RESEARCH REPORT 2015-03

APEC CLIMATE CENTER

Development of a Real-Time Verification System for the APCC Operational Multi-Model Ensemble Prediction

Young-Mi Min
Climate Prediction Team

APCC
APEC CLIMATE CENTER

APEC CLIMATE CENTER
RESEARCH REPORT

Development of a Real-Time Verification System for the APCC Operational Multi-Model Ensemble Prediction

Young-Mi Min
Climate Prediction Team

RESEARCH REPORT 2015-03

APCC
APEC CLIMATE CENTER

Preface

The multi-model ensemble (MME) technique was designed around the start of this century to quantify uncertainties in forecasts associated with model formulation and has been considered as an effective means of improving weather and climate forecasts. As a result, various MME prediction systems are currently utilized at several operational centers (e.g., the European Centre for Medium-Range Weather Forecasts (ECMWF), the International Research Institute for Climate and Society (IRI), National Center for Environmental Prediction (NCEP), Meteorological Service of Canada (MSC), World Meteorological Organization Lead Center (WMO LC), and the North American Multimodel Ensemble (NMME) that routinely provide MME seasonal forecasts.

Since its inception in 2005, The Asia-Pacific Economic Cooperation (APEC) Climate Center (APCC) has devoted considerable effort to developing a MME prediction system for producing improved and well-validated seasonal and regional forecasts in both deterministic and probabilistic frameworks for research and operational purposes. Currently, the APCC operates the largest ensembles in the world and has obtained essential experience in developing and assessing multi-model predictions. Along with seasonal forecasts, their verification information of retrospective forecasts (hindcast) is now also available online via their website. However, verification of real-time forecasts is not yet operational, although it is a very important issue from an operational perspective to investigate the ability of the APCC MME prediction system in presenting seasonal temperature and precipitation for real-time forecasts in a timely manner.

Motivated by this, a real-time verification system has been developed to assess the quality of the single-model and multi-model predictions in the APCC operational environment. Along with this, we have improved the

current verification system for hindcasts considering the recent improvements of the APCC operational prediction system. The present report also provides a preliminary documentation of the seasonal forecasts issued by the APCC operational multi-model forecasts, with a large set of predictions currently available in an operational context, particularly focusing on real-time predictions of temperature and precipitation. The developed real-time verification system has been successfully employed by the APCC as an operational tool and internally provided their information every month. All verification information of the hindcast and real-time forecasts will be available online beginning in 2015.

The APCC will continue to improve the accuracy of its long-range climate prediction information. The new service will bring us one step closer to meeting users' expectations as the public gains awareness on how climate affects their lives, and also strives to promote our mutual interests and scholarly exchange with the climate forecasting centers and institutes within the APEC region and beyond. Finally, I extend my thanks to you and I hope you enjoy this 2014 Research Report.

Dr. Chin-Seung Chung
Director / APEC Climate Center
March 2015

ABSTRACT

A real-time verification system for APCC participating single-model and multi-model predictions has been developed to assess the quality of the predictions in the APCC operational environment. This system is based on recommendations from the Commission for Basic Systems (CBS) of the World Meteorological Organization (WMO) Standardized Verification System for long-range forecasts (SVS-LRF) in terms of verification metrics and regions. The developed verification system for real-time forecast issues monthly three-month overlapping seasons for upcoming 3-month predictions (from a combination of 1-tier and 2-tier models) and 6-month predictions (from 1-tier models), with a 1-month lead time. In addition, efforts have been made to improve the current version of the verification system for retrospective forecasts (hindcasts) operationally implemented at the APCC by considering the recent improvements of the prediction system (e.g., extended lead time, SST and ENSO prediction). This report provides a detailed explanation of the developed/improved verification system for real-time forecasts/hindcasts and preliminary documentation of the current status of the APCC operational multi-model performance, particularly focusing on real-time predictions of temperature and precipitation for the period of 2008JFM–2013/14DJF. Skill comparisons of other operational centers providing seasonal forecasts are also discussed.

Contents

Development of a Real-Time Verification System for the APCC Operational Multi-Model Ensemble Prediction

1. INTRODUCTION

The Asia-Pacific Economic Cooperation (APEC) Climate Center (APCC) has facilitated the sharing of high cost climate data and information and promoted capacity-building to meet the growing societal and economic interests in monitoring and predicting seasonal climate variability and to minimize economic and human losses due to natural disasters. The APCC began issuing quarterly seasonal forecasts of global climate in September 2005. Since 2007, however, the APCC has issued monthly-rolling global predictions of temperature and precipitation for the upcoming three-month overlapping seasons that are disseminated to APEC member economies via their website (http://www.apcc21.org). Currently, four deterministic Multi-Model Ensemble (MME) methods[1] and one probabilistic MME method[2] based on the ensemble members of participating models are operationally employed for seasonal forecasts at the APCC. These forecasts utilize deterministic (based on the ensemble mean) and probabilistic (based on the full set of ensemble members) interpretations of the well-validated MME seasonal prediction system.

The participating models in the APCC operational MME prediction originally consisted of only operational centers from APEC member economies at the initial stage of the APCC, but have recently been diversified by voluntary participation from two non-APEC member economies (i.e., Italy and the United Kingdom). As a result, 16

---

[1] The first method is a simple averaged MME, where the contribution of each model is equally weighted (i.e., a simple composite method; SCM). The second is a calibrated MME obtained from the adjusted (or corrected) single-model predictions, based on a stepwise pattern projection method (SPM; Kug et al. 2008c). Others are empirically weighted MMEs with coefficients computed using multiple linear regression (Krishnamurti et al. 2000; Yun et al. 2003); and the use of multiple linear regression with the empirical orthogonal function-filtered dataset minimizes the residual error variance (i.e., a synthetic superensemble method; Yun et al. 2005).

[2] The operational probabilistic forecasts at the APCC are based on the MME, with the model weights being inversely proportional to the random errors in the forecast probability (Min et al. 2009). The APCC issues seasonal forecasts in the form of tercile-based categorical probabilities, i.e., the probabilities of below-normal (BN), near-normal (NN), and above-normal (AN) categories. The tercile-based categorical probabilities from an ensemble of forecasts, including the tercile information based on hindcasts, are estimated using a parametric estimator derived from a fitted Gaussian distribution (e.g., Kharin and Zwiers 2003; Boer 2005; Min et al. 2011).

prominent operational centers and research institutes from 10 countries currently participate in the APCC operational MME prediction by routinely providing their predictions in the form of ensembles of global forecast fields (Table 1).

Since its inception in 2005, the APCC has devoted considerable efforts in developing a MME prediction system for producing improved and well-validated global and regional forecasts in both probabilistic and deterministic frameworks for research and operational purposes (e.g., Kang et al. 2009; Min et al. 2009; Lee et al. 2011; Min et al. 2011; Sohn et al. 2012; Lee et al. 2013a; Lee et al. 2013b; Kang et al. 2014). As a result, there are several developments and improvements of the operational prediction system. One of the more recent developments is that the APCC has been forecasting sea surface temperature (SST) and El Niño-Southern Oscillation (ENSO) based on the MME prediction using state-of-the-art coupled general circulation models since 2012 (available at http://www.apcc21.org/eng/service/ 6mon/enso/japcc030701.jsp). Most recently, the member economies requested to extend the scope of the APCC's seasonal climate prediction services to better prepare for climate-related hazards in a timely manner. Upon receiving this request and to meet the needs of the APEC community, the APCC launched its operational 6-month lead MME seasonal prediction service in September 2013 (available at http://www.apcc21.org/eng/service/6mon /ps/japcc 030703.jsp).

Along with the developments and/or improvements of the operational prediction system, a standardized verification system has also been developed to validate the prediction system. Currently, an automated verification system for retrospective forecasts (hindcasts) has been operationally implemented for a 3-month lead MME prediction and its information is disseminated via our website. In terms of the verification of the APCC real-time multi-model forecast, only a few studies have been conducted (e.g., Min et al. 2009; Min et al. 2014). However, a verification system is not operational (i.e., for research purposes), although it is a very important issue from an operational perspective to investigate the ability of the APCC MME prediction system in presenting seasonal temperature and precipitation for real-time forecasts in a timely manner. Thus, one of the key issues in the APCC is to develop a real-time verification system to verify the operational prediction system and improve the accuracy of its long-range climate prediction information. Along with this, we also need to improve the current verification system for retrospective forecasts considering

the recent improvements of the prediction system (e.g., extended lead-time from 3 months to 6 months, and inclusion of SST and ENSO prediction). Motivated by the importance of the real-time forecast verification, this study attempted to develop a real-time verification system for the APCC MME prediction and to successfully implement it in an operational environment.

Currently, the APCC operates the largest multi-model ensemble prediction systems and has obtained essential experience in developing and assessing multi-model predictions. Most recently, Min et al. (2014) documented a comprehensive assessment of the APCC operational MME forecasts based on a large set of predictions. However, they only focused on deterministic forecasts with a 3-month lead multi-model prediction for temperature and precipitation, mostly focusing on the retrospective forecast. Now that the APCC has a verifiable history of more than 6 years for the real-time operational forecasts, their performance can be examined. Here, another purpose of this report is to provide a preliminary assessment of the current APCC operational seasonal forecasts in both deterministic and probabilistic frameworks, particularly focusing on the real-time forecasts, which has not been fully examined in previous studies.

The report is organized as follows: Section 2 describes the participating models in the APCC MME prediction, and their retrospective/real-time forecast datasets and corresponding observations. The developed real-time verification system is briefly introduced along with a brief explanation of the verification metrics and regions and preliminary assessment for the real-time forecasts in Section 3. Finally, Section 4 summarizes and discusses the results.

2. Data

a. Model Data

The APCC operational verification system for both retrospective and real-time forecasts includes procedures to measure the performance of the APCC participating single- and multi-model predictions. The models include the 17 dynamical climate prediction systems from the APCC MME producing centers. Table 2 presents a brief summary of model

specifications for five two-tier (COLA, CWB, HMC, IRI, and MGO) and twelve one-tier (APCC, BCC, MSC_CANCM3, MSC_CANCM4, CMCC, GloSea5, JMA, NASA, NCEP, PNU, POAMA and UKMO) systems, respectively. The models show a large range of model resolutions and ensemble sizes, and the model dataset is interpolated to a common resolution of 2.5°lon × 2.5°lat grid similar to that of the observed data.

Among them, only 12 dynamical seasonal prediction models (APCC, BCC, CWB, HMC, NASA, MSC_CANCM3, MSC_CANCM4, CMCC, JMA, NCEP, PNU, and POAMA) currently participate in the monthly rolling one-month lead seasonal mean MME predictions because their retrospective forecast datasets match the requirements of the Seasonal Prediction Model Intercomparison Project/Historical Forecast Project (SMIP/HFP), or the Coupled Model Intercomparison Project (CMIP). All models have generated ensemble retrospective forecasts for the common period of 1983–2003. Here, GloSea5 and UKMO also match the CMIP, but there are some difficulties in including them in the APCC MME prediction due to a relatively short hindcast period (1996–2009) as compared to the other prediction systems. For more detailed information on each prediction system, please refer to their references listed in Table 2.

b. Observed Data

The data used for verification of the retrospective and real-time forecasts of basic variables (e.g., temperature at 2m, geopotential height at 500hPa) in the APCC verification system were obtained from the NCEP-Department of Energy (DOE) reanalysis 2 data (Kanamitsu et al. 2002). The observed precipitation data used in the verification system is the Climate Anomaly Monitoring System and Outgoing longwave radiation Precipitation Index data (CAMS OPI; Janowiak and Xie 1999). The CAMS OPI is a precipitation analysis created by merging ground-based rain gauge observations with satellite rainfall estimates to obtain real-time monthly analyses of global precipitation. To evaluate Sea Surface Temperature (SST) and Nino indices, we use the optimum interpolation (OI) version 2 monthly mean SSTs (Reynolds et al. 2002), obtained from the Climate Diagnostics Center (CDC) of the National Oceanographic and Atmospheric Administration (NOAA).

3. Research Results

a. APCC Operational Verification System

As mentioned in the introduction, our activities include (a) developing an operational real-time verification system for both 3-month and 6-month lead forecasts and (b) improving the current retrospective (hindcast) verification system, currently only available for 3-month lead forecast, by considering the recent improvements of the prediction system (e.g., extended lead time, SST and ENSO prediction) with matching the procedures in the newly developed real-time verification system. Major issues in the verification system will be discussed in the following sub-sections. Note that full details of the verification system are given in **Appendix C**–Operation Manual for the APCC Operational Verification System.

1) REAL-TIME VERIFICATION SYSTEM

A newly developed real-time verification system is used to assess the real-time skill of individual models and multi-model ensembles for the upcoming 3-month (1-tier and 2-tier models) and 6-month (only 1-tier models) seasonal forecasts, issued once per month with near real-time updated observations.

(i) Key parameters

· **Variable:**

- The following variables of the model output are verified: temperature anomaly at 850hPa (t850), temperature anomaly at 2m (t2m), precipitation anomaly (prec), geopotential height anomaly at 500hPa (z500), and sea surface temperature anomaly (SST)

- In addition to these parameters, the Nino indices (e.g., Nino1+2/3/4/3.4), ENSO-Modoki index (EMI), and Indian Ocean Dipole (IOD) index, defined as the mean SST anomaly over each region (Table 3), are also verified.

· **Interval:** monthly and seasonal (i.e., monthly rolling 3-month means, e.g., JFM, FMA, MAM)

· **Lead Time:** 3-month and 6-month

· **Region:**

- Large-scale: Globe, Tropics, Southern Extratropics, and Northern Extratropics
- Regional-scale: East Asia, South Asia, South America, North America, Australia, Australia + South Pacific, Northern Eurasia, and Middle East

· **Metrics:**

- Deterministic: Anomaly Pattern Correlation Coefficient (ACC), Temporal Correlation Coefficient (TCC), Root Mean Square Error (RMSE), Mean Square Skill Score (MSSS)
- Probabilistic: reliability diagram, Relative Operating Characteristic (ROC), Heidke Skill Score (HSS), Brier Skill Score (BSS), Ranked Probability Skill Score (RPSS)

(ii) Verification regions

To more objectively define standard regions for the APCC operational real-time verification system, we followed recommendations from the Standardized Verification System (SVS) for long-range forecasting (LRF), developed by the Commission for Basic Systems (CBS) of the World Meteorological Organization (WMO; WMO 2002) and CORDEX (Coordinated Regional Climate Downscaling Experiment). First, the verification scores were produced over sub-regions to estimate large-scale verification statistics in order to evaluate the overall skill of the forecast, including the Globe (GL: 0–360°E, 90°S–90°N), Tropics (TR: 0–360°E, 20°S–20°N), and Northern (NE: 0–360°E, 20°–90°N) and Southern Extratropics (SE: 0–360°E, 20°S–90°NS).

In addition, to provide a regionalized assessment of the forecast system, we defined several sub-regions, focusing on the APEC member economies, which are more interest-

ed in the APCC. The target regions include East Asia (EA: 75°E–150°E, 15°N–60°N), South Asia (SA: 60°E–140°E, 10°S–35°N), North America (NAm: 190°E–310°E, 10°N–75°N), South America (SAm: 270°E–330°E, 60°S–10°N), Australia (AUS: 110°E–180°E, 50°S–0°N), Australia with some areas of South Pacific (AUS+SP: 110°E–260°E, 50°S–20°N), Northern Eurasia (NEu: 25°E–190°E, 40°N–80°N), and the Middle East (ME: 25°E–75°E, 10°N–45°N). Here, the Middle East is recently included in the APCC standard regions for the outlook and verification at the official request of the Qatar Meteorological Administration. The newly selected twelve standard regions of the APCC operational real-time verification are shown in Fig. 1. More detailed information on the definition of the standard regions can be also found in **Appendix C**.

(iii) Verification metrics

The verification procedures used to measure the performance of the seasonal forecasts in the real-time verification system are those recommended by WMO SVS-LRF. The real-time verification system is incorporated following three diagnostic measures; ROC, reliability diagrams and accompanying measures of sharpness (i.e., frequency histogram), and MSSS with associated decomposition. The recommended skill scores are very informative and powerful to assess the seasonal forecasts, but there are difficulties in communicating the information to the general public and/or end-users. In addition, it is recommended that, particularly, reliability diagrams should be constructed for large-sample probability forecasts aggregated over large-scale regions (e.g., globe, tropics, southern and northern extratropics) because of the sensitivity of the reliability diagram to small sample sizes. However, the APCC is more interested in regional-scale assessment, especially for APEC member economies. Thus, by considering these situations, we additionally use a variety of user-friendly metrics, which are widely used in operational centers to measure the seasonal forecast skill; TCC, RMSE, ACC, BSS, HSS, and RPSS. These metrics provide more easily understandable information for communicating the quality and potential value of seasonal forecast products. Detailed description of each verification measure can be found in **Appendix A**.

2) HINDCAST VERIFICATION SYSTEM

Currently, the Hindcast verification system is operationally implemented at the APCC, but only for 3-month lead MME forecasts, and its information is disseminated via our website. However, there are many points to be improved as shown below.

- Unavailable for monthly verification of probabilistic forecasts.
- Unavailable for verification of 6-month lead MME forecasts.
- Unavailable for verification of SST and ENSO forecasts.
- Inconsistency in target regions for deterministic and probabilistic verification.

We made efforts to improve the current hindcast verification system considering all aspects of the points above.

The APCC operational verification system for both real-time forecasts and hindcasts is summarized in Table 4, and example figures are attached in **Appendix B**. As compared with other operational centers, providing multi-model seasonal forecasts and their verification information (Table 5), the APCC verification system is unique and has many advantages, as listed below.

