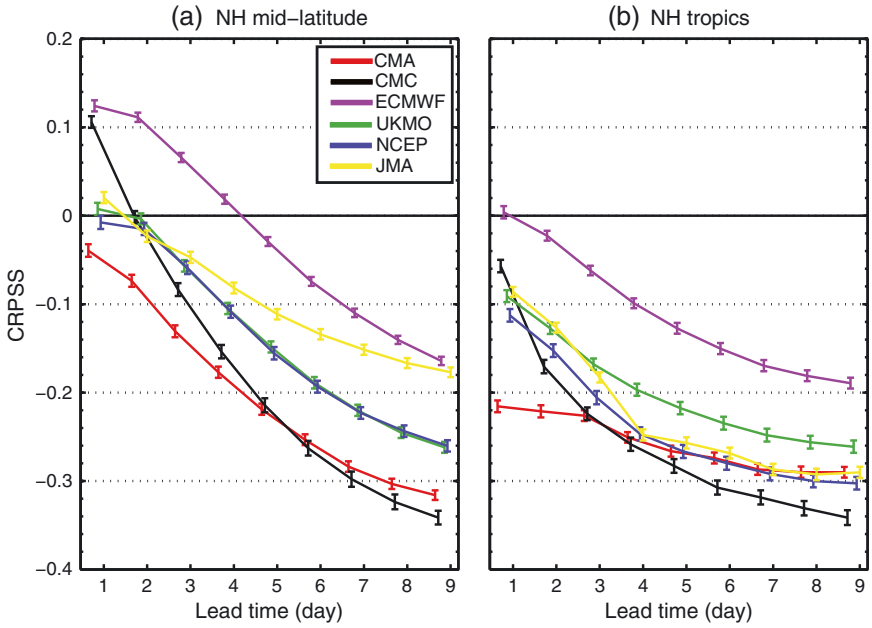

Figure 6. The CRPSS of PQPFs in (a) the NH midlatitude and (b) the NH tropics during JJA 2008–2012. Error bars represent 95% confidence intervals.

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 midlatitude 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 midlatitude. 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

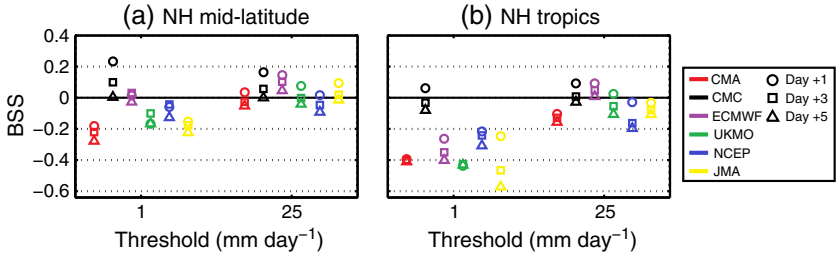

Figure 7. The BSS of PQPFs against different precipitation thresholds for different forecast lead times (day +1, +3, and +5) in (a) the NH midlatitude and (b) the NH tropics during JJA 2008–2012.

