Are CMIP5 Models Better than CMIP3 Models in Simulating Precipitation over East Asia?

SHOJI KUSUNOKI

Climate Research Department, Meteorological Research Institute, Tsukuba, Japan

OSAMU ARAKAWA*

Faculty of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Japan

(Manuscript received 19 August 2014, in final form 19 March 2015)

ABSTRACT

The performance of climate models participating in phases 5 and 3 of the Coupled Model Intercomparison Project (CMIP5 and CMIP3, respectively) is evaluated and compared with respect to precipitation over East Asia (20°–50°N, 110°–150°E). The target period covers the 20 years from 1981 through 2000. The CMIP5 and CMIP3 models underestimate precipitation amounts over East Asia in the warmer season (May–September), while they overestimate precipitation amounts in the colder season (October–April). Both sets of models have some difficulty in simulating the seasonal march of the rainy season over China, the Korean Peninsula, and Japan, and they also underestimate the precipitation intensity over East Asia. Nevertheless, the CMIP5 models show a higher reproducibility of precipitation over East Asia than the CMIP3 models with respect to the geographical distribution of precipitation throughout the year, seasonal march of the rainy season, and extreme precipitation events. Models with a higher reproducibility of annual precipitation tend to show a higher reproducibility of precipitation intensity for both the CMIP5 and CMIP3 models. Correlation analysis using all of the CMIP5 and CMIP3 models reveals that models with higher horizontal resolution tend to perform better than those with a lower resolution. The advantage of the CMIP5 models over the CMIP3 models in the simulation of the East Asian climate can be partly attributed to the improved representation of the west Pacific subtropical high in the CMIP5 models, especially during the summer.

1. Introduction

In response to a proposed activity of the World Climate Research Programme (WCRP), the Working Group on Coupled Modeling (WGCM) established phase 3 of the Coupled Model Intercomparison Project (CMIP3) as a standard protocol to be used in numerical experiments by climate models (Meehl et al. 2007). A set of experiments completed during CMIP3 made a significant contribution to the Fourth Assessment Report (AR4) of the Intergovernmental Panel on Climate Change (IPCC; Solomon et al. 2007) as the basis for investigating the response of climate to external forcings (Meehl et al. 2007). In the same manner, experiments conducted by phase 5 of CMIP (CMIP5; Taylor et al. 2012) contributed to the Fifth Assessment Report (AR5) of the IPCC (Stocker et al. 2013). A vast amount of model output produced by the CMIPs has been archived at the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and available for use by scientists. Hereafter, models that participated in CMIP3 and CMIP5 are referred to simply as CMIP3 models and CMIP5 models, respectively.

The performance of the atmosphere–ocean general circulation models (AOGCMs) that participated in CMIP5 with respect to the global-scale surface air temperature and precipitation distribution has apparently improved compared with the former generation of models in CMIP3, as indicated by Fig. 1 of the frequently asked questions (FAQ) in section 9.1 of Stocker et al. (2013). Sperber et al. (2013) demonstrated that the CMIP5 models are more skillful than the CMIP3 models

* Additional affiliation: Climate Research Department, Meteorological Research Institute, Tsukuba, Japan.

Corresponding author address: Shoji Kusunoki, Climate Research Department, Meteorological Research Institute, 1-1, Nagamine, Tsukuba, Ibaraki 305-0052, Japan.
E-mail: skusunoki@mri-jma.go.jp

DOI: 10.1175/JCLI-D-14-00585.1

in simulating various aspects of the Asian summer monsoon. The advantage of the CMIP5 models over the CMIP3 models was also reported by Ogata et al. (2014) for the Asian summer and winter monsoons. Watterson et al. (2014) evaluated the skill of the CMIP3 and CMIP5 models with regard to continental-scale seasonally averaged surface air temperature, precipitation, and mean sea level pressure. They introduced a combined skill measure from the three meteorological variables and found a modest improvement when comparing with the CMIP5 models with the CMIP3 models.

Song and Zhou (2014a) reported the improvement of simulated summer precipitation over East Asia by atmospheric general circulation models (AGCMs) incorporated within CMIP5 when compared with those of CMIP3. Song and Zhou (2014a) and He and Zhou (2014) indicated the erroneous northward shift of the western Pacific subtropical high (WPSH) in summer simulated by the AGCMs within CMIP3 and CMIP5 that distorts the precipitation climatology over East Asia. Song and Zhou (2014b) revealed that in CMIP5 AOGCMs are better than the AGCMs in simulating precipitation and low-level circulation over East Asia in summer, which indicates the importance of air–sea coupling in this region.

The target region of Sperber et al. (2013) and Ogata et al. (2014) covered the continental-scale area of Asia. However in this paper, we focus on model performance with respect to precipitation over the restricted area of East Asia, which is strongly characterized by a summer rainy season. Kusunoki and Arakawa (2012) investigated the ability of the CMIP3 models to simulate precipitation intensity in the East Asian rainy season and found an underestimation of intensity by these models. However, the ability to simulate extreme precipitation events including intense rainfall over East Asia has not yet been fully investigated for the CMIP5 models. The purpose of this paper is to compare the performance of CMIP5 models with that of the CMIP3 models over East Asia with respect to the geographical distribution of monthly precipitation, the seasonal march of the rainy season over China, the Korean Peninsula, and Japan, and extreme precipitation events.

In the remainder of this paper, section 2 describes the models we used in this study. Section 3 outlines the observational data used to verify model performance. Section 4 reports the performance of the models with regard to the geographical distribution of monthly mean precipitation. Section 5 considers model performance with respect to the seasonal march of the rainy season. Section 6 evaluates the ability of the models to reproduce extreme precipitation events. Section 7 discusses the reasons for the differences between the CMIP5 and CMIP3 models. We present conclusions in section 8.

2. Models

Table 1 shows information on 31 CMIP5 models used in this study. The majority of these models were AOGCMs. Several models were Earth system models (ESMs) that incorporate various biochemical cycles such as the carbon, sulfur, or ozone cycles. We only selected the models that archived daily precipitation data and used the Gregorian calendar. Eleven of the models used a realistic Gregorian calendar that included a leap year, but the other 20 models did not. The horizontal resolution of the CMIP5 models at 35°N ranged from 68 to 342 km, with an average of 193 km. None of these models incorporated the so-called "flux adjustment," which is a form of bias correction for momentum, heat, and salinity flux between atmosphere and ocean that is used to stabilize the climatology of the model. Large climate drifts can distort both the natural variability and the climate response to changes in radiative forcing, as indicated in section 8.2.7 of Solomon et al. (2007). Our target period was the 20-yr period from 1981 to 2000 in the "historical" experiment for the twentieth century. In the historical experiment of CMIP5, models are driven by the concentration of greenhouse gases, so that in the ESM models the concentration of carbon dioxide is prescribed and interactions associated with carbon flux are switched off.