- Available for various verification metrics.
- Available for various variables (not only basic variables, but also SST and ENSO indices).
- Available for both real-time forecasts and hindcasts.
- Available for both large-scale and regional-scale assessments.
- Available for each month and 3-month means for the upcoming 6 months.

b. Preliminary Results of APCC Real-Time Forecasts

To investigate the ability of the single-model and multi-model predictions in presenting seasonal mean temperature and precipitation for real-time forecasts, which is a very important issue from an operational perspective, we have evaluated the predictions for 12-running 3-month means during the period of 2008JFM–2013/14DJF. Note that the APCC has started monthly 3-month mean forecasts since November 2007.

1) APCC SINGLE-MODEL VS. MULTI-MODEL PREDICTIONS

In this section, we first present the overall performance of the APCC single-model and multi-model predictions for monthly rolling 3-month mean temperature at 850hPa (hereafter, temperature) and precipitation, with a 1-month lead time. Figures 2 and 3 show the time series of the anomaly pattern correlation averaged over the large-scale regions (globe, northern extratropics, and tropics) and most interested areas in the APCC, East Asia, for the period of JFM2008–DJF2013/14. Here, the SCM is the simple averaged MME with equal weighting from the participating single-model predictions at each target season. Note that the participating models in the APCC MME prediction are slightly different for each target season due to the operational situation at that time. The averaged performance of all single-model predictions is also displayed in the plots.

First, it was found that the general performance of the single-model and multi-model predictions was relatively more accurate in the tropics for both variables, as shown in many previous studies (e.g., Palmer et al. 2004; Min et al. 2009; Wang et al. 2009; Lee and Wang 2012; Jia et al. 2012; Min et al. 2014). Second, the spread of the individual model temperature correlations is generally larger than that of precipitation. The accuracy of an individual model prediction of comparatively well predictable temperature depends on the ability of individual models to simulate the climate system, with larger spreads reflecting differences in this ability, whereas for the less predictable precipitation, the accuracies of all of the individual models tended to be modest, with the differences in individual model skills being low. In the same sense, the larger spread of the individual

model skills in East Asia can be explained.

Third, the results clearly demonstrate that the SCM predictions generally perform better than any single-model predictions across all the 3-month overlapping seasons, regions, years, and variables. There are few cases in which the single-model performance is better or comparable to that of the multi-model. However, it should be noted that the main advantage of using a multi-model is not the large improvement compared to the best single-model in individual cases, but rather the consistently better performance of the multi-model when considering all aspects of the predictions (Doblas-Reyes et al. 2005; Hagedorn et al. 2005; Min et al. 2009; Wang et al. 2009; Kryjov 2012; Rodrigues et al. 2014; Min et al. 2014). Finally, the forecast skill of the multi-model predictions is consistently better than the averaged skill of all single-model predictions across all the seasons and regions. However, for some cases where the general performance of individual models is quite negative (especially in East Asia), the mean skill of individual models shows better performance than that of the multi-model. Because the skill of the multi-model prediction in terms of correlation is proportional to the averaged covariance of the single-model forecasts against observations and inversely proportional to the variance of the multi-model forecast, reduction of the single model variances provide a mutual offset of the single model errors (Yoo and Kang 2005). Thus, the multi-model ensemble provides a better skill if the skillful models were combined and the composite variance (i.e., the spread of the individual model skills) was smaller.

2) GENERAL PERFORMANCE OF APCC MME PREDICTION

(i) Spatial distribution of forecast skill

To illustrate the spatial distribution of the forecast skills for the period of 2008–2013, the temporal correlations between observations and one-month lead seasonal mean multi-model predictions of temperature and precipitation for the warm season (April–August) and cold season (October–February) are shown in Figure 4. Here, contour lines indicate that the temporal correlation is statistically significant at the 5% level using a Student

t-test. The estimated scores in the plot are first calculated for each grid-point and then averaged over the globe. Note that the 6-year period is relatively short for the collection of a sufficient number of real-time forecasts to obtain some quantitative estimates and make well-grounded conclusions for each season assessment (i.e., JJA, DJF). Thus, we use the whole time series of each 3-month mean forecast for the warm/cold season during the period of 2008–2013. For example, we use 30 cases for the warm season; 5 forecasts (AMJ, MJJ, JJA, JAS, and ASO) a year for 6 years produces 30 samples.

For temperature, the MME prediction for the cold season shows slightly wider coverage of the areas where the estimated temporal correlation is statistically significant at the 5% level as compared with that of the warm season. This is because ENSO reaches its peak during the boreal winter season and ENSO variability is likely the sole source of predictability of the seasonal forecast. For example, relatively high levels of skill for cold season are observed in the horseshoe region extending from the western Pacific (east of the Philippines) toward the extratropics in the northeastern and southeastern Pacific, and also in the entire Indian Ocean, South America, and the central North Atlantic. Most of these additional skills originate from the influence of ENSO via teleconnections during its peak phases (Wang et al. 2009; Barnston et al. 2010).

During the boreal warm season, the MME temperature prediction (with a temporal correlation value of 0.20) shows relatively lower skills than during the boreal cold season (with a temporal correlation value of 0.30). However, in the western and central Siberia, Maritime Continent, and most parts of North America, the forecast skill for the warm season is greater than that of the cold season for the period of 2008–2013. In general, in the tropical Pacific between 10°N-20°N (north of the Intertropical Convergence Zone (ITCZ)), the accuracy of warm season forecasts is better than that of the cold season because as the thermal equator moves northward for warm season, the northern hemisphere's tropics are more predictable than during cold season due to SST persistence. However, during the period of 2008–2013, most of the tropical regions for the warm season show relatively lower levels of skill than for the cold season. The possible causes might result from

comparably lower skill for warm season and higher skill for the cold season for the 6-year real-time forecast than usual (i.e., the overall skill of the hindcast for the period of 1983–2003; not shown here). However, further study is necessary to better understand why there is less predictability in the tropical Pacific for the JJA season during the 6-year real-time forecasts.

While the forecast skill for precipitation is generally modest, showing an estimated temporal correlation over the globe of 0.10 and 0.14 for the boreal warm and cold season, respectively. The area where the estimated scores are statistically significant is only confined to the tropical Pacific and the Maritime Continent. Comparison of the seasons shows that for cold season, the greater forecast skills are well extended into the global tropics and subtropics over the ocean, particularly for the western Pacific, affecting parts of southern Micronesia and Melanesia. As mentioned previously, the overall expansion of the forecast skill may be a result of the model's capacity to capture ENSO teleconnections (Wang et al. 2009).

(ii) Interannual variation of forecast skill

Figures 5 and 6 show the temporal evolution of the ROC score aggregated over the globe and tropics for the APCC probabilistic multi-model forecasts for temperature and precipitation during the period of 2008JFM to 2013/14DJF. The APCC operational seasonal forecasts are issued in the form of tercile-based categorical probabilities, i.e., the probability of Above-Normal (AN), Near-Normal (NN), and Below-Normal (BN) categories, with respect to climatology. For more detailed information of the APCC operational probabilistic MME forecasts, please refer to Min et al. (2009).

Results from Fig. 5 and 6 first indicate that the real-time forecasts of the APCC probabilistic MME predictions for temperature and precipitation over the globe and tropics are generally accurate, with aggregated ROC scores over 0.5 in most of the target seasons for the period of 2008–2013. Second, the time series of ROC scores for the APCC temperature forecasts show higher average levels than those of the precipitation forecasts. Third, results

for the NN category in both variables show consistently lower accuracy than the other two categories because its probability is relatively insensitive to signal perturbations. As a result, the probability of the NN category is close to the climatological probability, and thus, it is not easy to predict the outcome of an NN event with high confidence, consistent with many previous studies (e.g., van den Dool and Toth 1991; Kharin and Zwiers 2003; Min et al. 2009).

A more interesting feature is that the forecast skill over the 6-year period is strongly related to ENSO variability. That is, it shows a clear relationship between the forecast skill of the multi-model prediction and the absolute value of the Nino 3.4 SST anomaly (black contour lines in Fig. 5 and 6), especially in the tropics, with correlation values of 0.49 for temperature and 0.51 for precipitation of the AN category, which are statistically significant at the 5% level. This is also true for the BN category. That is, the highest skills for all variables and categories are featured by the seasonal forecasts for the 2009/10 and 2010/11DJF, coinciding with the peaks of a moderate El Niño and a strong La Niña based on the ONI (Oceanic Nino Index; Smith et al. 2008), computed using the Extended Reconstructed SST (ERSST.v3b) from NOAA's National Climate Data Center (NCDC; http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears.html)[3]. In addition, relatively low levels of forecast skill for both variables and categories tend to occur during the transitional or ENSO-neutral periods both globally and in the tropics, which is consistent with Livezey and Timofeyeva (2008), who identified ENSO variability as the sole source of seasonal precipitation forecast skill. In particular, the forecast skills across all seasons in 2012 and 2013 are relatively lower than other real-time years and even retrospective forecasts during the period of 1983–2003. This may be because there was

[3] The ONI defined as 3-month running mean of ERSST.v3b SST anomalies in the Nino 3.4 region (5°N-5°S, 120°-170°W), based on centered 30-year base periods (1971–2000). The episodes are defined when the threshold is met for a minimum of 5 consecutive over-lapping seasons. ßIt is defined as a strong/moderate/weak El Niño (La Niña) based on a threshold of + 1.5/1.0/0.5°C (-1.5/1.0/0.5°C) for the ONI.

transition and ENSO-neutral periods after just decaying phase of the moderate La Niña in 2011/12DJF throughout the whole period.

Temperature skill is also related to ENSO state but differently to precipitation; skill is generally highest near the end of and shortly following El Niño events. Barnston et al. (2010) demonstrates that the greatest impact of ENSO on temperature forecast skill occurs 4 months following the ENSO peak for both ENSO phases. This influence on forecast skill is attributable to a delayed temperature response in both the tropics and extratropics (Kumar and Hoerling 2003), which was earlier documented in the context of the atmospheric bridge (Lau and Nath 1996; Alexander et al. 2002). Note that we issue monthly rolling 3-month overlapping seasons. As a result, correlations between the Nino 3.4 SST anomaly and the tropical ROC score for temperature of the AN category for 1- and 2-month after 3-month overlapping seasons are high with values of 0.60 and 0.62, respectively (note that correlation between the absolute value of the Nino 3.4 anomaly and the tropical ROC score for temperature of the AN category with no time delay was 0.49). However, in the case of precipitation, a simultaneous positive relationship with both phases of ENSO was noted. As a result, correlation between the absolute value of the Nino 3.4 index and the forecast skill for precipitation is generally higher than that for temperature in all categories. Finally, to assess the consistency of their skills during the longer period and the 6-year period of real-time predictions, the ROC scores between the hindcasts and observations were also examined for each target season. Table 6 shows generally higher hindcast skill levels in terms of averaged score over the entire 21-year period for both variables.

(iii) Comparison of different multi-model prediction systems

The previous section with Fig. 2–6 described the overall performance of the APCC single-model prediction and their simple averaged multi-model prediction (SCM), both spatially and temporally. In this section, we will assess the forecast skills obtained from the different APCC operational MME prediction systems, based on the 6-year series of real-time forecasts. As mentioned in Section 1, the APCC currently operates four different deterministic MME prediction systems: SCM, SSE, MRG, and SPM. MRG and SSE

are empirically weighted MMEs with coefficients computed using multiple regression and multiple regression with the empirical EOF-filtered dataset, respectively. SPM is a calibrated MME obtained from the corrected single-model predictions, based on a stepwise pattern projection method, followed by simple averaging with equal weighting.

Figure 7 generally shows the zonally averaged temporal correlation for multi-model predictions of temperature and precipitation for the entire time-series of forecasts (12 forecasts a year for 6 years equaling 72 cases), obtained from four different MME methods. Results clearly indicate that the SCM and SPM methods consistently outperform the MRG and SSE methods over most latitudinal belts for both variables. Their high levels of forecast skills are well extended into the middle- and high-latitudinal zones, especially for temperature, and the high skill levels are widely found over the subtropical regions, whereas those obtained from multiple regression-based weighted MME methods are confined to the equatorial zone. A comparison of the SCM and SPM predictions for temperature shows that the positive effect of model correction using the SPM method is comparatively larger in the tropics (20°S-20°N) and near 60°S and 60°N. The forecast skills of SPM and SCM for precipitation are modest and comparable to each other. We also compared the forecast skills of different MME methods for each season and several regions: the globe, northern extratropics, tropics and East Asia, in terms of anomaly pattern correlation (Fig. 8). Similar results from Fig. 7 can be found for each season over several large-scale regions and from regional-scale assessment in East Asia.

In Fig. 7–8, there are two important points to be discussed. First, during the 6-year real-time forecast periods, SCM generally outperforms the multiple regression-based weighted MMEs over most latitudinal zones for both variables. This result is consistent with previous studies indicating that the quality of the simple averaged multi-model predictions have difficulty improving upon multiple linear regressions because of the relatively short time series used to estimate the regression coefficients in the seasonal forecasts (e.g., Peng et al. 2002; Doblas-Reyes et al. 2005; Rodrigues et al. 2014; Min et al. 2014). Another possible reason of the failure of multiple regression-based MMEs in real situations may result from

over-fitting. However the tests in avoiding of the overfitting problem performed by DelSole (2007) show that even the successful solution of the over-fitting problem by, for instance, the use of (multiple) ridge regression does not provide any improved simulation accuracy as compared with the simple averaged MME.

Second, the model correction using SPM has a positive effect in improving the multimodel temperature and precipitation predictions for the period of 2008JFM–2013/14DJF. Similar results were found by Min et al. (2014). They demonstrate that the calibrated MME prediction using the SPM method shows the capability of reducing errors and improving forecast skills in a large proportion of cases, mostly focusing on the 23-year retrospective forecast during the period of 1983–2003. The superiority of the SPM method in real-time forecasts, as shown here from Fig. 7–8, may be more meaningful than in Min et al. (2014). The reason why the calibrated MME predictions have higher correlation than others when considering all aspects of the predictions is that the current dynamical climate models are capable of capturing large-scale patterns related to local variability, though they have difficulty in correctly predicting local variability at each grid point. Therefore, the skill of SPM by correcting the (spatially shifted) model errors using the statistical correction technique, shows better accuracy than others for both the retrospective (Min et al. 2014) and real-time forecasts. This result is quite significant because it shows considerable promise for the use of the calibrated MME method for a series of operational forecast systems in a real-time basis by correcting the single-model prediction, although the use of the simple averaged MME is the practical way of utilizing the multi-model approach in an operational environment. However, there are many issues from an operational point of view. Thus, many issues concerning the calibration method still remain and a more integrated view of the advantages of the method considered is also required.

3) COMPARISON OF OTHER OPERATION CENTERS

To investigate the level of forecast skill of the APCC multi-model predictions, we finally compared the simulation results with those of other operational centers, NCEP/CPC (Climate Prediction Center) and IRI, which are two of the major operational centers

providing seasonal forecasts. Figure 9 shows the region of the CPC's official outlook and verification, focusing on the United States; they also provide its verification score in a digital image format via their website (http://www.cpc.ncep.noaa.gov/products/verification/summary/index.php?page=chart). The U.S. region is not included in the standard regions for the APCC verification system (Fig. 1), but for objective comparisons, we verified the APCC multi-model prediction over the U.S. defined by the CPC. The CPC official forecasts are manually created by a forecaster considering all predictions from CFS[4], CCA[5], ECCA[6], ENSO composites[7], OCN[8], CAS[9], and SMIR[10].

---

[4] An ensemble mean forecast from a fully coupled 1-tier dynamical model (CFS). Forecasters use an ensemble mean of 40 forecast members.

[5] Canonical Correlation Analysis (CCA) linearly predicts the evolution of patterns of temperature and precipitation based upon patterns of global SST, 700mb height, and U.S. surface temperature and precipitation from the past year for the most recent four non-overlapping seasons.

[6] ECCA utilizes the CCA method of projecting loading patterns onto predictor fields to make a linear prediction of temperature and precipitation. These loading patterns are statistically determined by maximizing the correlation between the predictors and predictands using data from 1953 to present.

[7] Averages of observational data stratified by El Niño and La Niña or ENSO-neutral conditions provide guidance for U.S. El Niño and La Niña effects by supplying historical frequencies of the three forecast classes in past years when (for the particular forecast season) the central equatorial Pacific was characterized by moderate or strong La Niña or El Niño conditions.

[8] The Optimal Climate Normal (OCN) method predicts temperature and precipitation on the basis of year-to-year persistence of the observed average anomalies for a given season during the last 10 years for temperature and the last 15 years for precipitation. OCN emphasizes long-term trends and multi-year regime effects.

[9] Constructed Analog on Soil (CAS) moisture is based on empirical orthogonal functions (EOF) from data over the lower 48 states beginning in 1932. This tool constructs a soil moisture analog from a weighted mean of past years. Then the temperature and precipitation in the same proportion is used to produce a forecast that is consistent with current soil moisture conditions.

[10] Screening Multiple Linear Regression (SMIR) uses the same predictor fields as CCA, but is applied to the single station patterns, whereas the multi-station anomaly patterns are done in CCA. Additionally, SMLR uses the two week MRF-based soil moisture forecast as a predictor.

Comparison of the seasonal forecasts for temperature shows that the APCC multimodel forecasts generally outperform the CPC's manual forecasts over the U.S. for the whole 6-year real-time forecasts, indicating that the area-averaged skills are 25.6 for the APCC forecast and 11.2 for the CPC forecast, respectively (Fig. 10). In particular, skill improvement of the APCC multi-model prediction over the U.S. is prominently featured in transition or ENSO-neutral periods (e.g., boreal late spring-summer of 2010, boreal summer season in 2011, and the whole period of 2013) that usually show relatively low levels of skill. On the other hand, in case of precipitation, the performance of the APCC and CPC forecast is generally comparable to each other; however, it is further noted that the overall skill of APCC over the U.S. is consistently positive in most of the seasons, except for few cases. The skill improvement of the APCC forecast is generally modest, with a slightly higher averaged score over the period of 2008–2013 as compared with the CPC forecast (i.e., 11.9 for the APCC forecast and 7.2 for the CPC forecast).