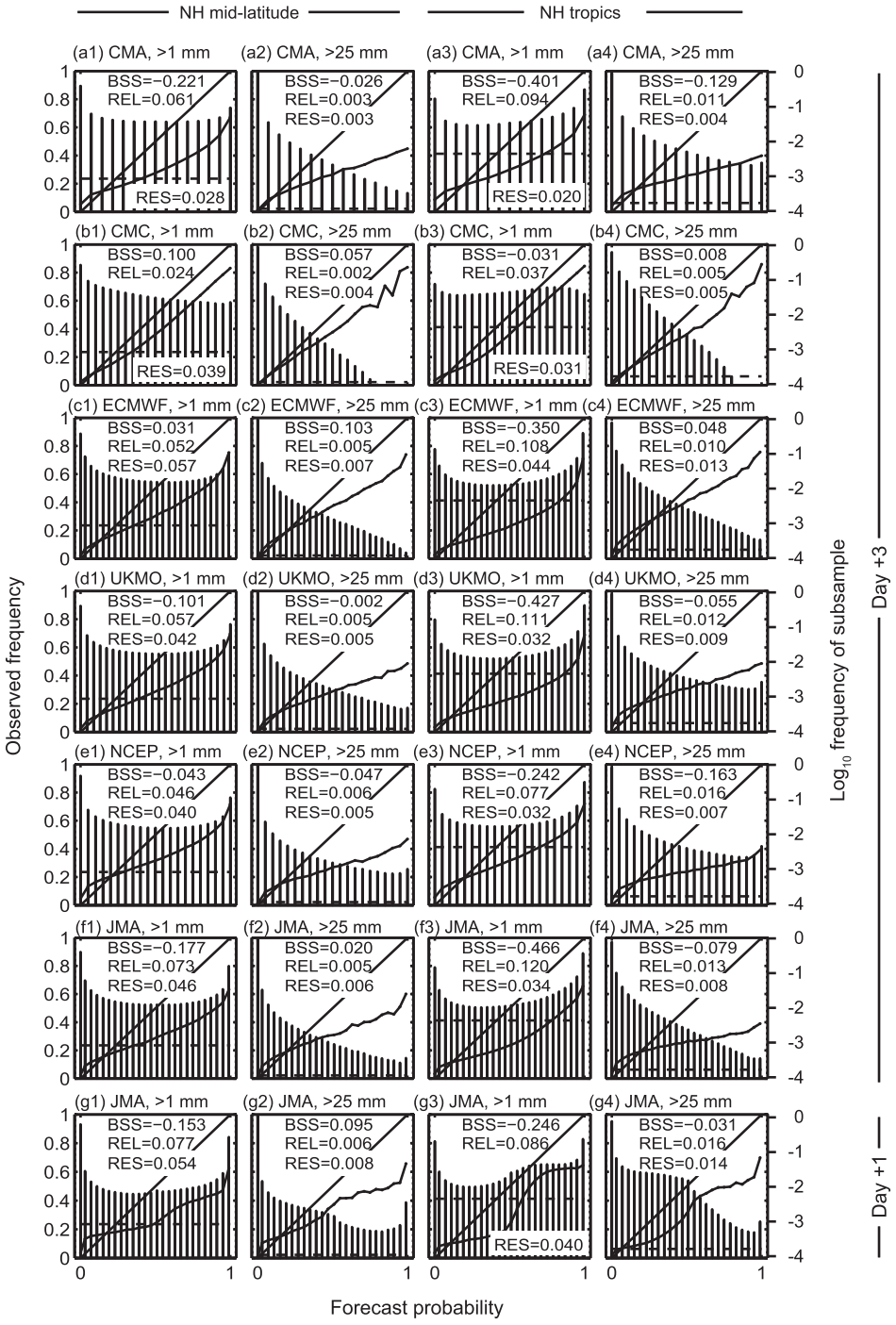

Figure 8. Reliability diagrams for day +3 and +1 PQPFs at the 1 mm d^{-1} and 25 mm d^{-1} thresholds in the NH midlatitude (left two columns) and the NH tropics (right two columns). The bar graphs show the subsample frequencies at the logarithm scale. The horizontal dash line represents the observed sample frequency. The BSS, and the reliability (REL), and resolution (RES) terms of the BS are shown as the numbers. For clarity, the 50 member ECMWF and JMA are converted into 26 probability bins.

obviously outperforms other centers at the 1 mm d^{-1} threshold, and CMC and ECWMF are more skillful at the 25 mm d^{-1} precipitation threshold. In addition, the BSS varies with the precipitation threshold and is sensitive to the conditional bias. ECMWF has the relatively low BSS at 1 mm d^{-1} in the NH tropics due to the poor reliability (Figure 8c3). The good reliability of CMC and the good resolution of ECMWF (Figures 8b1–8b4 and 8c1–8c4) contribute to higher BSSs in both EPSs. The conditional bias (reliability term) can be calibrated through

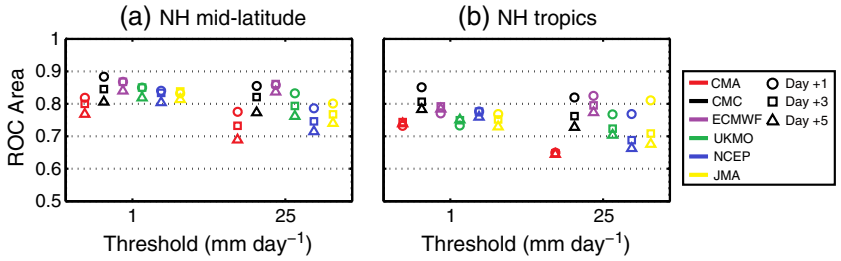

Figure 9. The area under the relative operating characteristic curve (ROCA) against different precipitation thresholds for different forecast lead times (day +1, +3, and +5) in (a) the NH midlatitude and (b) the NH tropics during JJA 2008–2012.

postprocessing, while the resolution term is associated with the model itself and difficult to be postprocessed. At the 1 mm d^{-1} threshold, the resolution terms (Figures 8d1, 8d3, 8e1, and 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 25 mm d^{-1} threshold, both the reliability and resolution terms of UKMO are better than or equal to those of NCEP (Figures 8d2, 8d4, 8e2, and 8e4), which leads to better BSSs of UKMO (Figure 7).