Table 2 shows information on 15 CMIP3 models used in this study. These models were all AOGCMs. As with the CMIP5 models, we only selected those models that archived daily precipitation data and used the Gregorian calendar. Six models used a realistic Gregorian calendar with a leap year, but the other nine models did not. The horizontal resolution of the CMIP3 models ranged from 103 to 455 km, with an average of 254 km. Four models used flux adjustment, suggesting that these models contain large errors with respect to some of the fluxes between atmosphere and ocean. Our target period was the 20-yr period from 1981 to 2000 in the 20th Century Climate in Coupled Models (20C3M) experiments, but some models did not cover the whole target period. The models used in this study are identical to those used by Kusunoki and Arakawa (2012).

Most numerical model experiments included ensemble simulations, but we used only the first member from the multiple simulations.

3. Observational data

a. Precipitation

Table 3 summarizes the observational precipitation data used in this study. Model skill was calculated against the One-Degree Daily (1DD) data of the Global Precipitation Climatology Project (GPCP), version 1.1

TABLE 1. Features of 31 CMIP5 models used in this study. Target period is from 1981 to 2000 (20 yr). The average grid size is 193 km. (Expansions of model name acronyms are available online at http://www.ametsoc.org/PubsAcronymList.)

No.	Label	Model [Table 9.A.1 of Stocker et al. (2013)]	AGCM resolution*	No. of grid points in AGCM		Longitudinal grid spacing (km) at 35°N
				Longitude	Latitude	
1	a	ACCESS1.0	G63L17	192	145	171
2	b	ACCESS1.3	G63L17	192	145	171
3	c	BCC_CSM1.1	T42L17	128	64	256
4	d	BCC_CSM1.1(m)	T106L17	320	160	102
5	e	BNU-ESM	T42L17	128	64	256
6	f	CanESM2	T42L22	128	64	256
7	g	CCSM4	T95L17	288	192	114
8	h	CESM1(BGC)	T95L17	288	192	114
9	i	CESM1(CAM5)	T95L17	288	192	114
10	j	CMCC-CESM	T32L33	96	48	342
11	k	CMCC-CM	T160L17	480	240	68
12	l	CMCC-CMS	T63L33	192	96	171
13	m	CNRM-CM5	T85L17	256	128	128
14	n	CSIRO Mk3.6.0	T63L18	192	96	171
15	o	EC-EARTH	T106L16	320	160	102
16	p	FGOALS-g2	T42L17	128	60	256
17	q	GFDL CM3	G47L23	144	90	228
18	r	GFDL-ESM2G	T47L17	144	90	228
19	s	GFDL-ESM2M	T47L17	144	90	228
20	t	INM-CM4	G59L17	180	120	182
21	u	IPSL-CM5A-LR	T31L17	96	96	342
22	v	IPSL-CM5A-MR	T47L17	144	143	228
23	w	IPSL-CM5B-LR	T31L17	96	96	342
24	x	MIROC5	T85L17	256	128	128
25	y	MIROC-ESM	T42L35	128	64	256
26	z	MIROC-ESM-CHEM	T42L35	128	64	256
27	A	MPI-ESM-LR	T63L25	192	96	171
28	B	MPI-ESM-MR	T63L25	192	96	171
29	C	MRI-CGCM3	T106L23	320	160	102
30	D	MRI-ESM1	T106L23	320	160	102
31	E	NorESM1-M	T47L17	144	96	228

* The letter G denotes gridpoint model. Two digits after G show corresponding spectral wavenumber. The digits after T denote the triangular truncation at the corresponding spectral wavenumber. Two digits after the letter L denote the number of vertical levels.

(v1.1), compiled by Huffman et al. (2001) from 1997 to 2008 (12 yr). The horizontal resolution of the precipitation data was 1.0° in longitude and latitude, corresponding to a grid spacing of about 90 km over East Asia. We selected these GPCP 1DD data for verification because daily data are required to calculate extreme precipitation events, although GPCP 1DD data do not cover the whole target period of our simulations (1981–2000). Pentad and monthly data are also created from daily data. For the verification process, the model data were interpolated onto the location of the grid points within the GPCP 1DD.

Since model skill depends on the choice of observational data (Sperber et al. 2013), we used additional data to evaluate the uncertainty associated with the observation. Pentad data from GPCP v1.2 and monthly data from GPCP v2.2 compiled by Adler et al. (2003) were used for the period 1981–2000 (20 yr). These data cover the whole target period of our simulations. The horizontal resolution of the additional data was 2.5° in longitude and latitude, corresponding to a grid spacing of about 210 km over East Asia.

We used the pentad and monthly data of the Climate Prediction Center (CPC) Merged Analysis of Precipitation (CMAP), version 1201 (v1201), and monthly CMAP v1201 data compiled by Xie and Arkin (1997) for the study period 1981–2000. The horizontal resolution this data was 2.5° in longitude and latitude.

Pentad data from the Tropical Rainfall Measuring Mission (TRMM) 3B42 product and monthly data from TRMM 3B43 product compiled by Huffman et al. (2007) were used for the period 1998–2010 (13 yr). The horizontal resolution of this dataset was 0.25° in longitude and latitude, corresponding to a grid spacing of about 25 km over East Asia. Regional coverage was restricted to a global belt extending from 50°S to 50°N.

TABLE 2. Features of 15 CMIP3 models used in this study. Target period is from 1981 to 2000 (20 yr). The average grid size is 254 km.

No.	Label	Model [Table 8.1 of Solomon et al. (2007)]	AGCM resolution[a]	No. of grid points in AGCM		Longitudinal grid spacing (km) at 35°N	Flux adjustment
				Longitude	Latitude		
1	a	BCCR-BCM2.0[b]	T42L31	128	64	256	No
2	b	CCSM3[c]	T85L26	256	128	128	No
3	c	CGCM3.1(T47)[d]	T30L31	96	48	392	Yes
4	d	CGCM3.1(T63)[e]	T42L31	128	64	256	Yes
5	e	CNRM-CM3	T42L45	128	64	256	No
6	f	CSIRO MK3.0	T63L18	192	96	171	No
7	g	ECHAM5/MPI-OM	T63L31	192	96	171	No
8	h	FGOALS-g1.0	T42L26	128	60	256	No
9	i	GFDL CM2.0	G47L24	144	90	228	No
10	j	GISS-AOM	G29L12	90	60	364	No
11	k	INM-CM3.0	G23L21	72	45	455	Yes
12	l	MIROC3.2(hires)	T106L56	320	160	103	No
13	m	MIROC3.2(medres)	T42L20	128	64	256	No
14	n	MRI-CGCM2.3.2	T42L30	128	64	256	Yes
15	o	PCM[f]	T42L26	128	64	256	No

[a] The letter G denotes gridpoint model. Two digits after G show corresponding spectral wavenumber. The digits after T denote the triangular truncation at the corresponding spectral wavenumber. Two digits after the letter L show vertical levels.
[b] Available data terminate in year 1998.
[c] Available data terminate in year 1999.
[d] Model name suggests T47, but the Program for Climate Model Diagnosis and Intercomparison (PCMDI) archives data had T30 resolution.
[e] Model name suggests T63, but PCMDI archives data had T42 resolution.
[f] Available data terminate in year 1999.