Each month IRI also issues a general circulation model multi-model based seasonal forecast of global precipitation and temperature. Note that the IRI's multi-model prediction is based on a two-tier system and issues a 0.5-month lead tercile-based categorical probabilistic forecast (Bengtsson et al. 1993). Fig. 11 and 12 indicate the skill comparison between the APCC and IRI multi-model prediction in terms of HSS for global and tropical temperature and precipitation. Different from the CPC, the verification information is only available online in a graphic format for IRI's seasonal forecasts from the late 1990's to present (http://iri.columbia.edu/our-expertise/climate/forecasts/verification/). To compare objectively, we arbitrarily added a horizontal line as a reference at a value of 40 for temperature and 20 for precipitation in the Fig. 11–12.

IRI's forecast skill over the entire period has been strongly related to ENSO variability. In particular, greater accuracy is shown in forecasts of El Niño than of La Niña (green lines indicate the Nino 3.4 index in Figs. 11–12). That is, the forecast skill is highest near the end of, and shortly following El Niño events, whereas the lowest skill occurs with the same timing for La Niña events, particularly in the tropics. This is consistent to the previous studies of

Goddard and Dilley (2005) and Barnston et al. (2010). As a result, the skill improvement of the APCC multi-model prediction is prominently featured in the La Niña events (i.e., boreal autumn and winter season in 2010/11 and 2011/12). This skill improvement of the APCC multi-model prediction is clearer for precipitation than temperature, particularly in the tropics. To conclude, the APCC multi-model prediction shows reasonably better accuracy than other operational centers in a large proportion of the predictions over the 6-year period. Note that the NCEP/CPC and IRI simulates temperature at 2m (t2m) whereas the APCC issues temperature at 850hPa. t2m shows slightly better skill than t850 because t2m is strongly related to surface temperature. For the oceans, SST is predicted very well due to strong persistence (Fig. 13). Nevertheless, skill improvement of the APCC multi-model prediction for t850, as shown here, is more significant.

4. Concluding Remarks

This study was motivated by the importance of evaluating the quality of climate predictions, which is an essential component of seasonal forecasting. Verification information of retrospective forecasts (hindcasts) is now available online for the APCC seasonal deterministic and probabilistic forecasts (http://www.apcc21.org/eng/service/6mon/ver/hid/japcc030704_lst.jsp). However, verification of real-time forecasts is not yet operational, although it is a very important issue from an operational perspective to investigate the ability of the APCC MME prediction system in presenting seasonal temperature and precipitation for real-time forecasts in a timely manner. Motivated by this, a real-time verification system has been developed to assess the quality of the singlemodel and multi-model predictions in the APCC operational environment. Along with this, we have improved the current verification system for hindcasts considering the recent improvements of the APCC operational prediction system (e.g., extended lead time, SST and ENSO prediction).

The developed real-time verification system is based on recommendations from WMO SVS-LRF in terms of verification metrics and regions. The system covers basic variables (t850, t2m, prec, z500, and SST) and SST indices (Nino1+2/3/3.4/4 indices, ENSO-Modoki index, and IOD index). The metrics used to measure the prediction skill of deterministic forecasts include the temporal correlation coefficient (TCC), anomaly pattern correlation coefficient (ACC), root mean square error (RMSE), and mean square skill score (MSSS). Probabilistic forecasts were assessed using a variety of metrics; relative operating characteristics (ROC), reliability diagram, Brier skill score (BSS), Heidke skill score (HSS) and ranked probability skill score (RPSS). The verification scores were produced over sub-regions to estimate large-scale verification statistics to evaluate the overall skill of the forecast system, including the Globe (GL: 0–360°E, 90°S-90°N), Tropics (TR: 0–360°E, 20°S–20°N), and Northern (NE: 0–360°E, 20°–90°N) and Southern Extratropics (SE: 0–360°E, 20°S–90°NS). In addition, to provide a regionalized assessment of the forecast system, we defined several sub-regions, focusing on the APCC member economies of greater interest in the APCC. The target regions include East Asia (EA: 75°E–150°E, 15°N–60°N), South Asia (SA: 60°E–140°E, 10°S–35°N), North America (NAm: 190°E–310°E, 10°N–75°N), South America (SAm: 270°E–330°E, 60°S–10°N), Australia (AUS: 110°E–180°E, 50°S–0°N), Australia with some areas of South Pacific (AUS+SP: 110°E–260°E, 50°S–20°N), Northern Eurasia (NEu: 25°E–190°E, 40°N–80°N), and the Middle East (ME: 25°E–75°E, 10°N–45°N). The developed verification system for real time forecast issues monthly rolling 3-month overlapping mean for upcoming 3-month and 6-month periods, with a 1-month lead time. Currently, both the improved hindcast verification system and the newly developed real-time forecast systems have been successfully employed by the APCC as an operational tool and internally provided their information every month since 2014. All verification information of the hindcast and real-time forecasts will be available online beginning in 2015.

The present study also provides a preliminary documentation of the seasonal forecasts issued by the APCC operational multi-model forecasts, with a large set of predictions

currently available in an operational context, particularly focusing on real-time predictions of temperature and precipitation for the period of 2008JFM–2013/14DJF. The results indicate that the simple averaged multi-model prediction with equal weighting generally perform better than any single-model predictions across all of the 3-month overlapping seasons, regions, years and variables. This is consistent with many previous studies, which indicate the advantages of MME prediction (e.g., Doblas-Reyes et al. 2005; Hagedorn et al. 2005; Min et al. 2009; Wang et al. 2009; Kryjov 2012; Rodrigues et al. 2014; Min et al. 2014). The results also indicate that the simple averaged MME generally outperforms multiple regression-based weighted MMEs over most latitudinal zones for all variables and seasons, whereas the calibrated MME shows the capability of improving the forecast skill in a large proportion of cases.

Forecast skills are higher in the tropics than in the extratropics for both temperature and precipitation. This is consistent with the higher signal-to-noise ratios at low latitudes documented for troposphere geopotential height (e.g., Shukla and Kinter 2006; Kumar et al. 2007) and for the associated surface climate (e.g., Rowell 1998; Peng et al. 2010). Forecast skill over the 6-year period is strongly related to ENSO variability, particularly for precipitation due to a simultaneous positive relationship with both phases of ENSO. The skill result found here is consistent with skill evaluations by other forecast-producing centers and with theoretical predictability studies. Over a period as brief as 6 years, the APCC real-time forecast skills for both variables are commonly slightly higher in the boreal cold season and lower in the boreal warm season as compared with hindcast skills. However, the averaged skills over the whole 6-year period for both variables are relatively lower than those over the 21-year period. To investigate the current level of forecast skill of the APCC multi-model prediction, we compared it with that of other operational centers, NCEP/CPC and IRI. The results indicate that the APCC multi-model prediction shows reasonably better skills than the other two operational centers in the predictions over the 6-year period, although this is a preliminary assessment with a relatively short period. However, more comprehensive assessment with a large set of predictions should be

necessary to make well-grounded conclusions. Finally, Table 7 shows the summary of the area-averaged HSS of the APCC operational multi-model temperature and precipitation forecasts for 1-month lead forecasts for the period of 2008JFM–2013/14DJF.

APPENDIX A

Verification Metrics

1. Deterministic Forecast

a. Anomaly Pattern Correlation Coefficient (ACC)

The ACC is a pattern correlation between predicted and analyzed anomalies defined as,

$$ACC = \frac{\sum_{i=1}^{N}(f_i - \bar{f})(o_i - \bar{o})}{\sqrt{\sum_{i=1}^{N}(f_i - \bar{f})^2 \sum_{i=1}^{N}(o_i - \bar{o})^2}} \quad ,$$

where the over bar indicates time averaging. ACC indicates spatial similarity between the forecast and observation map. The score always ranges from -1.0 to 1.0. If the forecast is perfect, the ACC score is equal to 1.0.

b. Root Mean Square Error (RMSE)

The RMSE indicates a measure of accuracy of the forecast (***f***) compared with the observation (***o***). RMSE is defined as,

$$RMSE = \sqrt{\frac{1}{W}\sum_{i=1}^{N} w_i (F_i - O_i)^2}$$

where w is the latitude weight, W is the summation of w, and the subscript i is a given grid point. RMSE indicates the total difference between the forecast and observation map. The score is always greater than or equal to 0. If the forecast is perfect, the RMSE score equals 0.

c. Temporal Correlation Coefficient (TCC)

The TCC, called a linear correlation coefficient, is used to measure the strength of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the Pearson product moment correlation coefficient. The TCC between two variables is defined as the covariance of the two variables divided by the product of their standard deviations:

$$TCC = \frac{\mathrm{cov}(y,o)}{s_y s_o} \qquad \mathrm{cov} = \frac{1}{(T-1)}\sum_{t=1}^{T}(y_t - \bar{y})(o_t - \bar{o})$$

$$s_y = \sqrt{\frac{1}{(T-1)}\sum_{t=1}^{T}(y_t - \bar{y})^2} \text{ , and } \qquad s_o = \sqrt{\frac{1}{(T-1)}\sum_{t=1}^{T}(o_t - \bar{o})^2} \text{ ,}$$

where, y_t and o_t are the forecast and observed variables in a specific time series t. $\bar{y}$ ($\bar{o}$) and S_y (S_0) are the climatological value and standard deviation of the forecast (observed) variable during the period T, respectively. The TCC of 1.0 (-1.0) denotes a perfect (inverse) linear relationship between the forecast and observation and that of zero indicates the absence of any linear association between them.

d. Mean Square Skill Score (MSSS)

A detailed description of the mean square skill score is provided by WMO (2002), so only a brief description is presented here. Let O_{ij} and f_{ij} (i = 1,…, n) denote the time series of observations and continuous deterministic forecasts, respectively, for a grid point or station j over the period of verification (POV). Then, their averages for the POV, O_i and f_i and their sample variances S^2_{oj} and S^2_{fj} are given by

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \bar{f}_j = \frac{1}{n}\sum_{i=1}^{n} f_{ij}$$

$$s_{xj}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2, s_{\bar{f}j}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(f_{ij} - \bar{f}_j\right)^2 .$$

The mean square error of the forecast is

$$MSE_j = \frac{1}{n}\sum_{i=1}^{n}\left(f_{ij} - x_{ij}\right)^2$$

For the case of cross-validated POV climatology forecasts where forecast/observation pairs are reasonably temporally independent of each other (only one year at a time is withheld), the mean squared error of climatology forecasts (Murphy, 1988) is

$$MSE_{cj} = \left(\frac{n}{n-1}\right)^2 S_{xj}^2$$

The mean squared skill score for j is defined as one minus the ratio of the squared error of the forecasts to the squared error for forecasts of climatology:

$$MSSS_j = 1 - \frac{MSE_j}{MSE_{cj}}$$

An overall MSSS is computed as,

$$MSSS = 1 - \frac{\sum_j w_j MSE_j}{\sum_j w_j MSE_{cj}}$$

where $\boldsymbol{W}_j$ is the unity of verifications at a given station and is equal to cos(θ_j), where θ_j is the latitude at grid point j on the latitude-longitude grid. For MSSS, a corresponding root mean squared skill score can be obtained easily from

$$RMSSS = 1 - \left(1 - MSSS\right)^{1/2}$$

MSSS for forecasts fully cross-validated (with one year at a time withheld) can be explained as

$$MSSS_j = \left\{ 2\frac{s_{fj}}{s_{xj}} r_{fxj} - \left(\frac{s_{fj}}{s_{xj}}\right)^2 - \left(\frac{\left[\bar{f}_j - \bar{x}_j\right]}{s_{xj}}\right)^2 + \frac{2n-1}{(n-1)^2} \right\} \Bigg/ \left\{ 1 + \frac{2n-1}{(n-1)^2} \right\}$$

where r_{fxj} is the product moment correlation of the forecasts and observations at point or station j.

$$r_{fxj} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(f_{ij} - \bar{f}_j\right)\left(x_{ij} - \bar{x}_j\right)}{s_{fj}\, s_{xj}}$$

The first three terms of the decomposition of MSSS are related to phase errors (through the correlation), amplitude errors (through the ratio of the forecast to observed variances) and overall bias error, respectively, of the forecasts. These terms provide the opportunity for those wishing to use the forecasts for input into regional and local forecasts to adjust or weight the forecasts as appropriate. The last term takes into account the fact that the 'climatology' forecasts are cross-validated as well.

2. Probabilistic Forecast

a. Brier Skill Score

The most common measure of accuracy of a probabilistic forecast is the mean-square error, which is normally referred to as the Brier score (BS; Brier 1950):

$$BS = \frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2$$

where n is the number of forecasts, f_i the forecast probability of occurrence for the ith forecast, and O_i is the ith observed probability, which is defined to be 1 if the event occurs

and 0 otherwise. The BS ranges from zero to one, with lower values indicating better forecasts. The BS can be decomposed as three terms as follows (Wilks 1995):

$$BS = \bar{o}(1-\bar{o}) + \frac{1}{n}\sum_{k=1}^{m} n_k (f_k - \bar{o}_k)^2 - \frac{1}{n}\sum_{k=1}^{m} n_k (o_k - \bar{o}_k)^2 ,$$

$$= BS_{unc} + BS_{rel} - BS_{res} ,$$

where $\bar{o}$ indicates the climatological probability of the event, m indicates the number of probability bins, f_k represents the forecast probability for bin k and $\bar{o}_k$ denotes the relative frequency of occurrence of the event when the forecast probability is f_k.

These terms are known as uncertainty, reliability, and resolution, respectively. The uncertainty term depends only on the variability of the observations, and cannot be influenced by anything the forecaster may do. The second term, reliability, summarizes the calibration, or conditional bias, of the forecasts. It consists of a weighted average of the squared differences between the forecast probabilities f_k and the relative frequencies of the forecast event in each sub-sample. For reliable (or well-calibrated) forecasts, all the squared differences in the reliability term will be near zero, and their weighted average will be small. The resolution indicates the ability of the forecasts to discern sub-sample forecast periods with different relative frequencies of the event. It is a weighted average of the squared differences between these sub-sample relative frequencies, and the overall sample climatological relative frequency. Thus, if the forecasts sort the observations into sub-samples having substantially different relative frequencies than the overall sample climatology, the resolution terms will be large. Consequently, if the forecasts sort the events into sub-samples with very similar event relative frequencies, the squared differences in the summation of the resolution terms will be small. In this case, the forecasts resolve the event only weakly, and the resolution term will be small.

The BSS (Wilks 1995; Jollife and Stephenson 2003) is defined with respect to a reference forecast (e.g., a climatological forecast):

$$BSS = 1 - \frac{BS}{BS_{ref}} = \frac{BS_{res} - BS_{rel}}{BS_{unc}} \quad .$$

The BSS is one for a perfect forecast, and zero or negative for inaccurate forecasts relative to the reference forecast.

b. Relative Operating Characteristics (ROC)

The ROC is a representation of the skill of a forecasting system in which the hit rate (HR) and the false-alarm rate (FAR) are compared. The FAR is the ratio of false alarms to the total number of nonoccurrences of an event:

$$FAR = \frac{b}{b + d} \quad .$$

The HR, which is also called the probability of detection (POD), is defined as:

$$HR = POD = \frac{a}{a + c} \quad ,$$

and represents the fraction of the event occurrences that were forecasted. The HR and FAR for the probability threshold P_n are defined as:

$$HR_n = \left(\frac{\sum_{i=n}^{N} O_i}{\sum_{i=1}^{N} O_i} \right), \text{ and } \qquad FAR_n = \left(\frac{\sum_{i=n}^{N} NO_i}{\sum_{i=1}^{N} NO_i} \right) \quad .$$

The HR and FAR are calculated for each probability threshold P_n, giving N grid-points on a graph of HR (vertical axis) against FAR (horizontal axis) to form the ROC curve. This curve, by definition, must pass through the points (0, 0) and (1, 1). The ROC curve for no-skill forecasts coincides with the 45° line from the origin (i.e., the diagonal

line where HR = FAR), and the curve for perfect forecasts connects the points (0, 0), (0, 1) and (1, 1); that is, the further the curve lies toward the upper left-hand corner (where HR = 1 and FAR = 1), the better. The area under the ROC curve (ROC score) is a commonly used summary statistic representing the skill of the probabilistic forecast system. Thus, the ROC score is equal to 1 for a perfect forecast and 0.5 for a no-skill forecast.

c. Reliability diagram

The technique for constructing the reliability diagram is somewhat similar to that for the ROC curve. Instead of plotting the HR against the FAR for the accumulated probability bins, the HR is calculated only from the sets of forecasts for each probability bin separately, and is plotted against the corresponding forecast probabilities. The HR for each probability bin is defined as:

$$HR_n = \frac{O_n}{O_n + NO_n} \quad .$$

In a perfect reliable system, the forecast probability is equal to the observed frequency; thus the graph is a straight line oriented at 45° to the axis (i.e., a diagonal line). If the curve lies below the diagonal line, the probabilities are underestimated. The more flat the curve is, the lower resolution the probabilities have.

Frequency histograms show the frequency of forecasts as function of the probability bin, indicating the sharpness of the forecast. The frequency of forecasts for each probability bin is defined as:

$$F_n = \frac{O_n + NO_n}{T} \quad ,$$

where T is the total number of forecasts and calculated as:

$$T = \sum_{n=1}^{N} (O_n + NO_n)$$

d. Heidke Skill Score (HSS)

The HSS compares how often the forecast category correctly matches the observed category above the number of correct "hits" expected by chance alone. This score utilizes the number of correct and incorrect category hits. A score of 100 indicates a perfect forecast and a score of -50 indicates a perfectly incorrect forecast. Scores greater than 0 indicate improvement compared to a random forecast and indicate skill.