The reliability diagrams of day +3 PQPFs at the 1 mm d^{-1} and 25 mm d^{-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 (Figures 8b1–8b4) 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 1 in 10,000) (Figures 8b2 and 8b4). In contrast, UKMO and NCEP are sharp, with more forecasts of extreme probabilities (Figures 8d1–8d4 and 8e1–8e4). The day +3 PQPFs from CMA have the worst resolution (smallest RES), while those from JMA and NCEP demonstrate the worst reliability (largest REL) for the 1 mm d^{-1} and 25 mm d^{-1} thresholds respectively. This indicates the relatively poorer model quality of CMA and larger conditional biases of JMA and NCEP (for light and heavy rain events respectively). In particular, for the day +1 PQPFs from JMA in the NH tropics (Figures 8g3 and 8g4), 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.

4.2.3. Discrimination Ability and Potential Economic Value

Compared with the reliability diagram, which is conditioned on the forecasts, the relative operating characteristic curve measures the discrimination ability of probabilistic forecasts conditioned on the observations. ROCA is usually 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. *Buizza et al.* [1999a] consider a ROCA of 0.7 as the limit of a useful prediction system. Figure 9 indicates that nearly all centers are useful for the day +1 to +5 lead times at the 1 and 25 mm d^{-1} precipitation thresholds in the NH midlatitude while some centers lack skill at the 25 mm d^{-1} precipitation threshold in the NH tropics. 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 a simple cost-loss model [*Zhu et al.*, 2002], the economic value (EV) here refers to a relative skill score (not actual economic loss). The EV compares the economic loss from the decision making, which is 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. The 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]. Figure 10 demonstrates the PEV for day +3 PQPFs at the 1 mm d^{-1} and 25 mm d^{-1} precipitation thresholds. Except the high PEV values of CMC at 1 mm d^{-1} precipitation threshold for high C/L ratio users, ECMWF has the highest PEV values. PQPFs from ECMWF outperform other centers more at the 25 mm d^{-1} precipitation threshold, 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–10c). 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 d^{-1} precipitation threshold, indicating the best reliability [*Jolliffe and Stephenson*, 2003]. The optimal probability thresholds of ECMWF are close to the diagonal line at the 25 mm d^{-1}

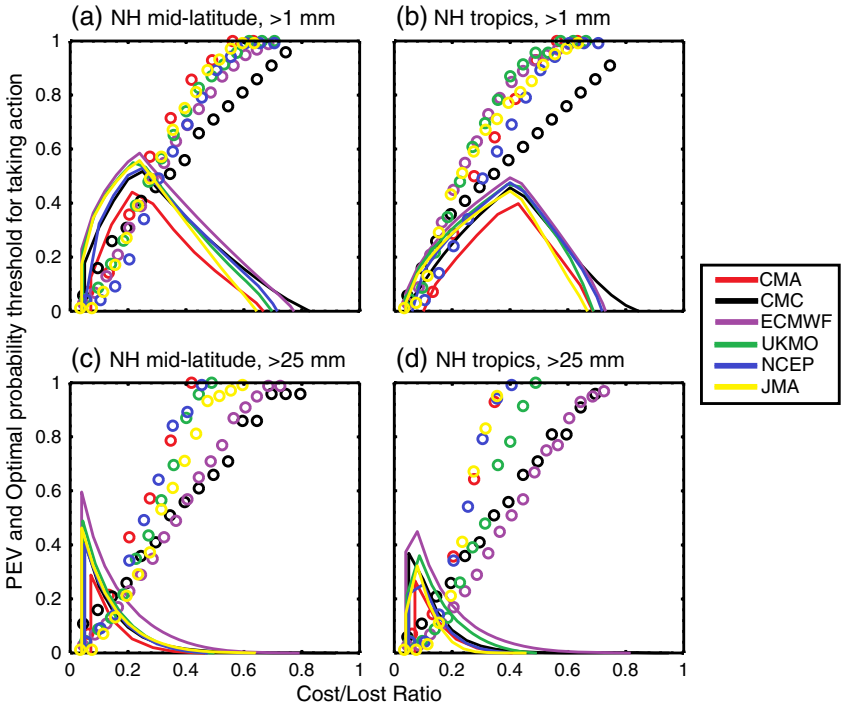

Figure 10. Potential economic value (PEV) curves and the optimal probability thresholds for taking action as a function of cost/loss ratio for day +3 PQPFs at different precipitation thresholds.