We also used the monthly data from the Asian Precipitation Highly Resolved Observational Data Integration Toward the Evaluation of Water Resources (APHRODITE) V1003R1 dataset compiled by Yatagai et al. (2009) for the period 1981–2000 (20 yr). These data only cover land area and have a horizontal resolution of 0.25° in longitude and latitude.

b. *Reanalysis data*

To verify the simulated large-scale circulation, we used the Japanese 55-year Reanalysis Project (JRA-55; Ebita et al. 2011) from 1981 to 2000 (20 yr). The horizontal resolution is 1.25° in longitude and latitude, corresponding to a grid spacing of about 105 km over East Asia. For the verification process, the model data were interpolated onto the location of the grid points used in the JRA-55 dataset.

4. Monthly precipitation

The rainy season over East Asia typically starts in May and ends in July, although this varies slightly with depending on the location. June is the middle of the rainy season, and Fig. 1 compares the precipitation climatology simulated by the CMIP5 models with the observations for June. In the observations (Fig. 1, top panels), the region of high precipitation amounts extends over the southern part of China, the East China

TABLE 3. Observational precipitation data used in this study.

Name	Temporal resolution	Spatial resolution	Period	Region	Reference
GPCP 1DD v1.1	Day	1.0°	1997–2008 (12 yr)	Global	Huffman et al. (2001)
GPCP 1DD v1.1	Pentad	1.0°	1997–2008 (12 yr)	Global	Huffman et al. (2001)
GPCP v1.2	Pentad	2.5°	1981–2000 (20 yr)	Global	Adler et al. (2003)
CMAP v1201	Pentad	2.5°	1981–2000 (20 yr)	Global	Xie and Arkin (1997)
TRMM 3B42	Pentad	0.25°	1998–2010 (13 yr)	50°S–50°N	Huffman et al. (2007)
GPCP 1DD v1.1	Month	1.0°	1997–2008 (12 yr)	Global	Huffman et al. (2001)
GPCP v2.2	Month	2.5°	1981–2000 (20 yr)	Global	Adler et al. (2003)
CMAP v1201	Month	2.5°	1981–2000 (20 yr)	Global	Xie and Arkin (1997)
TRMM 3B43	Month	0.25°	1998–2010 (13 yr)	50°S–50°N	Huffman et al. (2007)
APHRODITE V1003R1	Month	0.25°	1981–2000 (20 yr)	Land only	Yatagai et al. (2009)

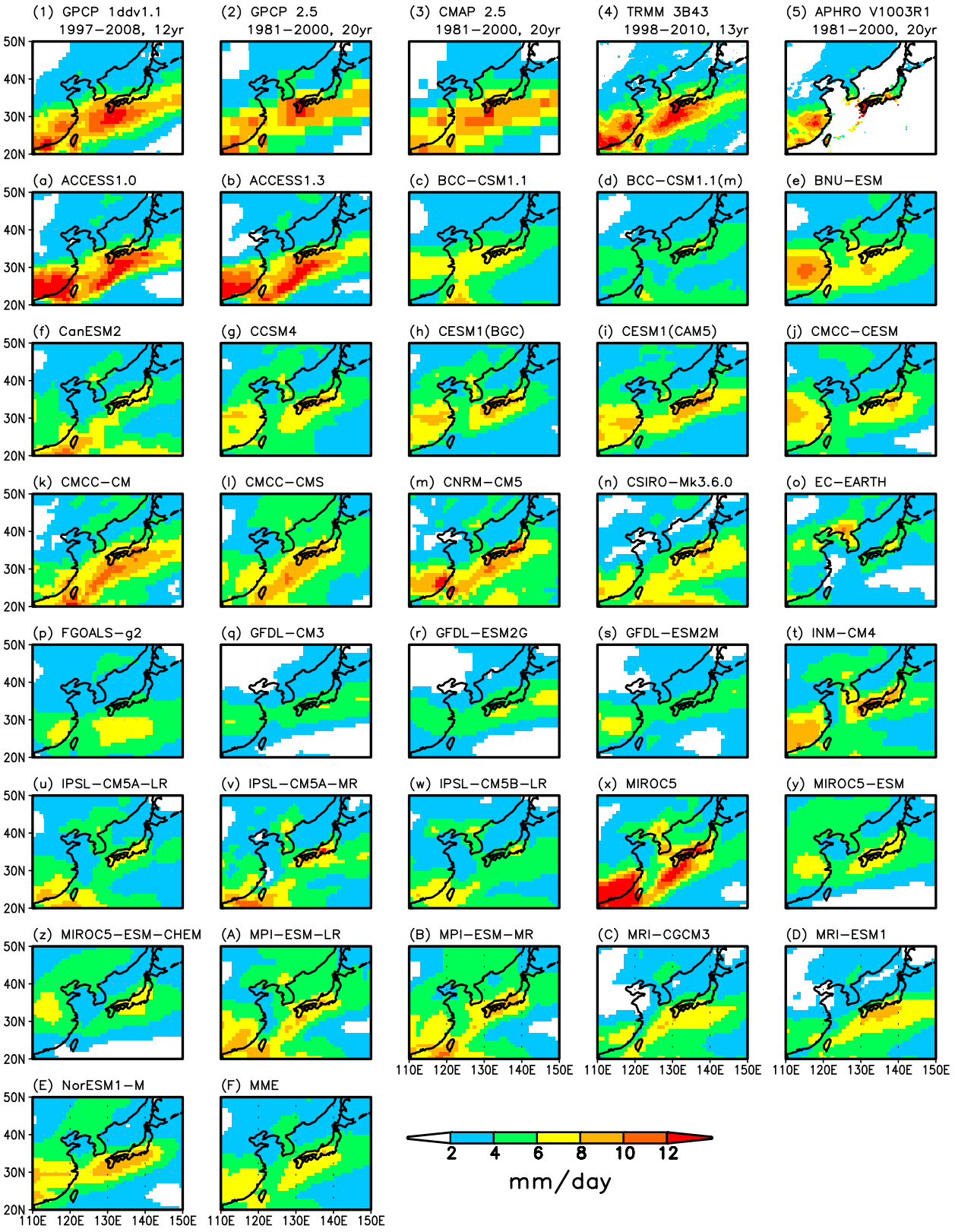

FIG. 1. Precipitation (mm day^{-1}) in June: (top) Observations (Table 3), (a–z,A–E) CMIP5 models (Table 1) for 1981–2000, and (F) MME of the CMIP5 models. The model data were interpolated onto the location of the grid points within the GPCP 1DD.