For monthly and seasonal forecasts, the equal chance forecast category is included in the scores. The equation for the score is

$$HSS = \frac{(hits + correct\ \ negative) - (expected\ correct)_{random}}{N - (expected\ correct)_{random}} x100$$

where $(expected\ correct)_{random} = \frac{1}{N}[(hits + misses)(hits + misses)(hits + false\ alarms) +$

$(correct\ negative + misses)(correct\ negative + false\ alarms)$

The Heidke score is considered as a simplified measure of forecast skill, easily understood by nontechnical users. Its main shortcoming is that it is insensitive to the details of the correspondence of the forecast probabilities with the relative frequencies of observed outcomes, such as probabilistic over- or under-confidence in the forecast probabilities.

e. Ranked Probability Skill Score (RPSS)

The ranked probability score (RPS) measures the squared forecast probability error, and therefore indicates to the extent to which the forecasts lack success in discriminating among differing observed outcomes, and/or have systematic biases of location and level of confidence. Thus, the score reflects the degree of a lack of discrimination, reliability, and/or resolution. The RPS is based on the squared probability error, cumulative across the three forecast categories in order from lowest to highest:

$$RPS = \frac{1}{ncat - 1} \sum_{icat=1}^{ncat} (Pcomfct_{icat} - Pcumobs_{icat})^2$$

where *icat* is the category number (1 for below normal, 2 for near normal, and 3 for above normal), *ncat* is the number of categories (3 in a tercile-based system), *Pcumfct* is the cumulative forecast probability up to category *icat*, and *Pcumobs* is the comparable term for cumulative observation "probability." The error is the squared difference between the cumulative categorical forecast probability and the corresponding cumulative observed probability in which 1 is assigned to the observed category and 0 is assigned to the other categories.

The RPSS is a skill score based on a comparison of the cumulative squared probability error (i.e., the RPS) for an actual set of forecasts, with the RPS resulting from constant issuance of all forecasts, and the climatology forecast has a probability of 0.333 for each category. The formula for RPSS is based on the RPS of the forecasts and the RPS of the constant climatology forecasts, as follows:

$$RPSS = 1 - \frac{RPS_{fct}}{RPS_{cli}}$$

where RPS_{fct} and RPS_{cli} are the RPS for the forecasts and for the climatology forecasts, respectively. When RPS_{fct} and RPS_{cli} are equal, RPSS is zero, and when RPS_{fct} is zero, RPSS reaches its maximum possible value of 1.

REFERENCES

Ahn, J. B., and H. J. Kim, 2013: Improvement of one-month lead predictability of the wintertime AO using a realistically varying solar constant for a CGCM. *Meteor. Appl.* DOI: 10.1002/met.1372.

Alessandri, A., A. Borrelli, S. Masina, P. Di Pietro, A. F. Carril, A. Cherchi, S. Gualdi, and A. Navarra, 2010: The INGV-CMCC seasonal prediction system: Improved ocean initial conditions. *Mon. Wea. Rev.*, **138**, 2930-2952.

Arribas, A., and Coauthors, 2011: The GloSea4 ensemble prediction system for seasonal forecasting. *Mon. Wea. Rev.*, **139**, 1891–1910.

Barnston, A. G., L. Shuhua, S. J. Mason, D. G. DeWitt, L. Goddard, and X. Gong, 2010: Verification of the first 11 years of IRI's seasonal climate forecast. *J. Appl. Meteor. Climate*, **49**, 493-520.

Barnston, A. G., S. J. Mason, L. Goddard, D. G. Dewitt, and S. E. Zebiak, 2003: Multimodel ensembling in seasonal climate forecasting at IRI. *Bull. Amer. Meteor. Soc.*, **84**, 1783-1796.

Beongtsson, L., U. Schlese, E. Roeckner, M. Latif, T. P. Barnett, and N. E. Graham, 1993: A two-tiered approach to long-range climate forecasting. *Science*, **261**, 1027-1029.

Collins, W. D., and Coauthors, 2006: The community climate system model version 3 (CCSM3). *J. Clim.*, **19**, 2122-2143.

Ding, Y., Y. Ni, X. Zhang, M. Li, W. Dong, Z. C. Zhao, Z. Li, and W. Shen, 2000: *Intro duction to the Short-term Climate Prediction Model System*. China Meteorological Press, Beijing, China (in Chinese).

DelSole, T., 2007: A Bayesian framework for multimodel regression. *J. Climate.*, **20**, 2810-2826.

Doblas-Reyes, F. J., R. Hagedorn, and T. N. Palmer, 2005: The rationale behind the success

of multi-model ensembles in seasonal forecasting: II. Calibration and combination, *Tellus, Ser*. A, **57**, 234–252.

Goddard, L., and M. Dilley, 2005: El Niño: Catastrophe or opportunity. *J. Climate*, **18**, 651-665.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting: I. Basic concept, *Tellus,* **57A**, 219–233.

Janowiak, J. E., and P. Xie, 1999: CAMS_OPI: a global satellite-raingauge merged product for real-time precipitation monitoring applications. *J. Climate*, **12**, 3335-3342.

Jeong, H. I., K. Ashok, B. G. Song, and Y. M. Min, 2008: Experimental 6-month hindcast and forecast simulation using CCSM3. APCC 2008 Technical Report, APEC Climate Center.

Jia, X., H. Lin, J. Y. Lee, and B. Wang, 2012: Season-dependent forecast skill of the dominant atmospheric circulation patterns over the Pacific North-American region. *J. Climate*, **25**, 7248-7265.

Kanamitsu, M., et al., 2002: NCEP-DOE AMIP-II Reanalysis (R-2). *Bull. Amer. Meteor. Soc.*, **83**, 1631–1643, doi:10.1175/BAMS-83-11-1631(2002)083<1631:NAR>2.3.CO;2.

Kang, H., C. K. Park, S.N. Hameed, and K. Ashok, 2009: Statistical downscaling of precipitation in Korea using multimodel output variables as predictors. *Mon. Wea. Rev.*, **137**, 1928–1938.

Kang, H. S., H. S. Park, J. G. Do, and K. M. Lee, 2013: http://super.kma.go.kr/eng/esrc/resources/ESRC-WP03-GloSea4.pdf.

Kang, S. C., J. Hur, and J. B. Ahn, 2014: Statistical downscaling methods based on APCC multi-model ensemble for seasonal prediction over South Korea. *Int. J. Climatol.*, doi:10.1002/joc.3952.

Kharin, V. V., and F. W. Zwiers, 2003: Improved seasonal probability forecast. *J. Climate*, **16**, 1684-1701.

Krishnamurti, T. N., C. M. Kishtawal, D. W. Shin, and C. E. Williford, (2000: Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate*, **13**, 4196-4216.

Kryjov, V. N., 2012: Seasonal climate prediction for North Eurasia. *Environ. Res.* Lett., 7, 015203, doi:10.1088/1748-9326/7/1/015203.

Kug, J. S., J. Y. Lee, I. S. Kang, B. Wang, and C. K. Park, 2008: Optimal multi-model ensemble method in seasonal climate prediction. Asia-Pacific. *J. Atmos.* Sci., **44**, 259-267.

Kumar, A., B. Jha, Q. Zhang, and L. Bounoua, 2007: A new methodology for estimating the unpredictable component of seasonal atmospheric variability. *J. Climate*, **20**, 3888-3901.

Lee, D. Y., J. B. Ahn, and K. Ashok, 2013a: Improvement of multi-model ensemble seasonal prediction skills over East Asian summer monsoon region using a climate filter concept. *J. Appl. Meteor. Climatol.*, **52**, 1127-1138.

Lee, D. Y., J. B. Ahn, K. Ashok, and A. Alessandri, 2013b: Improvement of grand multi-model ensemble prediction skills for the coupled models of APCC/ENSEMBLES using a climate filter. *Atmos. Sci.* Let., **14**, 139-145.

Lee, D. Y., K. Ashok, and J. B. Ahn, 2011: Toward enhancement of prediction skills of multimodel ensemble seasonal prediction: A climate filter concept. *J. Geophys. Res.*, **116**, D06116.

Lee, J. Y., and B. Wang, 2012: Seasonal climate prediction and predictability of atmospheric circulation, in Climate Models, edited by L. M. Druyan, pp. 19-42, InTech, Rijeka, Croatia, doi:10.5772/33782.

Lim, E. P., H. H. Hendon, S. Langford, and O. Alves, 2012: Improvements in POAMA2 for

the prediction of major climate drivers and south eastern Australian rainfall. CAWCR Tech. Rep. No. 051. Available from http://www.cawcr.gov.au/publications/technicalreports.php

Liou, C. S., J. H. Chen, C. T. Terng, F. J. Wang, C. T. Fong, T. E. Rosmond, H. C. Kuo, C. H. Shiao, and M. D. Cheng, 1997: The second generation global forecast system at the central weather bureau in Taiwan. *Wea. Forecasting*, **12**, 653–663.

Livezey, R. E., and M. M. Timofeyeva, 2008: The first decade of long-lead U.S. seasonal forecasts – Insights from a skill analysis. *Bull. Amer. Meteor. Soc.*, **89**, 843-854. Min, Y. M., V. N. Kryjov, and C. K. Park, 2009: A probabilistic multimodel ensemble approach to seasonal prediction. *Wea. Forecasting*, **24**, 812-828.

Min, Y. M., V. N. Kryjov, and C. K. Park, 2009: A probabilistic multimodel ensemble approach to seasonal prediction. *Wea. Forecasting*, **24**, 812-828.

Min, Y. M., V. N. Kryjov, and J. H. Oh, 2011: Probabilistic interpretation of regression-based downscaled seasonal ensemble predictions with the estimation of uncertainty. *J. Geophys*. Res., **116**, D08101.

Min, Y. M., V. N. Kryjov, and S. M. Oh, 2014: Assessment of APCC multi-model ensemble prediction in seasonal climate forecasting: Retrospective (1983-2003) and real-time forecasts (2008-2013). *J. Geophys*. Res., **119**, 12,132-12,150. doi:10.1002/2014JD022230.

Palmer, T. N., et al., 2004: Development of a European multi-model ensemble system for seasonal to inter-annual prediction. *Bull. Amer. Meteor*. Soc., **85**, 853–872, doi:10.1175/BAMS-85-6-853.

Peng, P., A. Jumar, A. G. Barnston, and L. Goddard, 2000: Simulation skills of the SST-forced global climate variability of the NCEP-MRF9 and the Scripps-MPI ECHAM3 models. *J. Climate*, 13, 3657-3679.

Peng, P., A. Kumar, H. van den Dool, and A. G. Barnson, 2002: An analysis of multi-model ensemble predictions for seasonal climate anomalies. *J. Geophys. Res.*, **107**(D23), 4710. doi:10.1029/2002JD002712

Reynolds, R. W., N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Wang, 2002: An improved in situ and satellite SST analysis for climate. *J. Climate*, **15**, 1609–1625.

Rodrigues, L. R. R., F. J. Doblas-Reyes, and C. A. S. Coelho, 2014: Multi-model calibration and combination of tropical seasonal sea surface temperature forecasts. *Climate Dyn.*, **42**, 597-616.

Rowell, D. P., 1998: Assessing potential seasonal predictability with an ensemble of multidecadal GCM *simulations. J. Climate*, **11**, 109-120.

Saha, S., and Coauthors, 2014: The NCEP Climate Forecast System Version 2. *J. Climate*, **27**, 2185–2208.

Shneerov, B. E., V. P. Meleshko, V. A. Matjugin, P. V. Spryshev, T. V. Pavlova, S. V. Vavulin, I. M. Shkolnik, V. A. Subov, V. M. Gavrilina, and V. A. Govorkova, 2002: The current status of the MGO global atmospheric circulation model(version-MGO-03). *MGO Proceeding*, **550**, 3–43.

Shukla, J., and J. L. Kinter III, 2006: Predictability of seasonal climate variations: A pedagogical view. Predictability of Weather and Climate, T. N. Palmer and R. Hagedorn, Eds., Cambridge University Press, 306-341.

Smith, et al., 2008: Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880-2006). *J. Climate*, **21**, 2283-2296, doi:10.1175/2007JCLI2100.1.

Sohn, S. J., Y. M. Min, J. Y. Lee, C. Y. Tam, I. S. Kang, B. Wang, J. B. Ahn, and T. Yamagata, 2012: Assessment of the long-lead probabilistic prediction for the Asian summer monsoon precipitation (1983-2011) based on the APCC multimodel system and a statistical model. *J. Geophys. Res.*, **117**, D04102.

Takaya, Y., T. Yasuda, T. Ose, and T. Nakaegawa, 2010: Predictability of the mean location of typhoon formation in a seasonal prediction experiment with a coupled general circulation model. *J. Meteor. Soc. Japan*, **88**, 799–812.

Trosnikov, I. V., V. D. Kaznacheeva, D. B. Kiktev, M. A. Tolstikh, 2005: Assessment of potential predictability of meteorological variables in dynamical seasonal modeling of atmospheric circulation on the basis of semi-Lagrangian model SL-AV. *Russian Meteor. Hydrol.*, **12**.

Van den Dool H, and Z. Toth, 1991: Why do forecast for "near normal" often fail? *Wea. Forecasting*, **6**, 76-85.

Wang, B., et al., 2009: Advance and prospectus of seasonal prediction: Assessment of the APCC/CliPAS 14-model ensemble retrospective seasonal prediction (1980-2004). *Climate Dyn.*, **33**, 93–117, doi:10.1007/s00382-008-0460-0.

World Meteorological Organization, 2002: *Standardised Verification System (SVS) for Long-Range Forecasts (LRF)*. New attachment II-9 to the manual on the GDPS, Vol. 1., WMO 485, 24 pp., Geneva, Switzerland.

Yoo, J. H., and I. S. Kang, 2005: Theoretical examination of a multi-model composite for seasonal prediction. *Geophys*. Res. Lett., **32**, L18707, doi:10.1029/2005GL023513.

Yun, W. T., L. Stefanova, A. K. Mitra V. V. Kumar, W. Dewar, and T. N. Krishnamurti, 2005: A multi-model superensemble algorithm for seasonal climate prediction using DEMETER forecasts. *Tellus*, **57A**, 280-289.

Yun, W. T., L. Stefanova, and T. N. Krishnamurti, 2003: Improvement of the multimodel superensemble technique for seasonal forecasts. *J. Climate*, **16**, 3834-3840.

Tables

Table 1. The Participating organizations and institutes in the APCC MME prediction.

Country	Organizations/Institutes
Australia	Australian Bureau of Meteorology (BoM)
Canada	Meteorological Service of Canada (MSC)
China	Beijing Climate Center (BCC)
Chinese Taipei	Central Weather Bureau of Chinese Taipei (CWB)
Italy	Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC)
Japan	Japan Meteorological Agency (JMA)
Korea	Korea Meteorological Administration (KMA) Pusan National University (PNU) APEC Climate Center (APCC)
Russia	Main Geophysical Observatory of Russia (MGO) Hydrometeorological Centre of Russia (HMC)
USA	Center for Ocean-Land-Atmosphere Studies (COLA) International Research Institute for Climate and Society (IRI) National Aeronautics and Space Administration (NASA) National Center for Environmental Prediction (NCEP)
United Kingdom	Europe Centre for Medium-Range Weather Forecasts (ECMWF)

Table 2. Description of 16 dynamical seasonal prediction models.

Model	Resolution	Ens. Size (H/F)	SST Specification (H/F)	Reference
APCC-CCSM3	T85L26	10/10	Predicted SST/ Predicted SST	Jeong et al. (2008)
BCC	T42L18	8/8	Predicted SST/ Predicted SST	Ding et al. (2000)
CMCC	T63L19	9/9	Predicted SST/ Predicted SST	Aleesandri et al. (2010)
COLA	1.875° × 1.865°, L18	10/10	Observed SST/ Predicted SST of IRI	Collins et al. (2006)
CWB	T42L18	10/10	Predicted SST/ Predicted SST	Liou et al. (1997)
GloSea5	0.83° × 0.56°, L85	12/42	Predicted SST/ Predicted SST	Kang et al. (2013)
HMC	1.125° × 1/1.40625°, L28	10/10	Persistent SST/ Persistent SST	Trosnikov et al. (2005)
IRI	T42L19	24/24	Observed SST/ Persistent SST	Barnston et al. (2003)
JMA	T95L40	5/51	Predicted SST/ Predicted SST	Takaya et al. (2010)
MGO	T42L14	6/10	Observed SST/ Persistent SST	Shneerov et al. 2002
MSC_ CANCM3	T63L31	10/10	Predicted SST/ Predicted SST	http://www.ec.gc.ca/ ccmac-cccma/ default.asp? lang=En&n=1299529F-1
MSC_ CANCM4	T63L31	10/10	Predicted SST/ Predicted SST	http://www.ec.gc.ca/ ccmac-cccma/ default.asp? lang=En&n=3701CEFE-1
NASA-GMAO	288 × 181 grid, L72	11/11	Predicted SST/ Predicted SST	http://www.gfdl.noaa.gov/ ocean-model
NCEP-CFSv2	T62L64	20/20	Predicted SST/ Predicted SST	Saha et al. (2014)
PNU	T42L18	10/10	Predicted SST/ Predicted SST	Ahn and Kim (2003)
POAMA	T47L17	33/33	Predicted SST/ Predicted SST	Lim et al. (2012)
UKMO	1.875° × 1.25°, L85	12/42	Predicted SST/ Predicted SST	Arribas et al. (2011)

Table 3. Definition of SST indices to be assessed in the APCC verification system.

Index	Definition
Nino 1+2	0-10°S, 80°-90°W
Nino 3	5°S-5°N, 90°-150°W
Nino 3.4	5°S-5°N, 120°-170°W
Nino 4	5S°-5°N, 150°-160°W
EMI	EMI=C-0.5 × (E+W) where, C: 10°S-10°N, 165°E-140°W E: 15°S-5°N, 110°W-70°W W: 10°S-20°N, 125°E-145°E
IOD	IOD=WIOD-EIOD where, WIOD: 10°S-10°N, 50°E-70°E EIOD: 10°S-0, 90°E-110°E

Table 4. Summary of the APCC operational verification system.