precipitation threshold but largely deviate from the diagonal line at the 1 mm d^{-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 EPS Upgrade

One aim 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 EPS upgrade. All the five centers have significant spread changes with 95% 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 d^{-1} in the NH midlatitude and the NH tropics respectively. All the five centers have significantly changed the day +3 spread/RMSE ratios except JMA in the NH tropics (Table 2). 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 EPS upgrades are similar with those of the day +3 forecasts.

Upgrading the EPS is expected to improve ensemble mean QPFs. RMSEs (Table 2) of the day +3 ensemble mean QPFs from UKMO are reduced significantly while there are no significant RMSEs changes for ECMWF after the major EPS upgrade. The reason is that UKMO has relatively larger RMSE before the EPS upgrade and achieves significant improvement in RMSE, while the RMSE of ECMWF is quite small compared to other centers and is hard to be improved further. Notably, CMC has increased the RMSE after the EPS upgrade, because oversized ensemble spread (Table 2) usually causes large forecast errors. Table 3 demonstrates the changes of RMSE and ETS of the ensemble mean QPFs at different lead times and precipitation thresholds before and after the major EPS upgrade. RMSEs of UKMO are significantly reduced for different lead times. RMSEs of CMC deteriorate after the EPS upgrade for most of the lead times while the 1 mm d^{-1} ETSs in the NH tropics are improved for day +2 to +9. This is because ETS is a dichotomous forecast score associated with the selected precipitation threshold and is insensitive to ensemble spread. ECMWF and UKMO have improved the 10 mm d^{-1} ETSs in the NH tropics after EPS upgrade. NCEP has improved the ETSs at heavier thresholds over 10 mm d^{-1} (except 25 mm d^{-1} ETS in the NH tropics).

Table 2. Average Ensemble Spread (mm d^{-1}), RMSE (mm d^{-1}), and Spread/RMSE Ratio of Five Centers and Their Differences With CMA for the Day +3 Forecasts Before and After the Major EPS Upgrade[a]

		NH Midlatitude			NH Tropics		
Score	Center	Before	After	Change	Before	After	Change
Spread	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**	−1.0	−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**
RMSE	CMC	7.0	7.5	**0.5**	11.6	12.3	**0.7**
	ECMWF	6.7	6.6	−0.1	11.1	11.0	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	**0.1**
	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	**−0.7**
	NCEP-CMA	0	−0.2	**−0.2**	0.2	−0.4	**−0.6**
	JMA-CMA	−0.5	−0.3	**0.2**	−0.5	−0.5	0
Spread/RMSE	CMC	0.85	1.24	**0.40**	0.97	1.19	**0.22**
	ECMWF	0.70	0.63	**−0.07**	0.63	0.49	**−0.14**
	UKMO	0.60	0.65	**0.05**	0.41	0.46	**0.05**
	NCEP	0.42	0.55	**0.13**	0.38	0.50	**0.12**
	JMA	0.45	0.49	**0.04**	0.42	0.43	0.01
	CMC-CMA	0.20	0.59	**0.39**	0.49	0.74	**0.25**
	ECMWF-CMA	0.05	−0.02	**−0.07**	0.15	0.04	**−0.10**
	UKMO-CMA	−0.05	0	**0.05**	−0.07	0	**0.08**
	NCEP-CMA	−0.23	−0.10	**0.13**	−0.10	0.04	**0.15**
	JMA-CMA	−0.19	−0.16	**0.03**	−0.06	−0.02	**0.04**

[a]Boldface represents the significant change with 95% confidence interval.