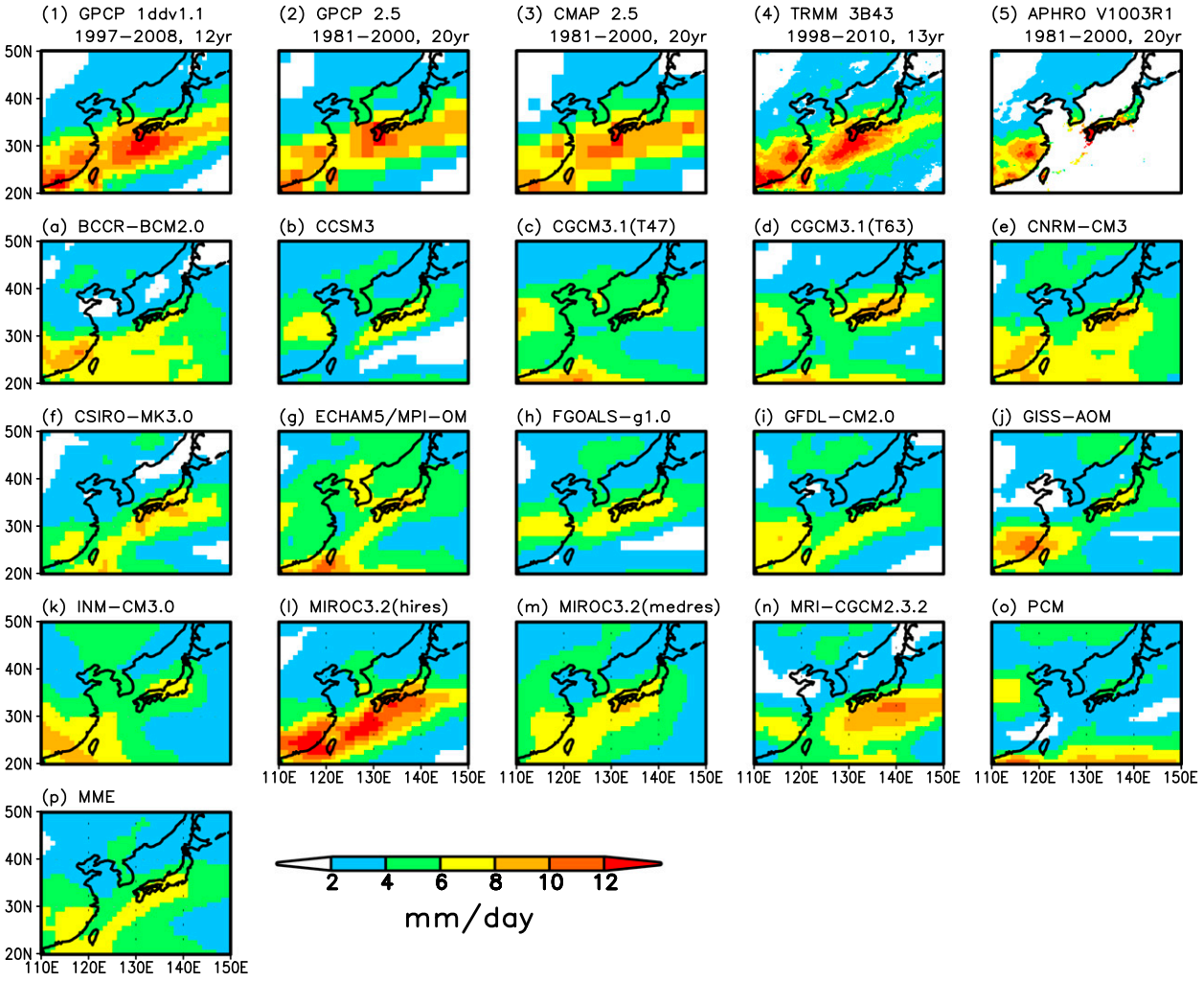

FIG. 2. As in Fig. 1, but for CMIP3 models (Table 2). (p) MME of the CMIP3 models.

Sea, the western part of Japan, and to the south of Japan. The maximum values of precipitation over the southern part of China shown by GPCP 1DD (top panel, first column), TRMM (top panel, fourth column), and APHRODITE (top panel, fifth column) are larger than those by GPCP 2.5° data (top panel, second column) and CMAP 2.5° data (top panel, third column), but this is mainly because of the finer grid size. In general, models (Figs. 1a–z,A–E) tend to underestimate this rainy zone over East Asia, but some models (Figs. 1a,b,x) show excessive precipitation. Underestimation of precipitation is also evident in the average multimodel ensemble (MME) of all 31 CMIP5 models (Fig. 1F).

Figure 2 compares the precipitation climatology simulated by the CMIP3 models with the observational data for June. As with the CMIP5 models, individual models (Figs. 2a–o) and the MME (Fig. 2p) tend to underestimate precipitation. Only the MIROC3.2(hires) model overestimates precipitation (Fig. 2l).

To quantitatively compare the performance of the CMIP5 models with that of the CMIP3 models, we calculated the skill score S proposed by Taylor (2001) against the GPCP 1DD data (Figs. 1 and 2, top panel, first column); S is defined by

$$S = \frac{4(1+R)}{\left(\sigma + \frac{1}{\sigma}\right)^2 (1+R_0)},$$

where R is the spatial correlation coefficient between observation and simulation, σ is the spatial standard deviation of the simulation divided by that of the observations, and R_0 is the maximum correlation attainable. Here we assumed that $R_0 = 1$. Note that S evaluates the spatial correlation coefficient as well as the spatial standard deviation. The value of S approaches unity in a perfect simulation. Although S is widely used to verify model performance in many climate modeling

studies, we must recognize that S cannot evaluate model bias because bias is corrected and excluded before the calculation of S. In the calculation, model data were interpolated onto the observation grid of 1° in longitude and latitude. The target region (20°–50°N, 110°–150°E) in Figs. 1 and 2 includes 40 grid points in longitude and 30 grid points in latitude, resulting in a total of $40 \times 30 = 1200$ grid points. Thus, the number samples for each two-dimensional field was 1200. Latitudinal weight was considered in the area averaging.

Figure 3 illustrates the Taylor diagram (Taylor 2001) of S for the CMIP5 and CMIP3 models as well as the bias and root-mean-square error (RMSE). The target month is June and the target region is the same as in Figs. 1 and 2. In the Taylor diagram (Fig. 3b), the radial distance from the origin is proportional to the standard deviation of a simulated pattern normalized by the observed standard deviation. The spatial correlation coefficient between the observed and simulated fields is given by the angle from the y axis. The skill of the GPCP 2.5° data and CMAP 2.5° data against the GPCP 1DD data were also plotted to estimate the uncertainty associated with the observations. We calculated the average skill of the individual models (AVM) as well as the skill of the MME. In the case of linear skill measures such as average and bias, the MME is identical to the AVM. In the case of nonlinear skill measure such as RMSE, the correlation coefficient and Taylor skill score S, the MME and AVM differ.