Parameters	Region	Deterministic forecast	Probabilistic forecast
Diagrams and scores to be produced for regions			
Basic variable (T850, T2m, PREC, Z500, SST)	GL, TR, NE, SE	MSSS ACC RMSE	ROC curve ROC score Reliability diagram
	EA, SA, NAm, SAm, AUS, AUS+SP, NEu, ME	ACC RMSE	HSS RPSS BSS
Nino Index	Nino1+2/3/3.4/4	TCC	N/A
ENSO-Modoki index	EMI	TCC	N/A
IOD index	IOD	TCC	N/A
Grid-point data for mapping			
Basic variable	Grid-point verification on a 2.5° × 2.5° grid	MSSS Difference map Ensemble spread map	ROC score HSS RPSS BSS

Table 5. Summary of verification information provided by other operational centers.

Institute	Advantage	Weakness
IRI	· Various metrics (ROC curve/score, reliability diagram, rate of return, likelihood skill score, RPSS, BSS, HSS, tendency diagram) · Relatively long periods	· Only available for all seasons together and multi-model prediction · Only available for global and tropical · Temperature and precipitation · Not available for hindcast verification
NMME (preliminary)	· Vailable for each month and 3-month mean · Available for individual models and multi-model	· Only two metrics (comparison map between obs. and prediction, ACC) · Only available for temperature and precipitation over North America · Not available for hindcast verification
WMOLC	· Various metrics (ACC, RMSE, MSS, GSS, ROC curve, reliability diagram, BSS) · Available for both real-time and hind-cast verification · Available for each month and 3-month mean · Available for individual models and multi-model	· Only available for global temperature and precipitation

Table 6. ROC scores of global and tropical temperature and precipitation for the AN and BN categories for the 6-year real-time forecast and the 21-year hindcast.

Variable	Region	6-year real-time forecast (AN/BN)	Hindcast (AN/BN)
Temperature	Globe	0.62/0.63	0.67/0.67
	Tropics	0.69/0.68	0.73/0.72
Precipitation	Globe	0.59/0.58	0.60/0.59
	Tropics	0.64/0.62	0.65/0.64

Table 7. Area-averaged HSS of the APCC multi-model temperature (precipitation) forecast for 1-month lead forecasts for the period of 2008JFM–2013/14DJF.

Region	MAM	JJA	SON	DJF
Globe	20.7 (6.8)	33.7 (6.9)	38.7 (4.7)	29.4 (5.9)
N. Extratropics	23.3 (4.8)	53.8 (4.8)	50.1 (2.0)	28.0 (0,0)
S. Extratropics	8.6 (3.9)	20.8 (5.6)	26.6 (3.5)	24.6 (7.0)
Tropics	32.9 (15.1)	16.7 (12.6)	37.6 (11.5)	38.0 (14.2)
E. Asia	29.7 (12.4)	42.2 (4.1)	41.0 (3.4)	19.9 (6.0)
S. Asia	34.4 (20.6)	14.4 (12.0)	27.0 (7.9)	26.7 (10.6)
N. America	17.9 (10.0)	50.8 (12.1)	29.2 (7.5)	12.1 (10.7)
S. America	20.9 (7.5)	15.0 (8.1)	34.7 (0.6)	27.8 (7.6)
Australia	29.9 (13.6)	33.6 (5.4)	33.1 (14.7)	34.8 (16.9)
Australia + S. Pacific	37.5 (15.8)	28.1 (16.6)	38.2 (19.9)	33.9 (19.6)
Northern Eurasia	26.9 (-2.3)	74.1 (4.4)	65.3 (4.4)	36.3 (-4.6)
Middle East	33.7 (19.4)	24.4 (-0.2)	26.9 (-2.8)	32.9 (5.9)

Appendix Tables

Table 1. Contingency table showing the four possible outcomes of a forecast of a discrete variable.

		Observed		
		Yes	No	Total
Forecast	Yes	a	b	a + b
	No	c	d	c + d
Total		a + c	b + d	n = a + b + c + d

a = the number of times that an observed event was correctly forecasted (called hits);

b = the number of times that no event occurred but the forecast called for an occurrence (called false alarms);

c = the number of times that an observed event is forecasted to not occur (called miss);

d = the number of times that an event was correctly forecasted to not occur (called a correct negative);

n = the number of forecasts

Table 2. ROC contingency table for probabilistic forecasts with different probability bins.

Bin number	Forecast probability	Observed occurrence	Observed non-occurrence
1	0-P_1(%)	O_1	NO_1
2	P_2-P_3 (%)	O_2	NO_2
…	…	…	…
n	P_n - P_{n+1} (%)	O_n	NO_n
…	…	…	…
N	P_N-100 (%)	O_N	NO_N

n = the number of the n^{th} probability interval or bin n; n goes from 1 to N;

P_n = the lower probability limit for bin n;

P_{n+1} = the upper probability limit for bin n;

N = the number of probability intervals of bins;

$O_n = \sum W_i(O_i)$, (O) is 1 when an event corresponding to a forecast in bin n is observed as an occurrence; otherwise it is 0. The summation is over all forecasts in bin n, at all grid points or stations;

$NO_n = \sum W_i(NO_i)$, (NO) is 1 when an event corresponding to a forecast in bin n is not observed; otherwise it is 0. The summation is over all forecasts in bin n, at all grid points i or stations i;

W_i = 1 when verification is done at stations or at single grid points; $W_i = cos(\theta_i)$ at grid point i , when verification is done on a grid; θ_i is the latitude at grid point i

Figure Captions

Figure 1. Selected target areas for the APCC operational real-time verification system.

Figure 2. Time series of the anomaly pattern correlation for 1-month lead, 3-month mean single-model, and multi-model prediction of temperature over the (a) Globe, (b) Northern Extratropics, (c) Tropics and (d) East Asia for the period of 2008JFM–2013/14DJF.

Figure 3. Same as Fig. 2, except for precipitation.

Figure 4. Spatial distribution of temporal correlation for the simple averaged multi-model ensemble with equal weighting of temperature and precipitation for the warm seasons (AMJ, MJJ, JJA, JAS, and ASO) and cold seasons (OND, NDJ, DJF, JFM, and FMA) for the period of 2008–2013. The area-averaged scores are also displayed in the plot. The contour lines indicate that the estimated score is statistically significant at the 5% level using a one-tailed Student t-test.

Figure 5. Time series of ROC score aggregated over (a) the globe and (b) tropics for three-categorical probabilistic multi-model forecasts (above-, near-, and below-normal) for the period of 2008JFM to 2013/14DJF for temperature. The black line shows the absolute amplitude of the Nino 3.4 SST observations. The 21-year hindcast skill for each category is also displayed with a dashed line for comparison.

Figure 6. Same as Fig. 5, except for precipitation.

Figure 7. Zonal mean temporal correlation of the four different MME prediction systems for (a) temperature and (b) precipitation for the period of 2008JFM–2013/14DJF. The horizontal black lines correspond to the estimated score with statistical significance at the 1% level using the one-tailed Student t-test. The vertical lines and markers indicate the range of the highest and lowest skill and the averaged skill across the globe obtained from each prediction system.

Figure 8. Anomaly pattern correlation of the four different multi-model prediction systems

for (a) temperature and (b) precipitation over the Globe, Northern Extratropics, Tropics, and East Asia, during the period of 2008JFM–2013/14DJF.

Figure 9. Verification region for the United States defined by the CPC in NCEP.

Figure 10. Time series of HSS for the United States of the APCC multi-model forecasts and CPC manual forecasts for the period of 2008JFM to 2013/14DJF for temperature and precipitation. The red and blue lines show the averaged skills for APCC and CPC forecasts over the whole period.

Figure 11. Time series of HSS for global and tropical predictions for temperature of the APCC and IRI multi-model forecasts for the period of 2008JFM to 2013/14DJF.

Figure 12. Same as Fig. 11, except for precipitation.

Figure 13. HSS for temperature at 2m (t2m) and 850hPa (t850) over the globe and tropics for the period of 1983–2003 for each season.

Appendix Figure Captions

Figure 1. Area-averaged RMSE (left) and ACC (right) of single-model and multi-model predictions of temperature at 2m for 2014 JFM.

Figure 2. Time series of ACC for single-model and multi-model predictions of temperature at 2m in the JJA season during the period of 1983–2005.

Figure 3. Difference map of the forecasts from the single-model (left) and multi-model (right) forecasts with observations of temperature at 2m for 2014 JFM.

Figure 4. Ensemble spread map of the NCEP (left) and MSC_CANCM4 (right) simulations of temperature at 2m for 2014 JFM.

Figure 5. MSSS and its correlation map of the multi-model prediction (SCM) of temperature at 2m for the boreal autumn season (SON) during the period of 1983-2005.

Figure 6. Spatial distribution of ROC scores for three categorical probabilistic multi-model predictions (above-, near-, and below-normal) in the JJA season during the period of 1983–2005.

Figure 7. Aggregated ROC curve and score over the globe for three categorical probabilistic multi-model predictions (above-, near-, and below-normal) of precipitation for 2013 DJF.

Figure 8. Reliability diagram and frequency histogram over the globe for three categorical probabilistic multi-model predictions (above-, near-, and below-normal) of precipitation for 2013 DJF.

Figure 9. Regionally averaged HSS for three categorical probabilistic multi-model predictions of precipitation for 2013 DJF with corresponding hindcast skill.

Figure 10. Spatial distribution for three categorical probabilistic multi-model predictions of HSS of precipitation in the JJA season during the period of 1983–2005.

Figure 11. Spatial distribution of BSS for three categorical probabilistic multi-model

predictions (above-, near-, and below-normal) in the JJA season during the period of 1983–2005.

Figure 12. Regionally averaged BSS over the globe for three categorical probabilistic multi-model predictions (above-, near-, and below-normal) of precipitation for 2014 JJA.

Figure 13. Regionally averaged RPSS for three categorical probabilistic multi-model predictions of precipitation for 2013 DJF with corresponding hindcast skill.

Figure 14. Spatial distribution of RPSS for three categorical probabilistic multi-model predictions of precipitation for 2014MAM.

Figure 15. Hovmoller diagram (averaged between 5°S–5°N) of MSC_CANCM3 SST prediction for the period of 2014FMAMJJ.

Figure 16. Spatial distribution of observations and multi-model ensemble predictions for the period of 2014MAMJJA.

Figure 17. Comparison between the observed and forecasted Nino 3.4 index obtained from the multi-model prediction for the period of 2014MAMJJA.

Figure 18. Temporal correlation coefficient of the Nino 3.4 index obtained from the multi-model prediction for JASOND during the period of 1983–2003.

Figures

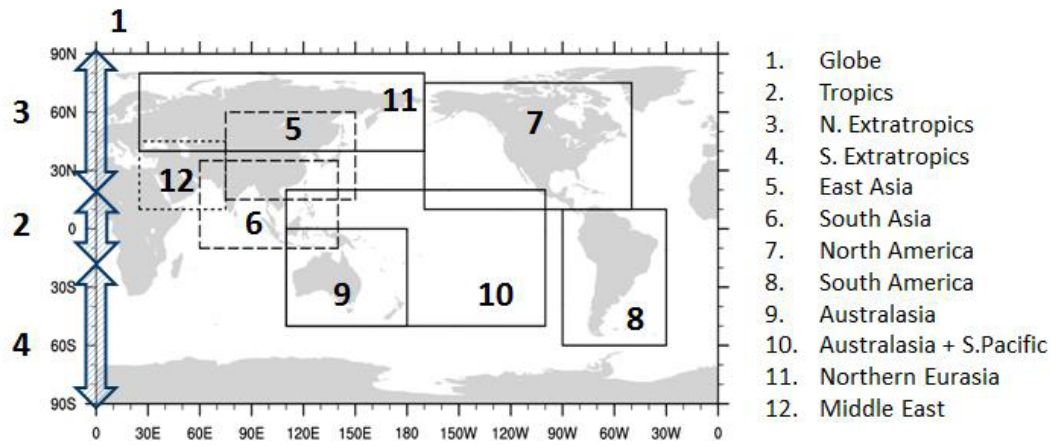

Figure 1. Selected target areas for the APCC operational real-time verification system.

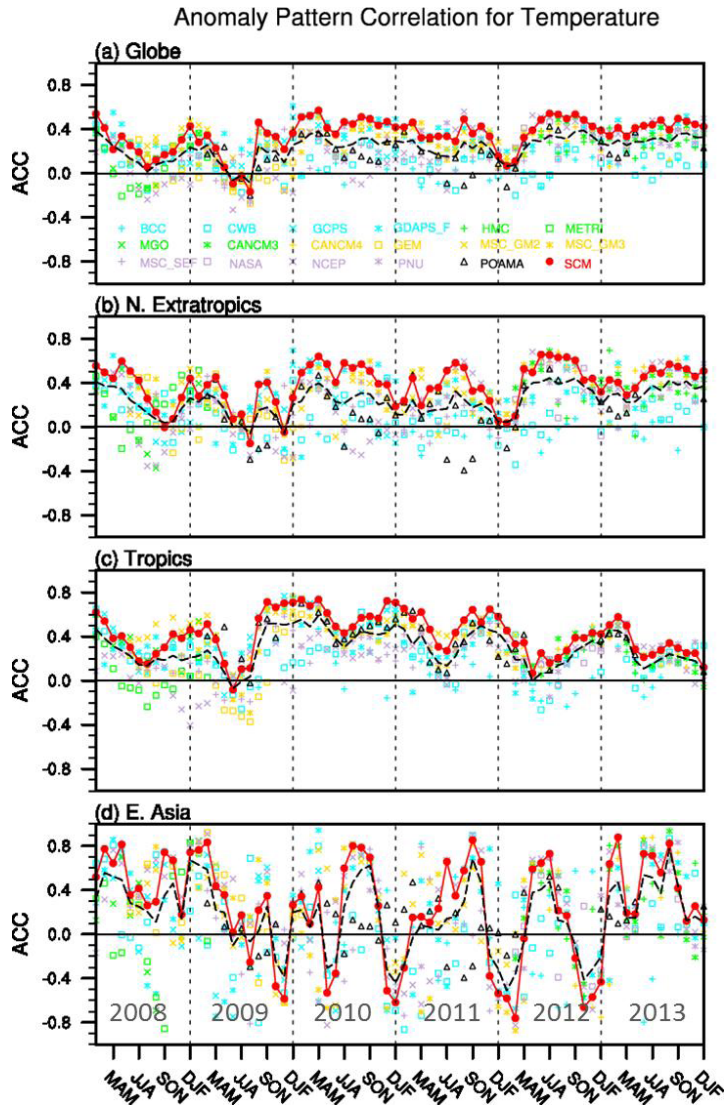

Figure 2. Time series of the anomaly pattern correlation for 1-month lead, 3-month mean single-model, and multi-model prediction of temperature over the (a) Globe, (b) Northern Extratropics, (c) Tropics and (d) East Asia for the period of 2008JFM–2013/14DJF.

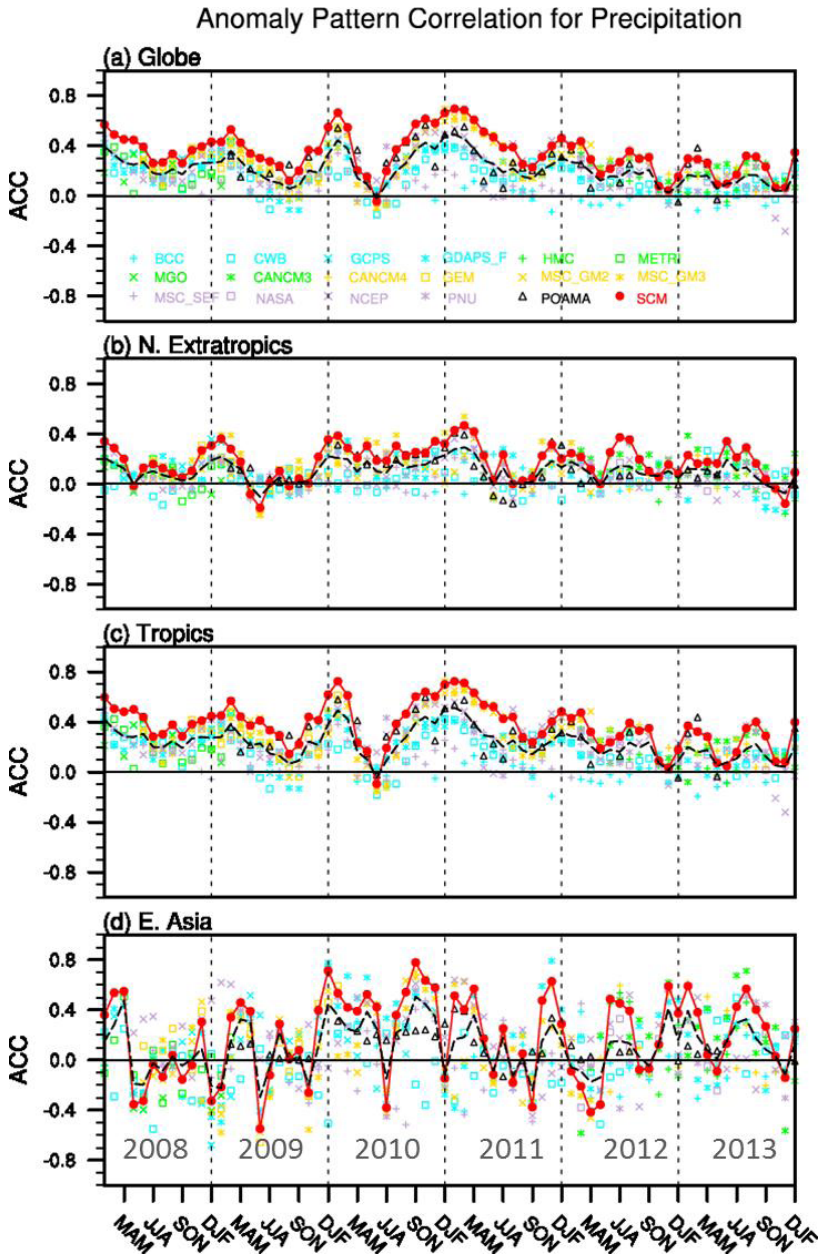

Figure 3. Same as Fig. 2, except for precipitation.