At the same time, the PQPFs are expected to be improved when the ensembles are upgraded. All the centers have significantly changed CRPSS for the day +3 PQPFs except JMA (Figures 11a–11e). The CRPSSs of ECMWF, UKMO, and NCEP are improved significantly after major EPS upgrades as the gaps between the two time series become larger (Figures 11b–11e). However, the CRPSS of CMC becomes even lower than that from the static version of CMA after the EPS upgrade (Figure 11a). The deterioration of CRPSS of CMC is probably due to its remarkably increased ensemble spread. Situations in the NH tropics and of other lead times are similar as those of the day +3 PQPFs (Table 3). Unlike CRPSS that more depends on precipitation amount, the BSS is sensitive to the selected precipitation threshold. The 1 mm d^{-1} BSS of CMC has been significantly improved despite of its low CRPSS. Generally, the PQPF skill (CRPSS and BSS) of NCEP has been improved, while the PQPF skill of JMA has not been changed much after the EPS upgrade. ECMWF has improved BSSs at light to moderate thresholds, while UKMO has improved those at moderate to heavy thresholds.

5. Summary and Discussions

This study provides a comprehensive verification of ensemble mean QPFs and PQPFs from six operational global EPSs in the NH midlatitude 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

Table 3. The Forecast Lead Times With Significant Changes of the Ensemble Spread, Spread/RMSE Ratio, RMSE, ETS, CRPSS, and BSS due to the Major EPS Upgrade With 95% Confidence Interval[a]

Score	NH Region	CMC	ECMWF	UKMO	NCEP	JMA
SPREAD	midlatitude	1–9 ↑	2–9 ↓	3–6 ↑	1–9 ↑	2–9 ↑
	tropics	1–9 ↑	2–9 ↓	3–6 ↑	1–9 ↑	3–9 ↑
SPREAD/RMSE	midlatitude	1–9 ↑	1 ↑ 2–9 ↓	1–9 ↑	1–9 ↑	2–9 ↑
	tropics	1–9 ↑	2–9 ↓	1–9 ↑	1–9 ↑	5,6,8,9 ↑
RMSE	midlatitude	2–9 ↑	1 ↓	1–9 ↓	6–9 ↓	-
	tropics	3–9 ↑	1 ↓	1–9 ↓	3–9 ↓	-
ETS (1 mm d^{-1})	midlatitude	1 ↓	-	1,8 ↓	-	-
	tropics	1 ↓ 2–9 ↑	4–9 ↓	3–9 ↓	2–9 ↓	-
ETS (10 mm d^{-1})	midlatitude	-	1 ↑	-	1–9 ↑	6,7 ↑
	tropics	7 ↑	1–9 ↑	1–9 ↑	1–3, 9 ↑	1 ↑
ETS (25 mm d^{-1})	midlatitude	7–9 ↑	-	-	1–9 ↑	4–9 ↑
	tropics	1 ↓ 7,9 ↑	1 ↑	-	-	1, 2 ↑
ETS (50 mm d^{-1})	midlatitude		-	-	1–7 ↑	1 ↑
	tropics	6,8,9 ↑	-	1–4 ↓	1–6 ↑	1 ↑
CRPSS	midlatitude	1–9 ↓	1–9 ↑	2–7 ↑	1–9 ↑	-
	tropics	1–9 ↓	1–9 ↑	1–9 ↑	1–9 ↑	-
BSS (1 mm d^{-1})	midlatitude	19 ↑	1–3 ↑	-	4–9 ↑	-
	tropics	1–9 ↑	1–5 ↑	-	1–8 ↑	-
BSS (10 mm d^{-1})	midlatitude	1 ↑	1–8 ↑	2–9 ↑	1–9 ↑	-
	tropics	1,5–9 ↑	1–9 ↑	1–9 ↑	1–9 ↑	-
BSS (25 mm d^{-1})	midlatitude	-	1–9 ↑	1–9 ↑	1–9 ↑	-
	tropics	-	1–6 ↑	1–9 ↑	2–9 ↑	-
BSS (50 mm d^{-1})	midlatitude	3,4,6,8 ↓	1,2 ↑	1–9 ↑	7–9 ↑	-
	tropics	4–9 ↓	1 ↑	1–9 ↑	1 ↓ 3–9 ↑	1 ↑

[a]The up (down) arrow represents an increase (decrease) change.

precipitation thresholds. Performance changes due to the major EPS 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 exhibit 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 multiphysics 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 of 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 weigh 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.