Figure 3a shows that most models from CMIP5 and CMIP3 have a negative bias, whereas ACCESS1.0 (red letter a) and MIROC5 (red letter x) from CMIP5 and MIROC3.2(hires) (blue letter l) from CMIP3 have a positive bias. The bias in the CMIP3 MME (blue circle) is slightly smaller than that for the CMIP5 MME (red circle). Conversely, the RMSE of the CMIP5 MME is slightly smaller than that of the CMIP3 MME. In both the CMIP5 and CMIP3 models, the RMSE of the MME (red and blue circles) is smaller than the AVM (red and blue squares), and this is caused by the cancellation of model errors by averaging. The MME average can be expected to outperform individual models in climate simulations (Lambert and Boer 2001; Gleckler et al. 2008; Reichler and Kim 2008) as well as seasonal forecasts (Kusunoki et al. 2001; Palmer et al. 2004; Hagedorn et al. 2005).

In the Taylor diagram (Fig. 3b), the majority of points fall inside the quadrant of radius one, which means that the standard deviation of most models is smaller than the observations. This implies that the simulated geographical distributions of precipitation are smoother than the observations. In contrast the ACCESS1.0 (red letter a), ACCESS1.3 (red letter b), and MIROC5 (red

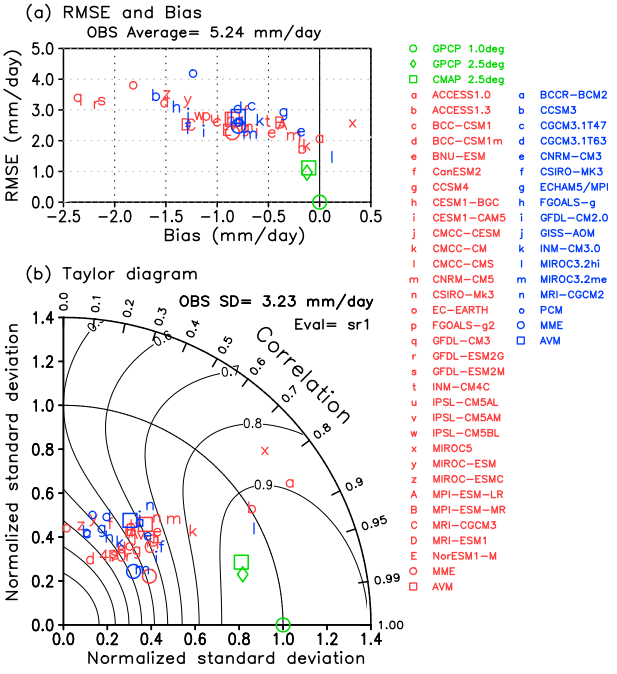

FIG. 3. Skill of precipitation climatology for June simulated by models verified against the GPCP 1DD v1.1 data (green circle). Green, red, and blue symbols denote observations, CMIP5 models, and CMIP3 models, respectively. The target domain is same as in Figs. 1 and 2. (a) RMSE vs bias (mm day^{-1}). The domain average of the observation is shown above the panel. (b) Taylor diagram displaying pattern statistics (Taylor 2001). The standard deviation of the observation in the domain is shown above the panel. The contours show the value of Taylor skill score S. Red and blue circles indicate the MME. Red and blue squares indicate the average skill of the individual models (AVM).

letter x) models from CMIP5, and the MIROC3.2(hires) (blue letter l) model from CMIP3 show spatial variability that is greater than, or equivalent to, the observations. The spatial correlation coefficients R of these models tend to be larger than those of other models. In terms of the MME and AVM, the Taylor skill score S (contour plot) of the CMIP5 models (red circle and square) is larger than the CMIP3 models (blue circle and square). The advantage of the MME over AVM is recognized in the spatial correlations coefficients R for both the CMIP5 and CMIP3 models, but not for the Taylor skill score. The differences in RMSE, bias, and S among the three observations (green symbols) are smaller than those of this models. This provides positive evidence that verification against GPCP 1DD data is reasonable and appropriate.

Figure 4 summarizes and visualizes both the bias and Taylor skill score S of the CMIP5 models for all months. The magnitude and sign of the bias are indicated using a color scale, and the value of S is expressed by the size of the circle. During the warmer season, from May to