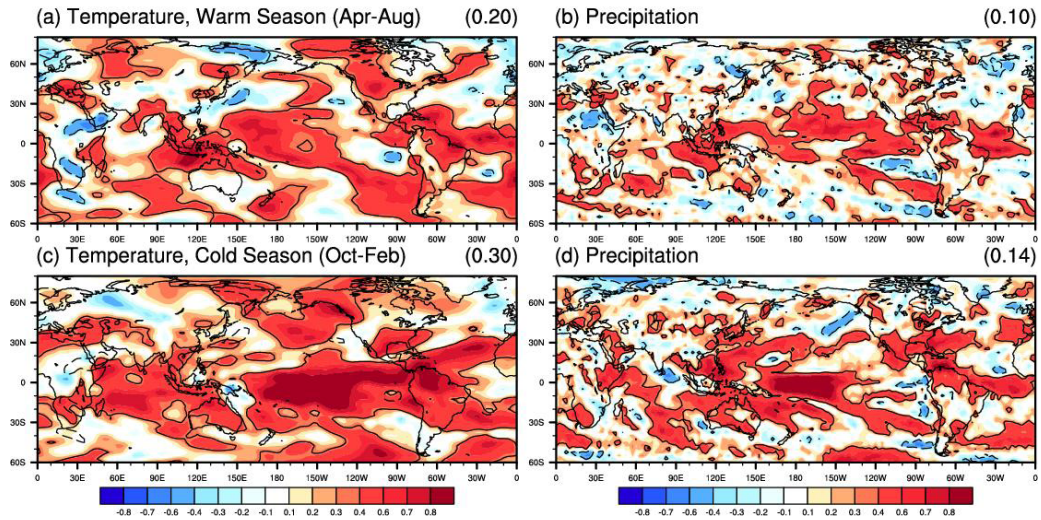

Figure 4. Spatial distribution of temporal correlation for the simple averaged multi-model ensemble with equal weighting of temperature and precipitation for the warm seasons (AMJ, MJJ, JJA, JAS, and ASO) and cold seasons (OND, NDJ, DJF, JFM, and FMA) for the period of 2008–2013. The area-averaged scores are also displayed in the plot. The contour lines indicate that the estimated score is statistically significant at the 5% level using a one-tailed Student t-test.

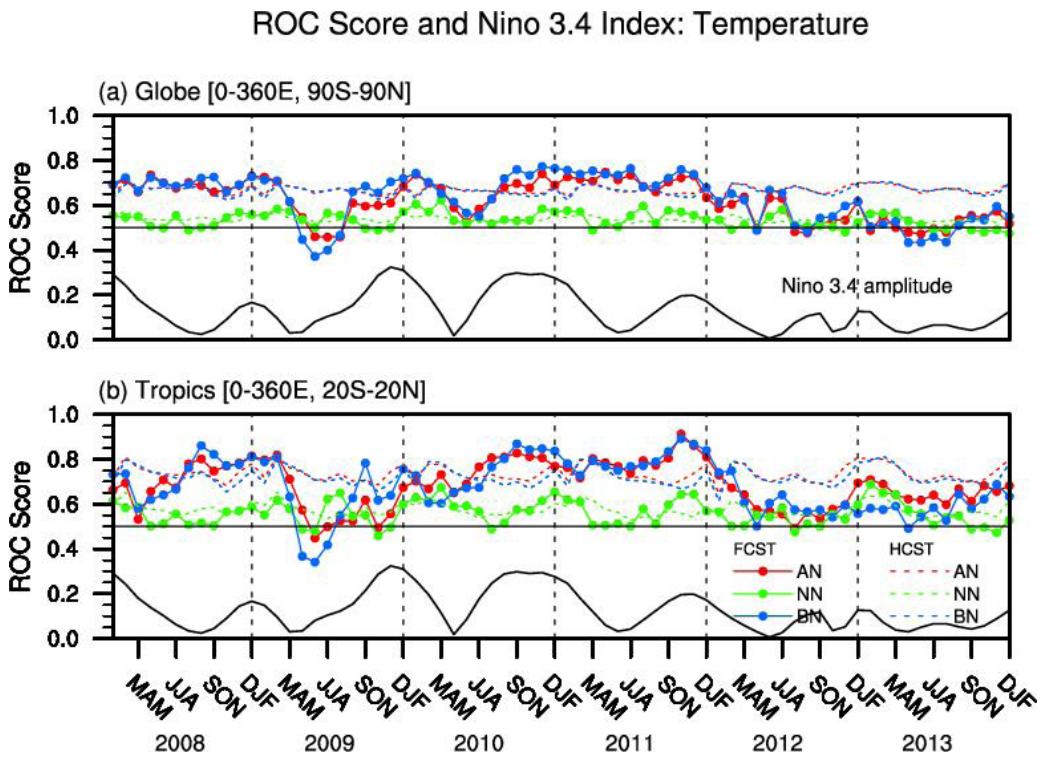

Figure 5. Time series of ROC score aggregated over (a) the globe and (b) tropics for three-categorical probabilistic multi-model forecasts (above-, near-, and below-normal) for the period of 2008JFM to 2013/14DJF for temperature. The black line shows the absolute amplitude of the Nino 3.4 SST observations. The 21-year hindcast skill for each category is also displayed with a dashed line for comparisown.

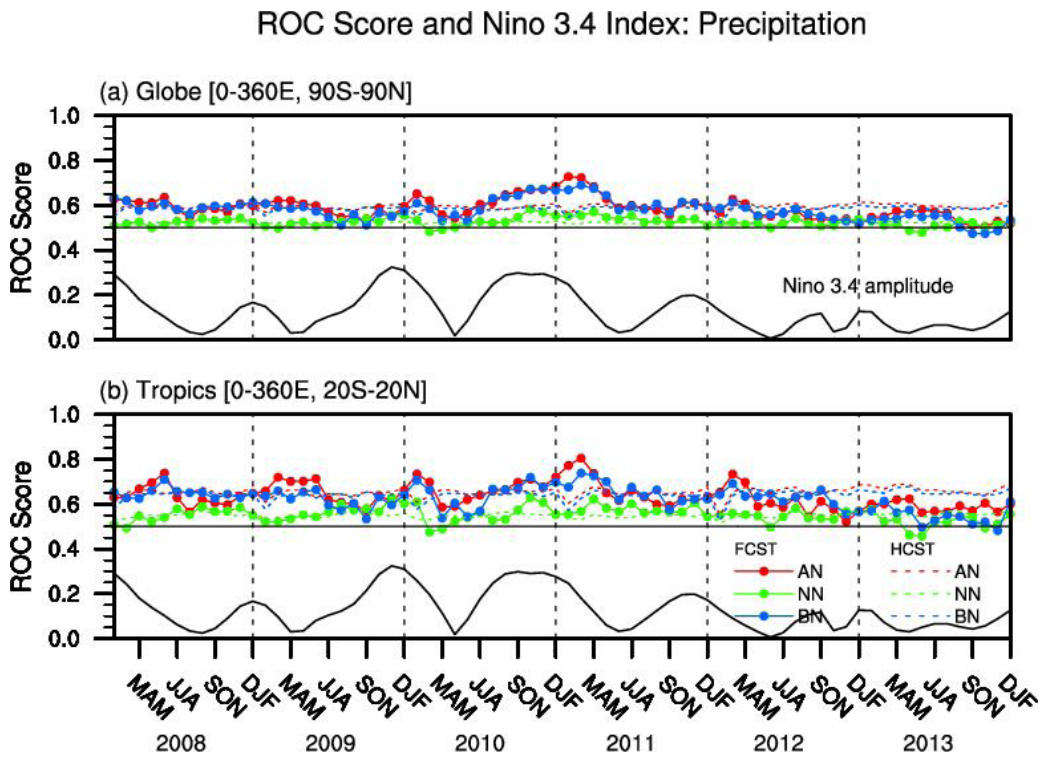

Figure 6. Same as Fig. 5, except for precipitation.

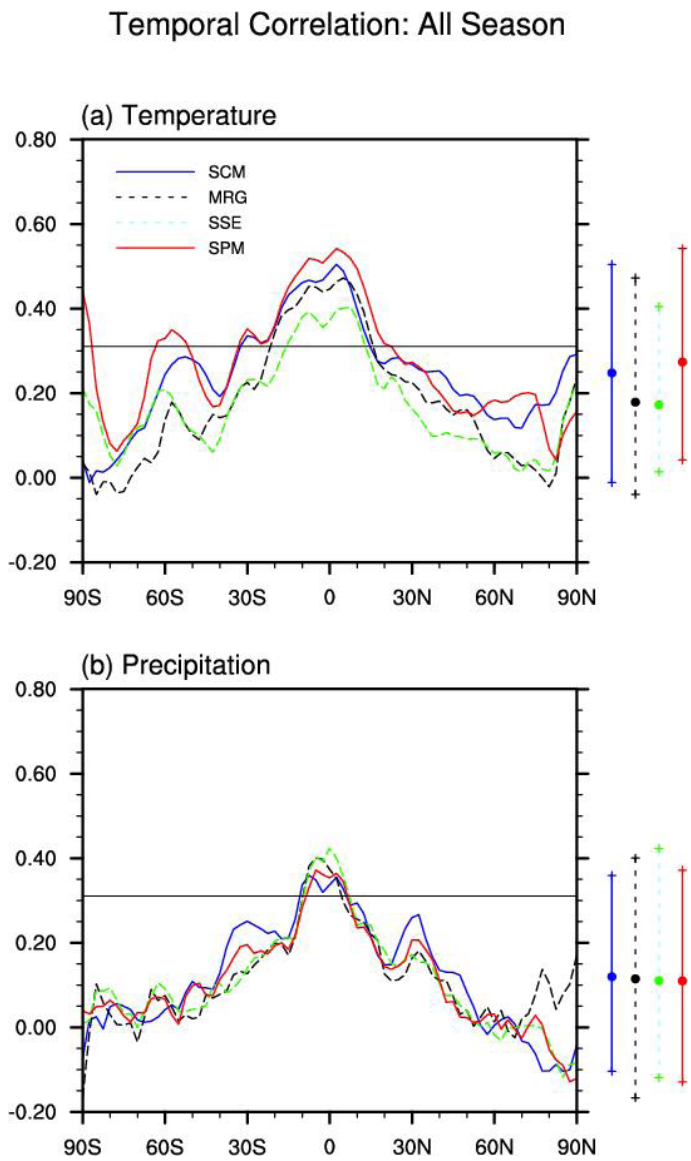

Figure 7. Zonal mean temporal correlation of the four different MME prediction systems for (a) temperature and (b) precipitation for the period of 2008JFM–2013/14DJF. The horizontal black lines correspond to the estimated score with statistical significance at the 1% level using the one-tailed Student t-test. The vertical lines and markers indicate the range of the highest and lowest skill and the averaged skill across the globe obtained from each prediction system.

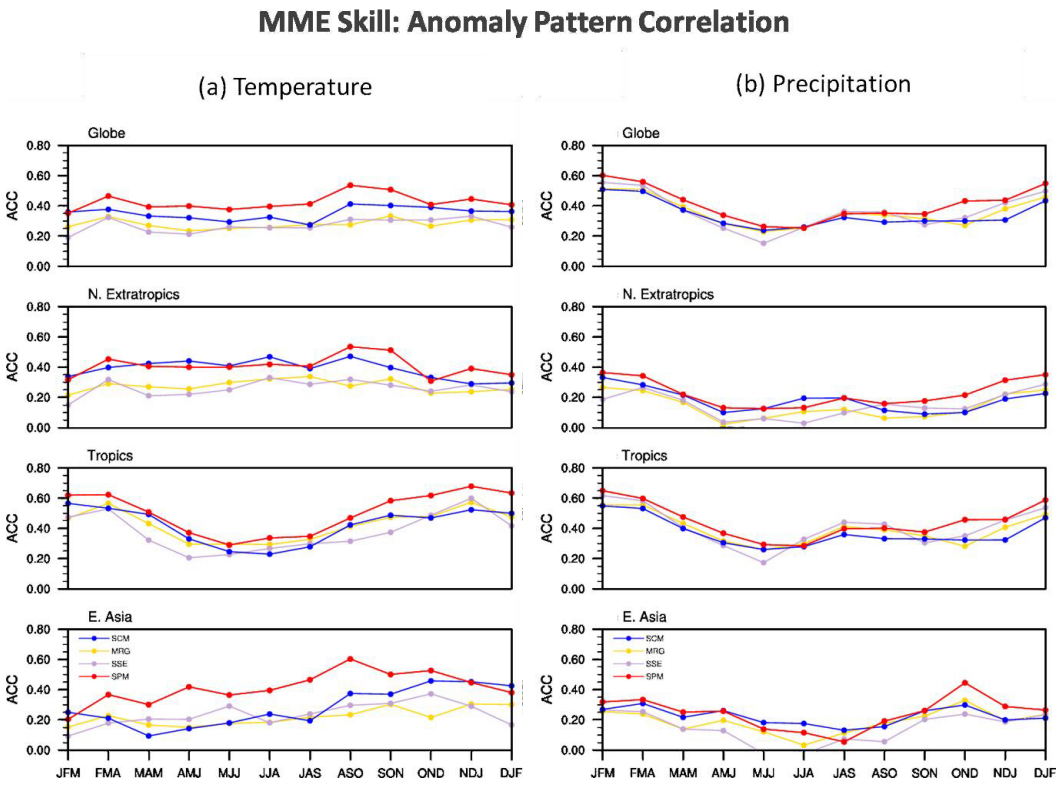

Figure 8. Anomaly pattern correlation of the four different multi-model prediction systems for (a) temperature and (b) precipitation over the Globe, Northern Extratropics, Tropics, and East Asia, during the period of 2008JFM–2013/14DJF.

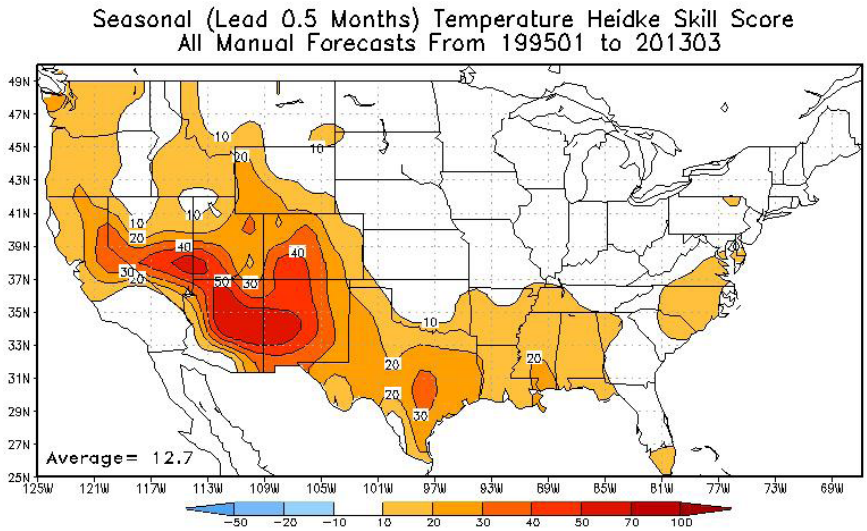

Figure 9. Verification region for the United States defined by the CPC in NCEP.

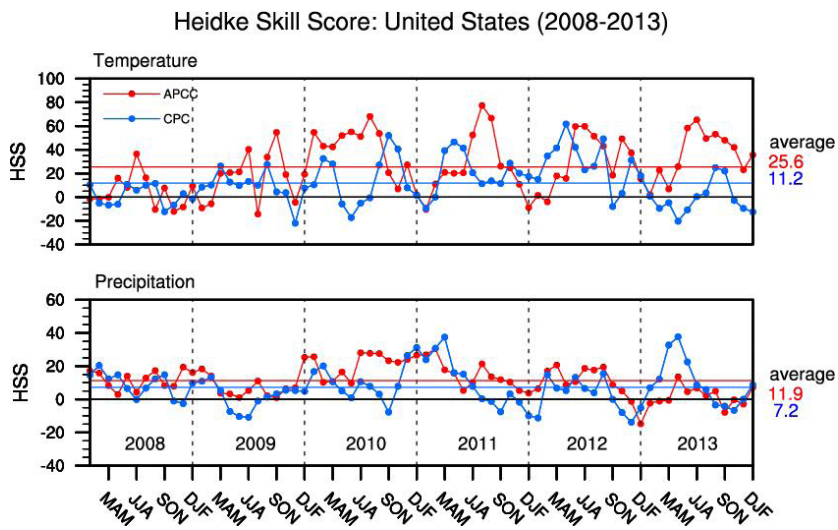

Figure 10. Time series of HSS for the United States of the APCC multi-model forecasts and CPC manual forecasts for the period of 2008JFM to 2013/14DJF for temperature and precipitation. The red and blue lines show the averaged skills for APCC and CPC forecasts over the whole period.