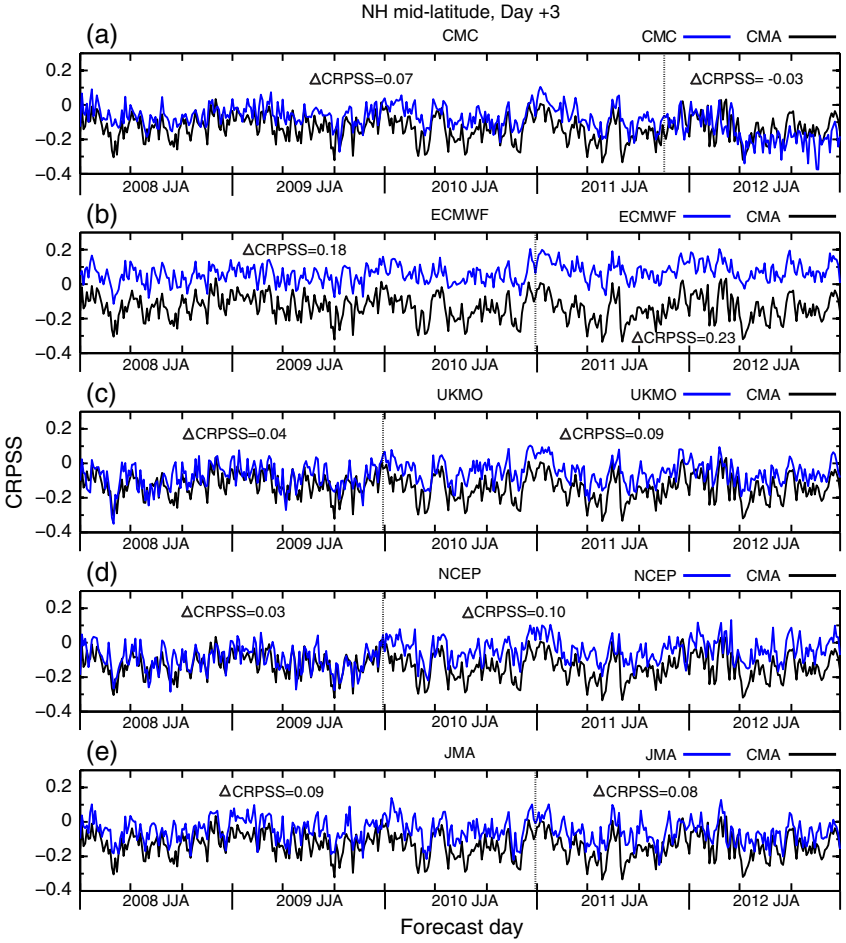

Figure 11. Time series of CRPSS for the day +3 PQPFs in the NH midlatitude. Inside numbers indicate the averaged CRPSS differences between each center and CMA during the periods before and after the major EPS upgrade split by the vertical dotted line. (a–d) The changes are significant with 95% confidence interval except (e) the CRPSS change of JMA.

The ultimate goal of verification study is to improve the performance of QPFs and PQPFs. The postprocessing work and the development of the EPSs (improvements in data assimilation, initial perturbations, and model components and configurations) 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 postprocessing 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 EPS quality can benefit more from the postprocessing 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 5 years, except for the CMA EPS. Therefore, the performance changes related to the major EPS 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 EPS upgrade 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 UKMO has reduced RMSEs and increased CRPSSs. The improvements in ETS and BSS vary with selected precipitation thresholds

and lead times. The EPS 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 but 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 outperforms the old version, based on the day +6 and +7 BSs of different precipitation thresholds and the 2.5 and 15 mm d^{-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, ECMWF, UKMO, and NCEP have improved the CRPSSs and BSSs of different thresholds for different 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.

Appendix A: The Calculation of Area-Weighted Verification Scores

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

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} w_i \cdot (x_i - y_i)^2}{\sum_{i=1}^{N} w_i}} \tag{A1}$$

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). The overall Brier Score (BS) and continuous ranked probability score (CRPS) can be derived similarly from the area-weighted averages of the BSs and CRPSs on each grid point.

The Pearson correlation [*Wilks*, 2006] is modified to the spatial correlation (SC) to measure the similarity of two patterns:

$$\mathrm{SC} = \frac{\sum_{i=1}^{N} w_i \cdot (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N} w_i \cdot (x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{N} w_i \cdot (y_i - \overline{y})^2}} \tag{A2}$$

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

$$\overline{x} = \frac{\sum_{i=1}^{N} w_i \cdot x_i}{\sum_{i=1}^{N} w_i} \tag{A3}$$

$$\overline{y} = \frac{\sum_{i=1}^{N} w_i \cdot y_i}{\sum_{i=1}^{N} w_i} \tag{A4}$$