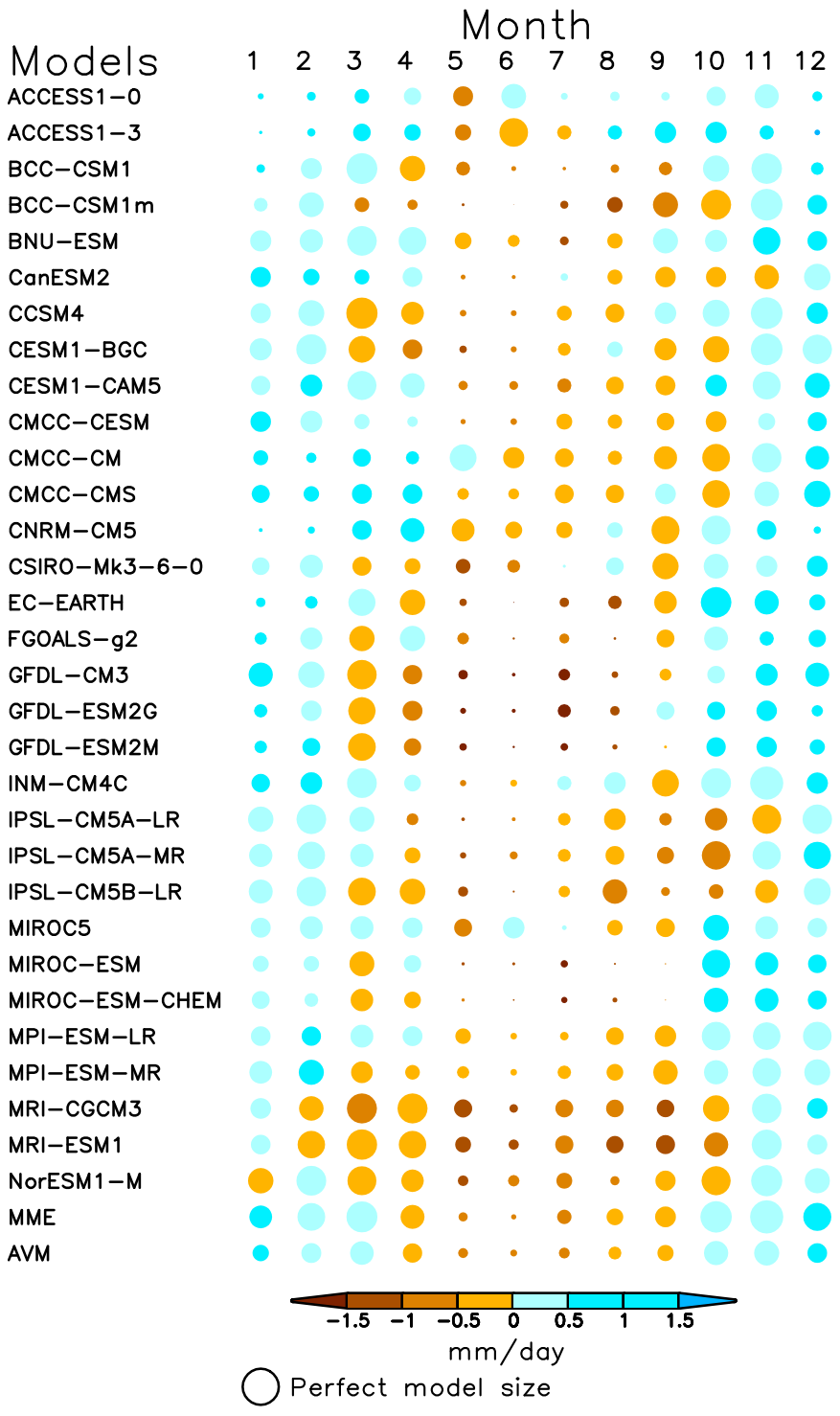

FIG. 4. Bias (color) and S (size of circle) of CMIP5 models for all months. Diameter of a circle is proportional to the third power of S. Size of circle is much more sensitive to third power of S than to the linear case.

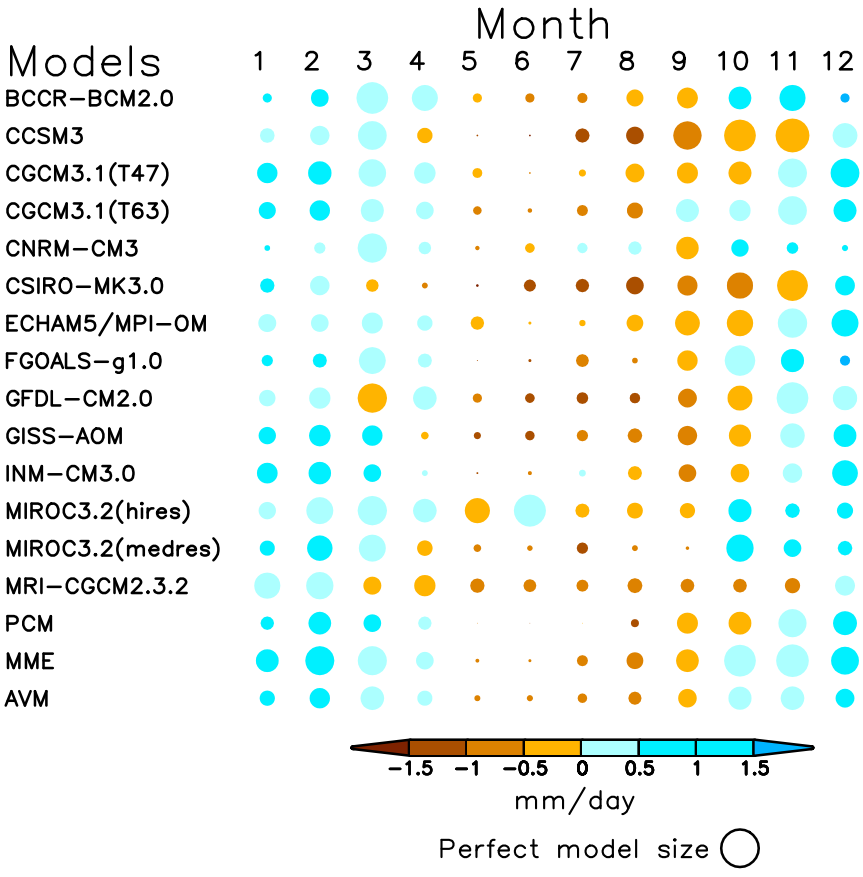

FIG. 5. As in Fig. 4, but for CMIP3 models.

September, the models tend to underestimate precipitation (brown shades). Moreover, the Taylor skill scores of the models are low (small circles). During the colder season, from November to February, the models tend to overestimate precipitation (blue shades). The Taylor skill scores of the models are relatively higher in the colder season (large circles) than the warmer season. This seasonal dependence of an individual model's skill is also evident in the MME and AVM. The Taylor skill score (size of circle) of the MME is higher than the AVM, except for May and June, but the advantage of the MME over the AVM is not clear for the bias (color).

Figure 5 is the same as Fig. 4, but for CMIP3 models. As with the CMIP5 models, the CMIP3 models underestimate precipitation and show a reduced Taylor skill score in the warmer season, but overestimate precipitation and have a larger Taylor skill score in the colder season. Precipitation during the cold season over East Asia is largely influenced by the activity of storm tracks, and the higher skill in the cold seasons might be partly attributed to the higher reproducibility of storm tracks for that seasons.

It is difficult to judge whether the CMIP5 models are better or worse than the CMIP3 models simply by comparing Figs. 4 and 5. Therefore, the MME and AVM are directly compared in Fig. 6 by means of the RMSE. The advantage of the CMIP5 models is evident in all months for the MME (red) and AVM (green). We further confirmed the statistically significance of these differences using the bootstrap method (see the appendix). We found that the statistically significant advantage of the CMIP5 models is limited to the warmer season. Errors in warmer months such as May, June, and July are larger than those in the colder months. The underestimation of precipitation and lower Taylor skill score contribute to a larger RMSE in the warmer months. As for each month, the RMSE value for the MME (red) tends to be smaller than those of AVM (green) as already indicated by Fig. 3a for June case. A similar plot for the Taylor skill score (figure not shown) gave the same results, but with less robustness compared with Fig. 6.

The MIROC5 model from CMIP5 shows a considerably larger positive bias in June precipitation compared with other CMIP5 models (Fig. 1x; red letter x in Fig. 3a; blue circle for June in Fig. 4). This large bias might lower

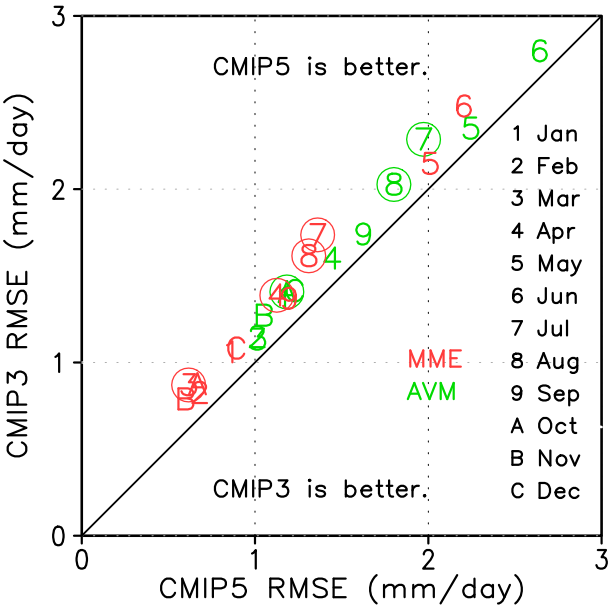

FIG. 6. Comparison of RMSE (mm day^{-1}) between the CMIP5 models and the CMIP3 models for all months. The MME and AVM are indicated in red and green, respectively. Circles indicate a significance level exceeding 90% based on the bootstrap method (see the appendix).