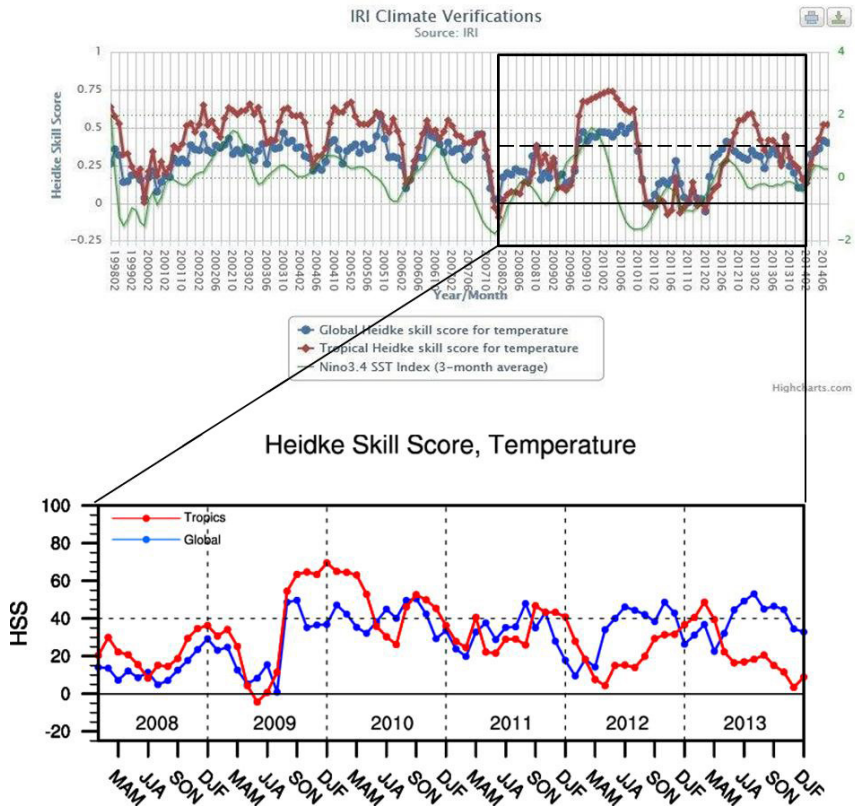

Figure 11. Time series of HSS for global and tropical predictions for temperature of the APCC and IRI multi-model forecasts for the period of 2008JFM to 2013/14DJF.

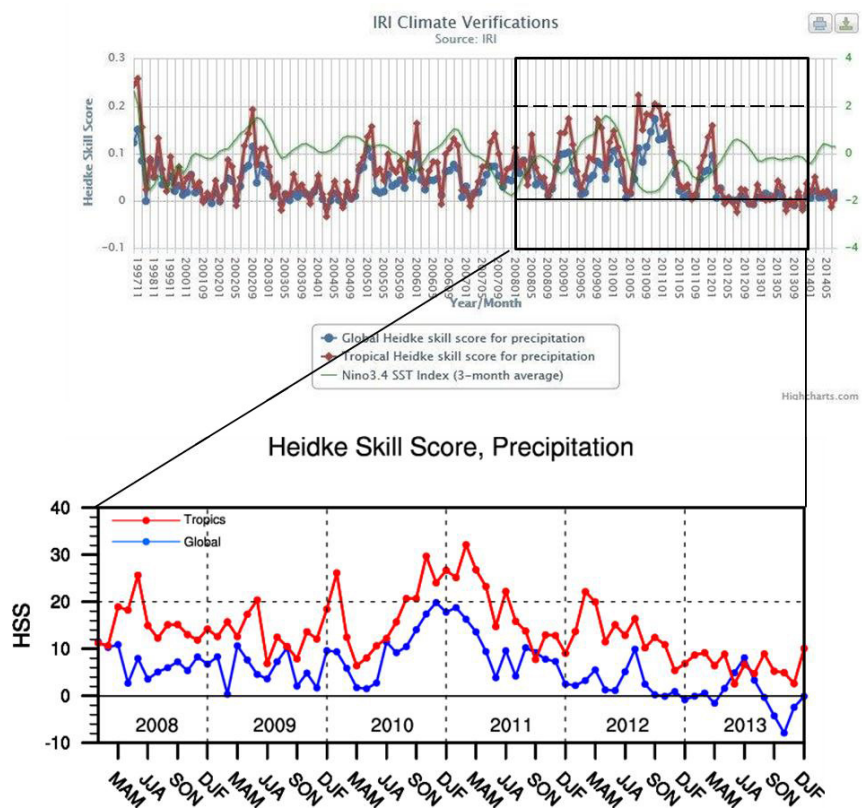

Figure 12. Same as Fig. 11, except for precipitation.

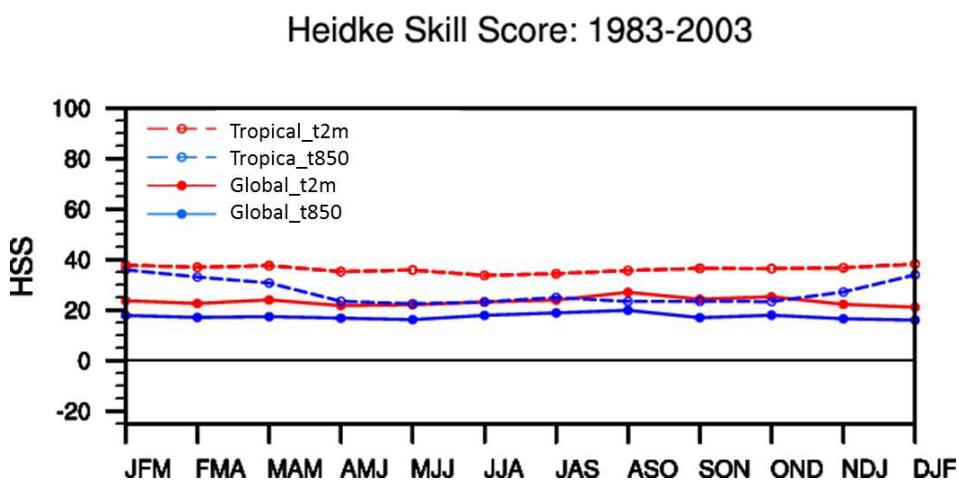

Figure 13. HSS for temperature at 2m (t2m) and 850hPa (t850) over the globe and tropics for the period of 1983–2003 for each season.

Appendix Figures

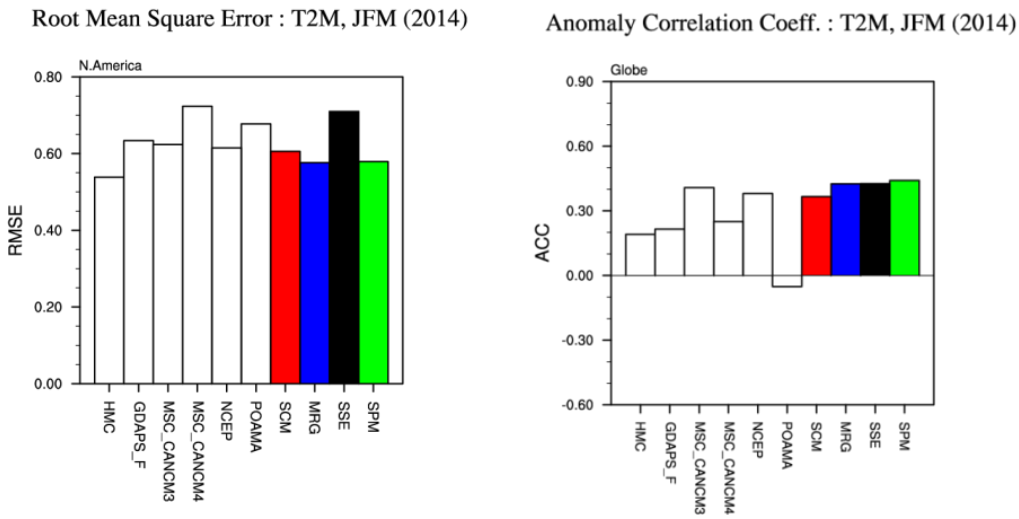

Figure 1. Area-averaged RMSE (left) and ACC (right) of single-model and multi-model predictions of temperature at 2m for 2014 JFM

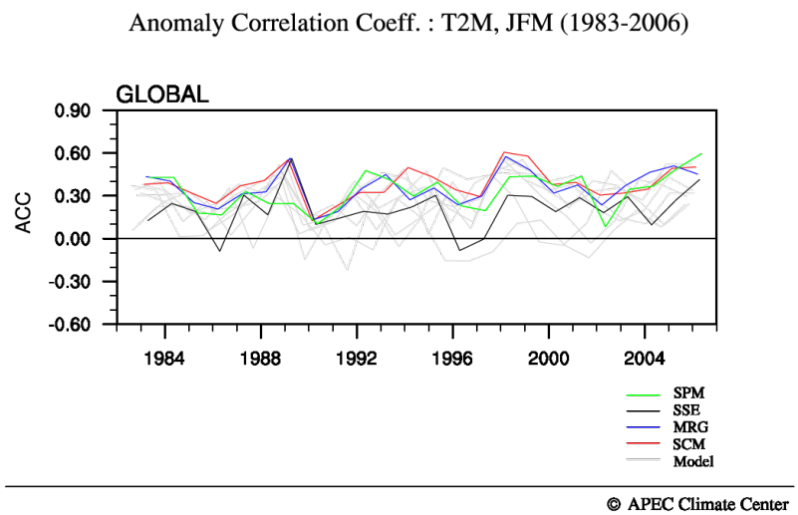

Figure 2. Time series of ACC for single-model and multi-model predictions of temperature at 2m in the JJA season during the period of 1983–2005.

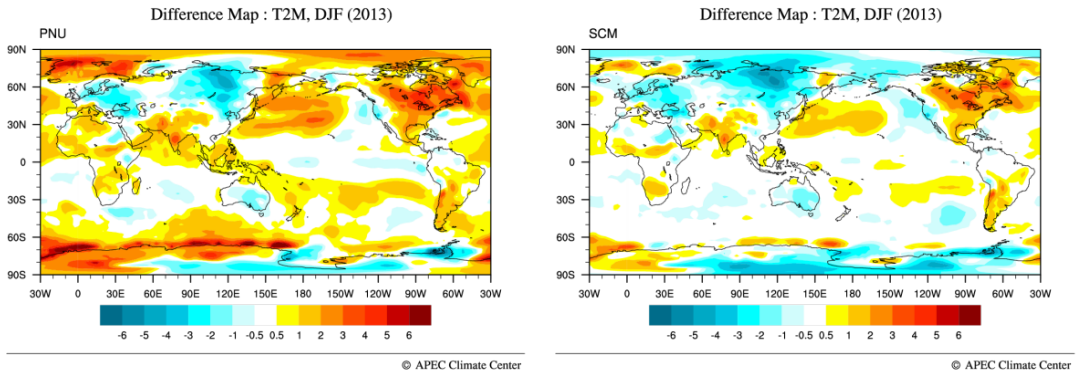

Figure 3. Difference map of the forecasts from the single-model (left) and multi-model (right) forecasts with observations of temperature at 2m for 2014 JFM.

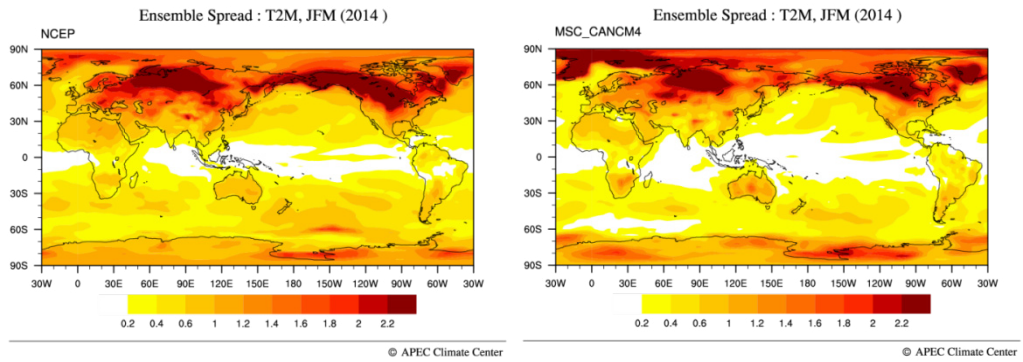

Figure 4. Ensemble spread map of the NCEP (left) and MSC_CANCM4 (right) simulations of temperature at 2m for 2014 JFM.

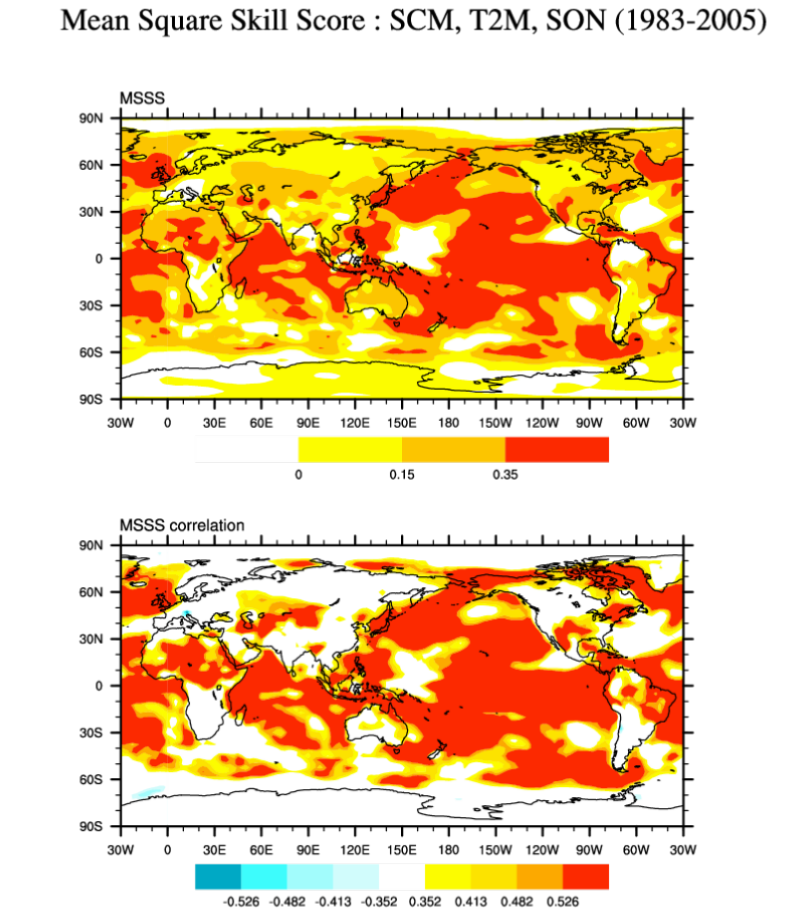

Figure 5. MSSS and its correlation map of the multi-model prediction (SCM) of temperature at 2m for the boreal autumn season (SON) during the period of 1983-2005.

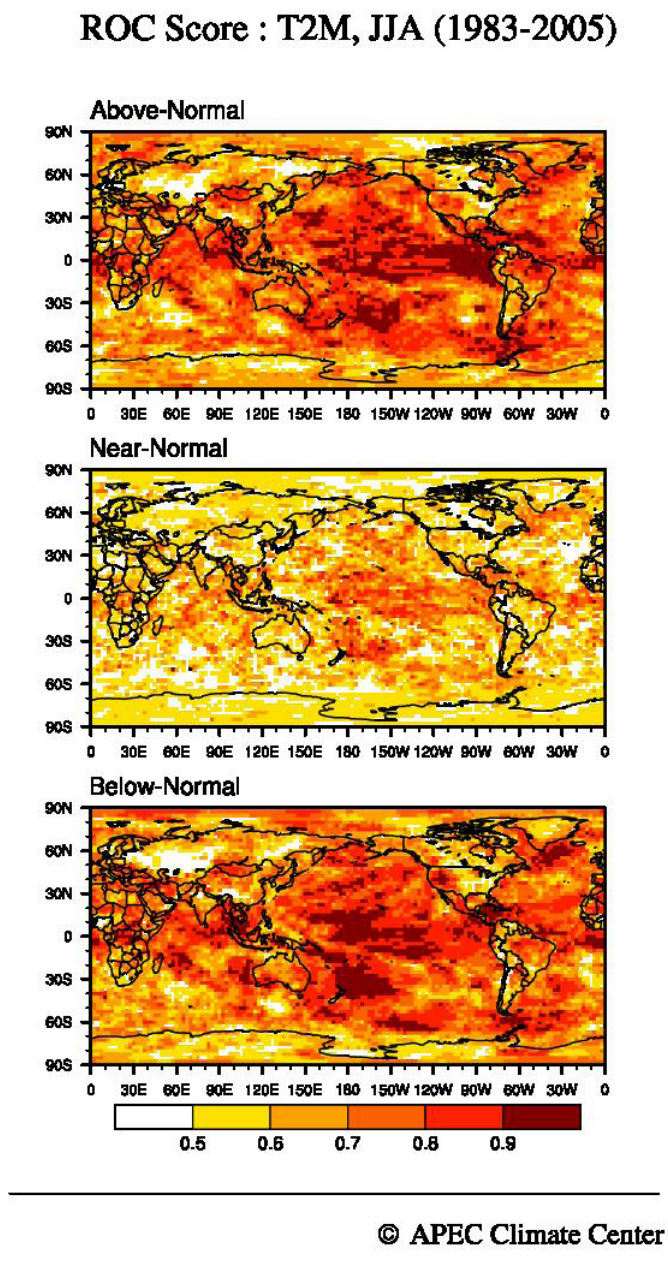

Figure 6. Spatial distribution of ROC scores for three categorical probabilistic multi-model predictions (above-, near-, and below-normal) in the JJA season during the period of 1983–2005.