To calculate the area-weighted dichotomous scores such as equitable threat score (ETS) and false alarm ratio (FAR), the variables in the 2×2 contingency table for the kth rain event are calculated first as

$$\mathrm{hit}_k = \frac{\sum_{i=1}^{N} w_i \cdot A_k^i \cdot B_k^i}{\sum_{i=1}^{N} w_i} \tag{A5}$$

$$\mathrm{mis}_k = \frac{\sum_{i=1}^{N} w_i \cdot \left(1 - A_k^i\right) \cdot B_k^i}{\sum_{i=1}^{N} w_i} \tag{A6}$$

$$\mathrm{fal}_k = \frac{\sum_{i=1}^{N} w_i \cdot A_k^i \cdot \left(1 - B_k^i\right)}{\sum_{i=1}^{N} w_i} \tag{A7}$$

$$crj_k = \frac{\sum_{i=1}^{N} w_i \cdot \left(1 - A_k^i\right) \cdot \left(1 - B_k^i\right)}{\sum_{i=1}^{N} w_i} \tag{A8}$$

where $A_j^i = 1$ if the jth event is forecasted for the ith sample or otherwise $A_j^i = 0$ and B_k^i is similar but for the observed kth event. Then the dichotomous scores can be derived, e.g., $\mathrm{FAR}_k = \mathrm{fal}_k/(\mathrm{hit}_k + \mathrm{fal}_k)$.

Acknowledgments

In this paper, the TIGGE data are from the ECMWF portal (http://tigge-portal.ecmwf.int/) and the TRMM data are available at NASA's Earth Observing System Data and Information System (EOSDIS) (ftp://disc2.nascom.nasa.gov/data/TRMM/Gridded/3B42_V7/). The authors appreciate Roberto Buizza of ECMWF and anonymous reviewers for their constructive suggestions. 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), and Natural Science Foundation of China (41175087). We also thank the support of High Performance Computing Center of Nanjing University.

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, *J. Atmos. Sci.*, *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, *Mon. Weather Rev.*, *129*(3), 420–436, doi:10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2.

Bougeault, P., et al. (2010), The THORPEX interactive grand global ensemble, *Bull. Am. Meteorol. Soc.*, *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, *Q. J. R. Meteorol. Soc.*, *134*(632), 703–722, doi:10.1002/qj.234.

Buizza, R. (2008), The value of probabilistic prediction, *Atmos. Sci. Lett.*, *9*(2), 36–42, doi:10.1002/asl.170.

Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli (1999a), Probabilistic predictions of precipitation using the ECMWF ensemble prediction system, *Weather Forecasting*, *14*(2), 168–189, doi:10.1175/1520-0434(1999)014<0168:PPOPUT>2.0.CO;2.

Buizza, R., M. Milleer, and T. Palmer (1999b), Stochastic representation of model uncertainties in the ECMWF ensemble prediction system, *Q. J. R. Meteorol. Soc.*, *125*(560), 2887–2908, doi:10.1256/smsqj.56005.

Buizza, R., M. Leutbecher, and L. Isaksen (2008), Potential use of an ensemble of analyses in the ECMWF ensemble prediction system, *Q. J. R. Meteorol. Soc.*, *134*(637), 2051–2066, doi:10.1002/qj.346.

Buizza, R., M. Leutbecher, L. Isaksen, and J. Haseler (2010), Combined use of EDA-and SV-based perturbations in the EPS, *ECMWF Newsl.*, *123*, 22–28. [Available online at http://old.ecmwf.int/publications/newsletters/pdf/123.pdf.]

Fritsch, J. M., et al. (1998), Quantitative precipitation forecasting: Report of the eighth prospectus development team, U.S. Weather Research Program, *Bull. Am. Meteorol. Soc.*, *79*(2), 285–299. [Available at http://journals.ametsoc.org/doi/pdf/10.1175/1520-0477%281998%29079%3C0285%3AQPFROT%3E2.0.CO%3B2.]

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*), *Tech. Rep.*, Development and Operations Divisions at CMC and Meteorological Research Division, Dorval, Quebec, Canada. [Available at http://collaboration.cmc.ec.gc.ca/cmc/CMOI/product_guide/docs/lib/op_systems/doc_opchanges/technote_geps_20110906_e.pdf.]

Hamill, T. M. (1999), Hypothesis tests for evaluating numerical precipitation forecasts, *Weather Forecasting*, *14*(2), 155–167, doi:10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

Hamill, T. M. (2012), Verification of TIGGE multimodel and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the contiguous United States, *Mon. Weather Rev.*, *140*(7), 2232–2252, doi:10.1175/MWR-D-11-00220.1.