the skill of the MME and AVM. To evaluate the contribution of MIROC5 to the MME and AVM, we calculated the skill of the MME and AVM excluding MIROC5. In the case of the MME for June, the RMSE without MIROC5 was reduced by about 5% relative to the RMSE with MIROC5. In contrast, in the case of the AVM, the RMSE calculated without MIROC5 increased by about 20% relative to that calculated with MIROC5. MIROC5 shows a very high spatial correlation coefficient and *S* value (red letter x in Fig. 3b; large blue circle for June in Fig. 4). Therefore, excluding MIROC5 leads to a lower skill in terms of the AVM. A similar effect is also seen for MIROC3.2(hires) from CMIP3. As there are fewer CMIP3 models (15) than that of CMIP5 models (31), the relative contribution of MIROC3.2(hires) is larger than that of MIROC5. Therefore, the effect of excluding MIROC3.2(hires) on the skill of the MME and AVM is larger than that of MIROC5.

5. Seasonal march of the rainy zone

The rainy zone over East Asia moves northward during the rainy season from May to July (Wang and LinHo 2002). This rainy season and its associated rainbands are called the mei-yu in China, changma over the Korean Peninsula, and baiu in Japan. The top panel in Fig. 7 defines the three regions used to calculate the longitudinal averages for precipitation. Figure 8 compares the observed seasonal march of the rainy zones with the CMIP5 and CMIP3 MMEs. For China (mei-yu), the observations (Figs. 7a,d,g) show a stationary rainy zone around 25°N throughout the period from April to August. Part of this rainy zone begins to migrate north in June and reaches around 40°N in July. The precipitation maximum of this migrating section in the GPCP 1DD dataset (Fig. 7a) is larger than that in the GPCP 2.5° (Fig. 7d) and CMAP 2.5° (Fig. 7g) data, but this is mainly because of the finer grid size. The CMIP5 MME (Fig. 7j) partly captures the observed pattern of the seasonal march, but the simulated precipitation is generally underestimated across the whole period. In particular, the part of northward migration seen in the model is much weaker than in the observation. The CMIP3 MME (Fig. 7m) shows a similar defect to the CMIP5 MME (Fig. 7j), but with a larger underestimation of precipitation and a weaker northward branch of the rainy zone.

For the Korean Peninsula and the southern island of Japan (Okinawa), the observations (Figs. 7b,e,h) show that the rainy zone begins to migrate northward in May and reaches around 45°N in July. The stationary rainy zone around 25°N seen over China does not develop over the Korean Peninsula. The differences among the observations (Figs. 7b,e,h) are relatively small. The CMIP5 MME (Fig. 7k) shows smaller precipitation amounts and a weaker northward migration of the rainy zone compared with the observations (Figs. 7b,e,h). The CMIP3 MME (Fig. 7n) shows a similar defect to the CMIP5 MME (Fig. 7k), but with a larger underestimation of precipitation and a weaker northward branch of the rainy zone. The degree of underestimation of precipitation over the Korean Peninsula for the CMIP5 (Fig. 8k) and CMIP3 MMEs (Fig. 7n) is much larger than for China (Figs. 7j,m).

For Japan (baiu), the observed seasonal march of the rainy zone (Figs. 7c,f,i) is almost the same as for the Korean Peninsula (Figs. 8b,e,h). The CMIP5 MME (Fig. 8l) shows smaller precipitation amounts and a weaker northward migration of the rainy zone compared with the observations (Figs. 7e,f,i). The CMIP3 MME (Fig. 7o) shows a similar defect to the CMIP5 MME (Fig. 7l), but with a larger underestimation of precipitation and a weaker northward branch of the rainy zone. The simulated precipitation maximum in June around 35°N over Japan (Figs. 7l,o) is larger than that over the Korean Peninsula (Figs. 7k,n), but the simulated northward migration of the rainy zone over Japan (Figs. 7l,o) is slower and shorter than both the observations (Figs. 7e,f,i) and those simulated over the Korean Peninsula (Figs. 7k,n).