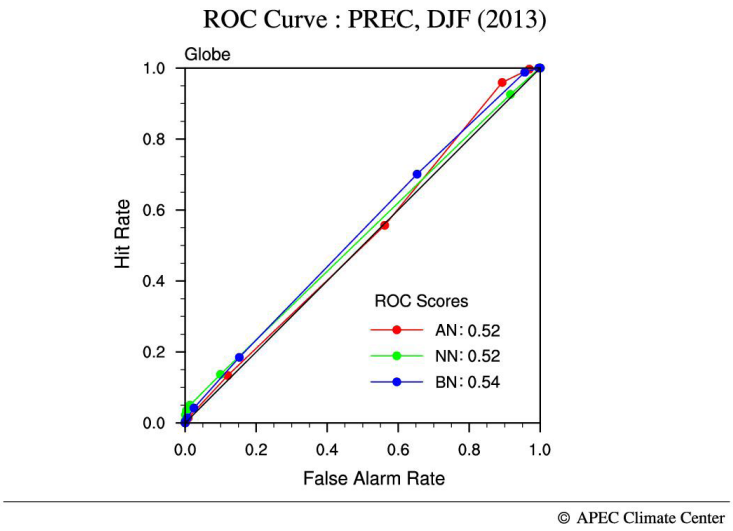

Figure 7. Aggregated ROC curve and score over the globe for three categorical probabilistic multi-model predictions (above-, near-, and below-normal) of precipitation for 2013 DJF.

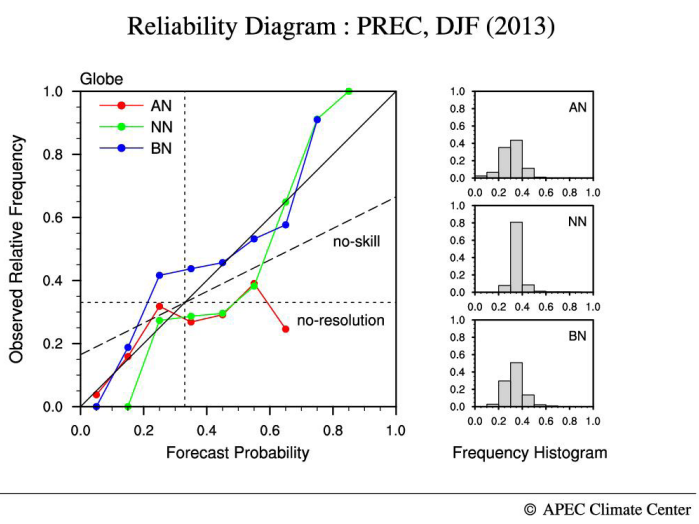

Figure 8. Reliability diagram and frequency histogram over the globe for three categorical probabilistic multi-model predictions (above-, near-, and below-normal) of precipitation for 2013 DJF.

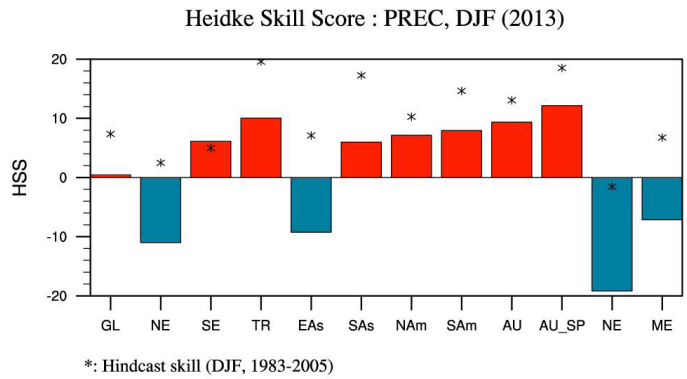

Figure 9. Regionally averaged HSS for three categorical probabilistic multi-model predictions of precipitation for 2013 DJF with corresponding hindcast skill.

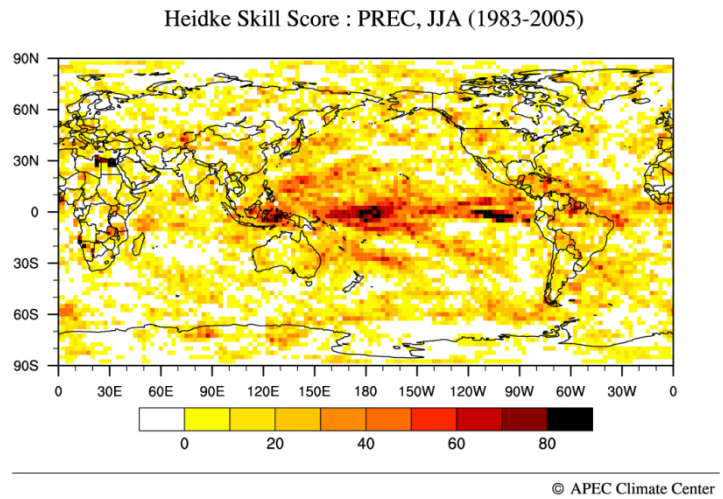

Figure 10. Spatial distribution for three categorical probabilistic multi-model predictions of HSS of precipitation in the JJA season during the period of 1983–2005.

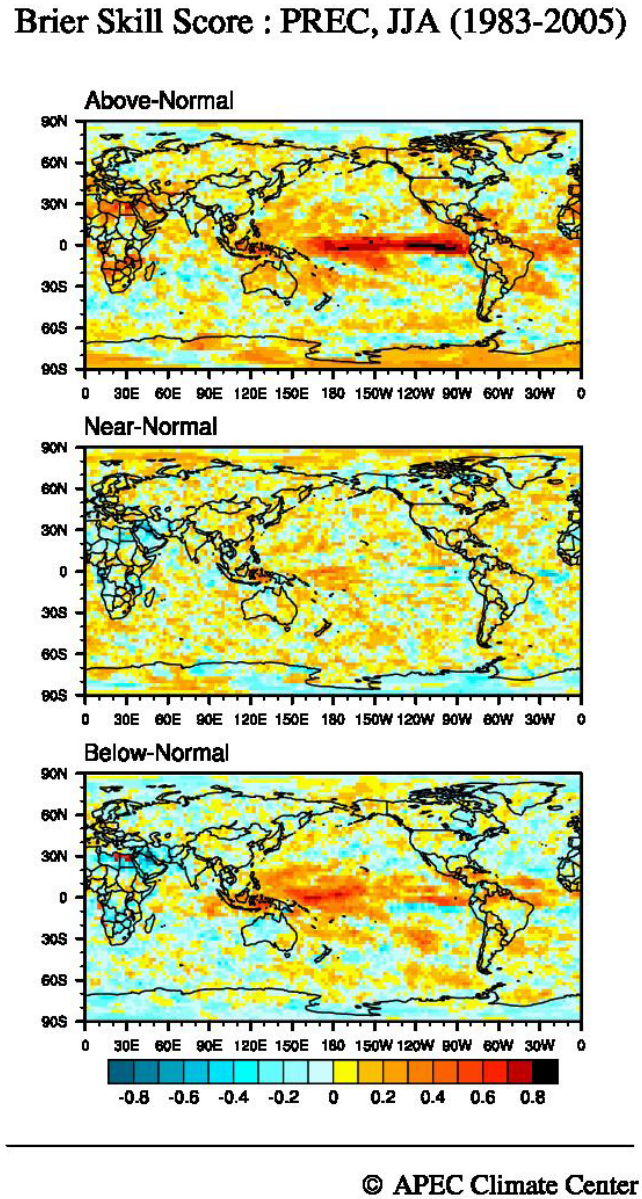

Figure 11. Spatial distribution of BSS for three categorical probabilistic multi-model predictions (above-, near-, and below-normal) in the JJA season during the period of 1983–2005.

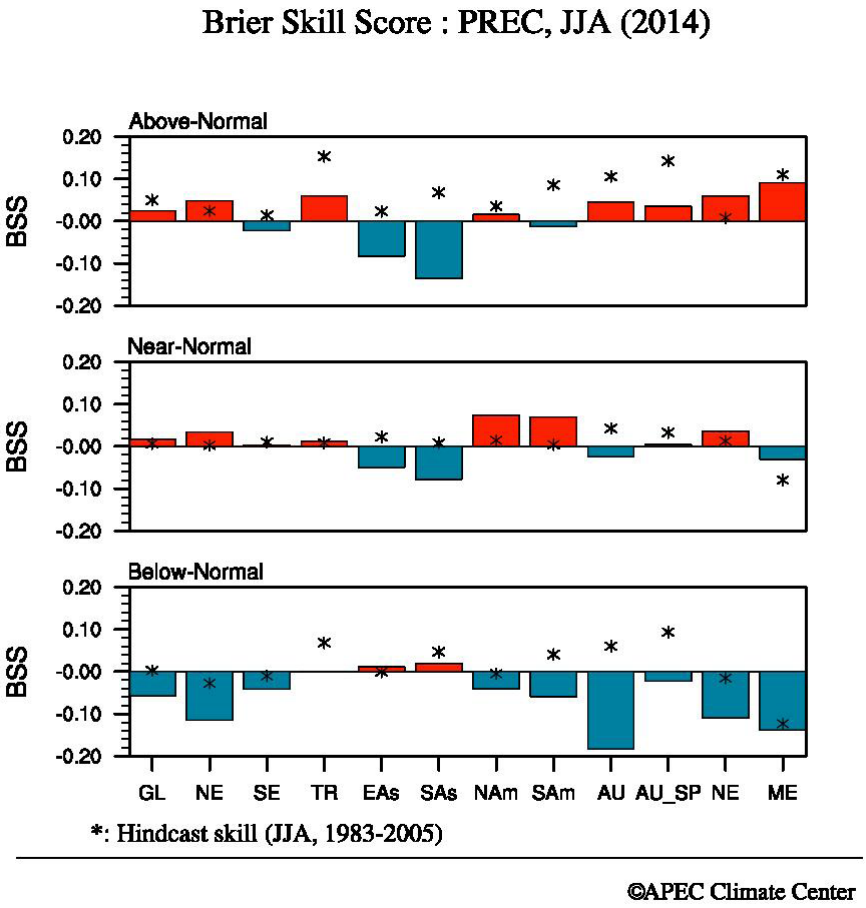

Figure 12. Regionally averaged BSS over the globe for three categorical probabilistic multi-model predictions (above-, near-, and below-normal) of precipitation for 2014 JJA.

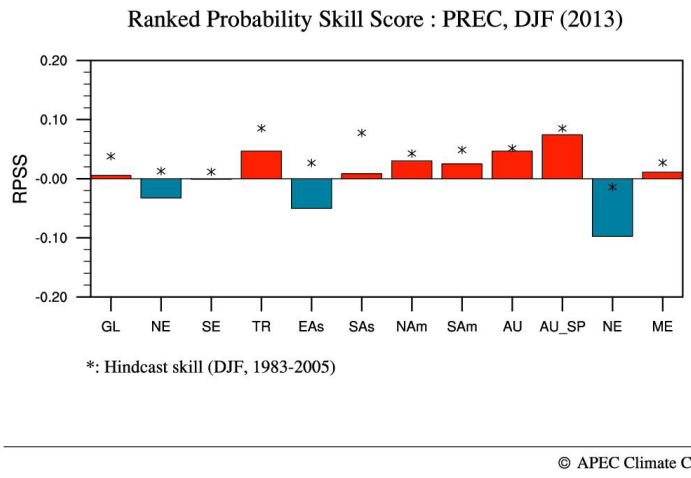

Figure 13. Regionally averaged RPSS for three categorical probabilistic multi-model predictions of precipitation for 2013 DJF with corresponding hindcast skill.

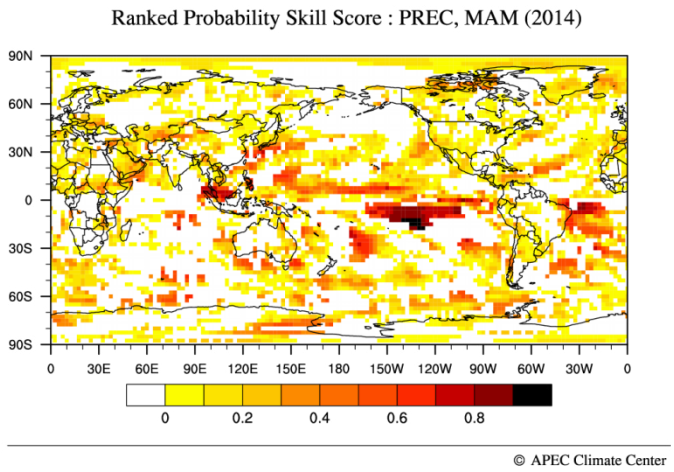

Figure 14. Spatial distribution of RPSS for three categorical probabilistic multi-model predictions of precipitation for 2014MAM.

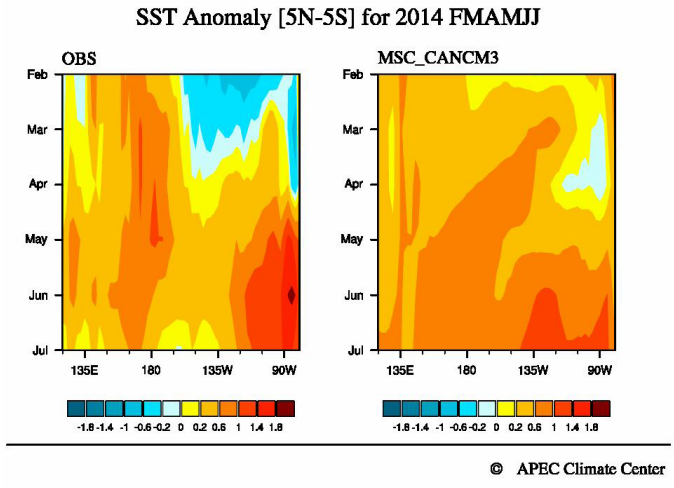

Figure 15. Hovmoller diagram (averaged between 5°S–5°N) of MSC_CANCM3 SST prediction for the period of 2014FMAMJJ.

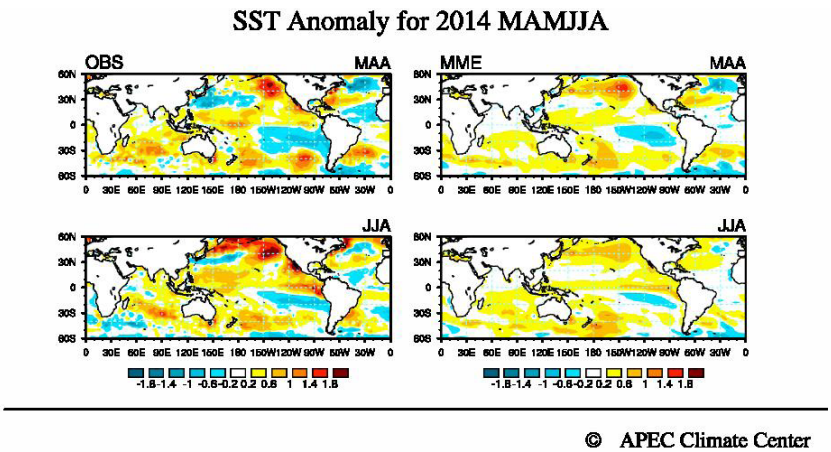

Figure 16. Spatial distribution of observations and multi-model ensemble predictions for the period of 2014MAMJJA.

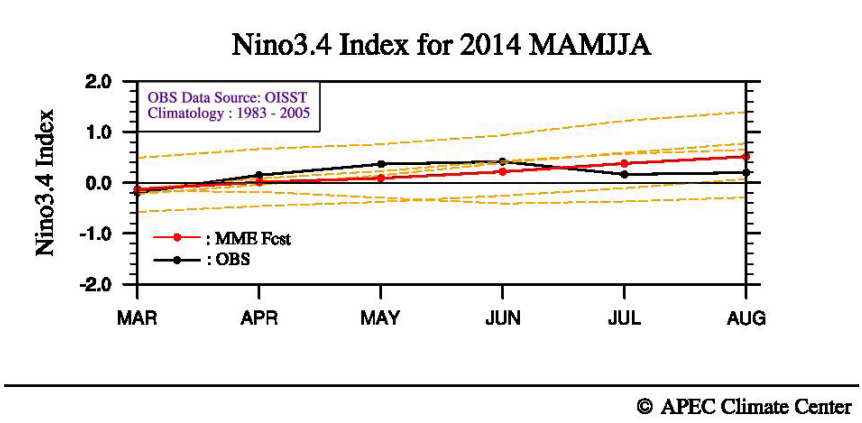

Figure 17. Comparison between the observed and forecasted Nino 3.4 index obtained from the multi-model prediction for the period of 2014MAMJJA.

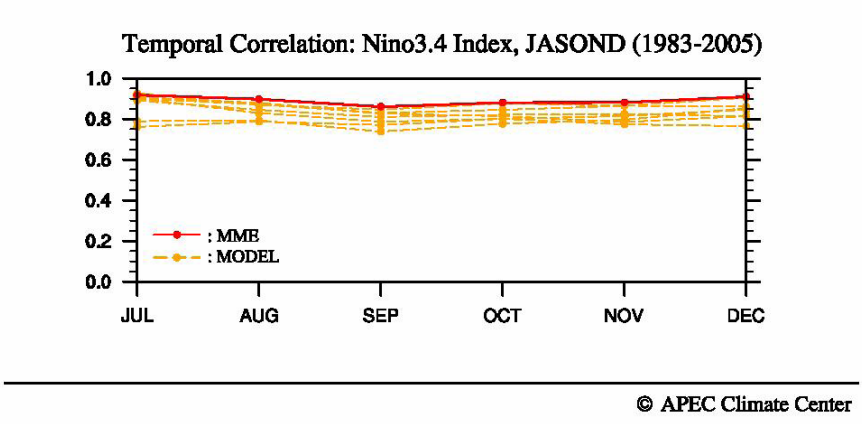

Figure 18. Temporal correlation coefficient of the Nino 3.4 index obtained from the multi-model prediction for JASOND during the period of 1983–2003.

RESEARCH REPORT 2015-03

Development of a Real-Time Verification System for the APCC Operational Multi-Model Ensemble Prediction

Young-Mi Min Climate Prediction Team

APEC Climate Center
12 Centum 7-ro, Haeundae-gu, Busan 612-020, Republic of Korea
Tel: +82-51-745-3900 Fax: +82-51-745-3949

www.apcc21.org

www.facebook.com/apcc21 www.twitter.com/apcc21 www.flickr.com/apcc21
www.youtube.com/APECClimateCenter21 www.plus.google.com/+APECClimateCenter21