Hamill, T. M., and J. Juras (2006), Measuring forecast skill: Is it real skill or is it the varying climatology?, *Q. J. R. Meteorol. Soc.*, *132*(621C), 2905–2923, doi:10.1256/qj.06.25.

Hamill, T. M., J. S. Whitaker, and S. L. Mullen (2006), Reforecasts: An important dataset for improving weather predictions, *Bull. Am. Meteorol. Soc.*, *87*(1), doi:10.1175/BAMS-87-1-33.

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, *Meteorol. Appl.*, *16*(1), 91–101, doi:10.1002/met.132.

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, *Atmos. Sci. Lett.*, *11*(2), 132–138, doi:10.1002/asl.270.

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, Am. Meteorol. Soc., New Orleans, Louisiana. [Available at https://ams.confex.com/ams/pdfpapers/134165.pdf.]

Hou, D., et al. (2012), Climatology-calibrated precipitation analysis at fine scales: Statistical adjustment of stage IV towards CPC gauge based analysis, *J. Hydrometeorol.*, doi:10.1175/jhm-d-11-0140.1.

Houtekamer, P., H. L. Mitchell, and X. Deng (2009), Model error representation in an operational ensemble Kalman filter, *Mon. Weather Rev.*, *137*(7), 2126–2143, doi:10.1175/2008MWR2737.1.

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, *J. Hydrometeorol.*, *8*(1), 38–55, doi:10.1175/JHM560.1.

Jolliffe, I. T., and D. B. Stephenson (2003), *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, John Wiley, Chichester, West Sussex, England.

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, *Adv. Atmos. Sci.*, *26*(5), 813–839, doi:10.1007/s00376-009-8162-z.

Leutbecher, M. (2005), On ensemble prediction using singular vectors started from forecasts, *Mon. Weather Rev.*, *133*(10), 3038–3046, doi:10.1175/MWR3018.1.

Murphy, A. H. (1973), A new vector partition of the probability score, *J. Appl. Meteorol.*, *12*(4), 595–600, doi:10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

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, *Geophys. Res. Lett.*, *35*, L10404, doi:10.1029/2008GL033837.

Richardson, D., R. Buizza, and R. Hagedorn (2005), First workshop on the THORPEX interactive grand global ensemble (TIGGE), *Tech. Rep.*, 1–39 pp., WMO World Weather Research Programme, Reading, U. K. [Available at http://www.wmo.int/pages/prog/arep/wwrp/new/documents/TIGGEFirstWorkshopReport.pdf.]

Sakai, R., M. Kyouda, M. Yamaguchi, and T. Kadowaki (2008), A new operational one-week Ensemble Prediction System at Japan Meteorological Agency, Rep., CAS/JSC WGNE Research Activities in Atmospheric and Oceanic Modelling, Montreal, Canada. [Available at http://www.wcrp-climate.org/WGNE/BlueBook/2008/individual-articles/06_Sakai_Ryota_JMAWEPS.pdf.]

Schumacher, R. S., and C. A. Davis (2010), Ensemble-based forecast uncertainty analysis of diverse heavy rainfall events, *Weather Forecasting*, *25*(4), 1103–1122, doi:10.1175/2010waf2222378.1.

Toth, Z., and E. Kalnay (1997), Ensemble forecasting at NCEP and the breeding method, *Mon. Weather Rev.*, *125*(12), 3297–3319, doi:10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2.

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, *J. Geophys. Res.*, *113*, D10110, doi:10.1029/2007JD008750.

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, Ser. A*, *60*(1), 62–79, doi:10.1111/j.1600-0870.2007.00273.x.

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 Forecasting*, *26*(6), 957–974, doi:10.1175/WAF-D-10-05060.1.

Wilks, D. S. (2006), *Statistical Methods in the Atmospheric Sciences*, 2nd ed., Academic Press, San Diego, Calif.

Yamaguchi, M., and S. J. Majumdar (2010), Using TIGGE data to diagnose initial perturbations and their growth for tropical cyclone ensemble forecasts, *Mon. Weather Rev.*, *138*(9), 3634–3655, doi:10.1175/2010MWR3176.1.

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, *Mon. Weather Rev.*, *133*(1), 279–294, doi:10.1175/MWR-2858.1.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne (2002), The economic value of ensemble-based weather forecasts, *Bull. Am. Meteorol. Soc.*, *83*(1), 73–83, doi:10.1175/1520-0477(2002)083<0073:TEVOEB>2.3.CO;2.