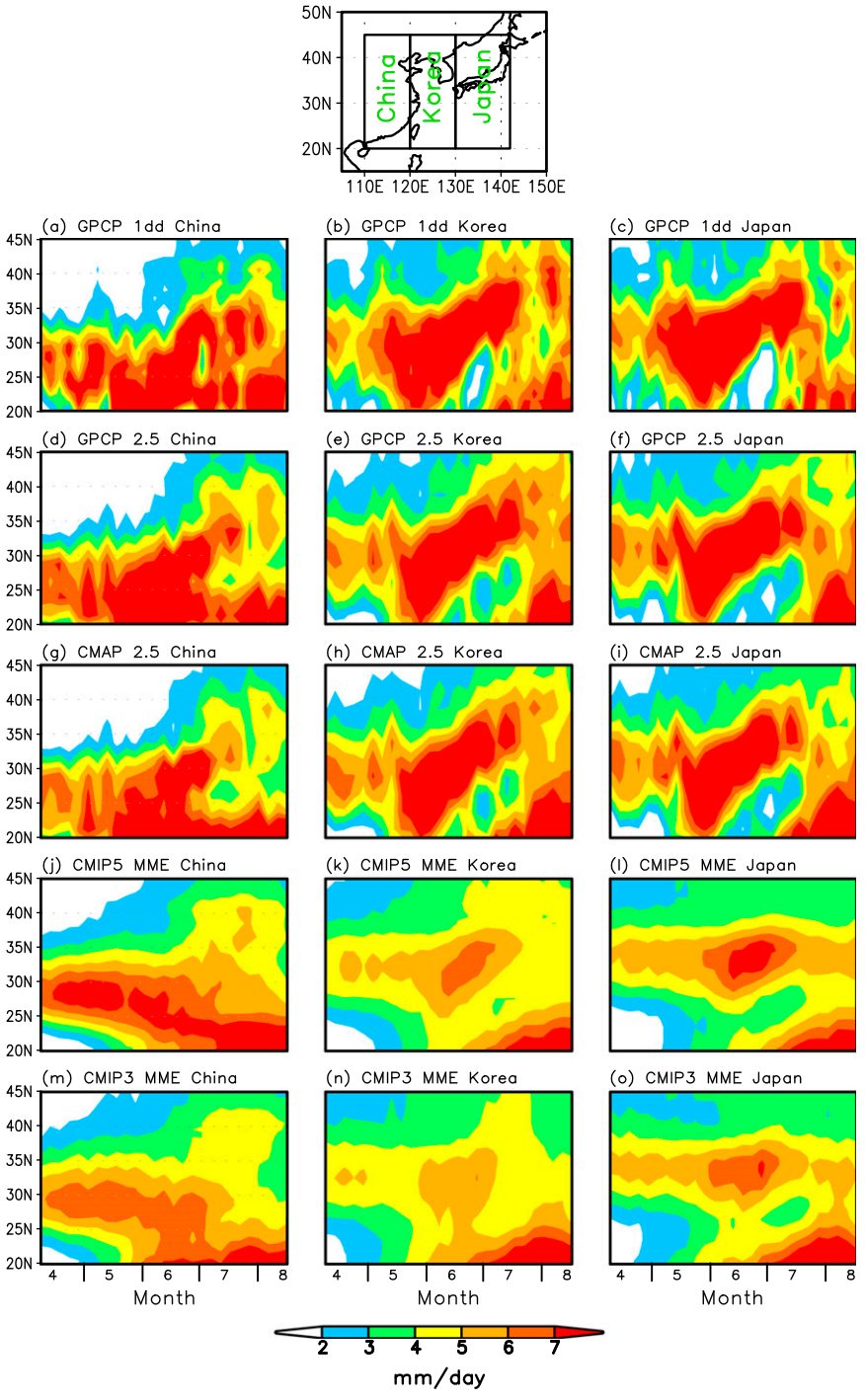

FIG. 7. Time–latitude cross sections of climatological pentad mean precipitation (mm day^{-1}) averaged for China (20°–45°N, 110°–120°E), the Korean Peninsula (20°–45°N, 120°–130°E), and Japan (20°–45°N, 130°–142°E). (top) Map of the target regions. The target period is from pentad 20 (6–10 April) to 46 (14–18 August). (a) Observed GPCP 1DD data for the Chinese region. (b),(c) As in (a), but for the Korean Peninsula region and Japanese region, respectively. (d)–(f) As in (a)–(c), but for GPCP 2.5° data. (g)–(i) As in (a)–(c), but for CMAP 2.5° data. (j)–(l) As in (a)–(c), but for CMIP5 MME. (m)–(o) As in (a)–(c), but for CMIP3 MME.

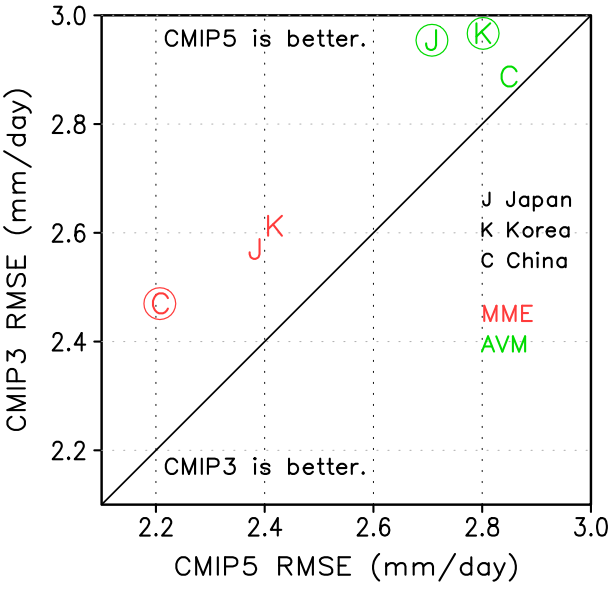

FIG. 8. Comparison of RMSE (mm day^{-1}) between CMIP5 models and CMIP3 models for time–latitude cross sections of precipitation (Fig. 7). RMSEs were calculated against GPCP 1DD data. The MME and AVM are indicated in red and green, respectively. Circles indicate a significance level exceeding 90% based on the bootstrap method (see the appendix).

Figure 8 compares the MME and AVM of the CMIP5 and CMIP3 models using the RMSE. The underestimation of precipitation mainly leads to increase RMSE. The advantage of the CMIP5 models is evident in all three regions for the MME and AVM. The errors in the MME and AVE over the Korean Peninsula for both the CMIP5 and CMP3 models are the largest among the three regions, suggesting some difficulty in simulating the rainy season of the Korean Peninsula. As for each region, the errors in the MME (red) tend to be smaller than those of the AVM (green), which is similar to Fig. 6. The statistically significant advantage of the CMIP5 models is evident for the MME over China and the AVM over Japan and the Korean Peninsula. A similar plot of the Taylor skill scores (figure not shown) shows almost the same results, giving robustness to our analyses.

Despite of the improvement in the CMIP5 models revealed in this section, we should recognize that the models still fail to capture the northward migration of the rainy zone over East Asia. Previous studies using a very high horizontal resolution AGCM with a 20-km grid size (Kusunoki et al. 2006; Kitoh and Kusunoki 2008) stressed that the realistic reproduction of the seasonal march of the rainy zone requires a model with a higher horizontal resolution. One possible reason for is that a model with a higher horizontal resolution is more likely to resolve small-scale structures embedded in the mei-yu, changma, and baiu rainbands. Another possible reason is that the higher horizontal resolution model can realistically represent the high elevation of the Tibetan Plateau. Kitoh (2004) conducted idealized model experiments in which the elevation of the Tibetan Plateau was systematically changed from 0% to 100% of actual values. Summer rainy season over East Asia was not simulated in the lower elevation scenario, but the rainfall reproducibility of precipitation amounts and the seasonal march of the rainy season improved as elevation was increased. These experiments highlight the importance of the orographic effect of the Tibetan Plateau on the formation of the summer rainy season over East Asia. Models with a higher horizontal resolution model more realistically represent the elevation of the Tibetan Plateau than do the lower-resolution models. This may lead to the better reproduction of the northward migration of the rainy season.

6. Extreme precipitation events

The ability of models to simulate extreme precipitation events was investigated for the four indices SDII, R5d, PMAX, and CDD defined in Table 4. The indices SDII, R5d, and PMAX measure precipitation intensity, whereas CDD is a measure of dryness and drought. These indices are based on annual statistics. Figure 9 compares observations from the GPCP 1DD data and simulations by the CMIP5 and CMIP3 MMEs. The observed precipitation intensity indicated by SDII (Fig. 9a) shows that the region of intense rainfall extends broadly over East Asia to the south of 40°N with maximum values over southern China, the Korean Peninsula, and Japan. The CMIP5 MME (Fig. 9e) underestimates most

TABLE 4. Indices of extreme precipitation events.

Index	Expanded index name (unit)	Definition
SDII	Simple daily precipitation intensity index (mm day^{-1})	Total annual precipitation divided by the number of rainy days (precipitation ≥1 mm)
R5d	Maximum 5-day precipitation total (mm)	Maximum of consecutive 5-day precipitation
PMAX	Maximum 1-day precipitation (mm)	Maximum of precipitation in one day
CDD	Consecutive dry days (day)	Maximum number of consecutive dry days (precipitation <1 mm)

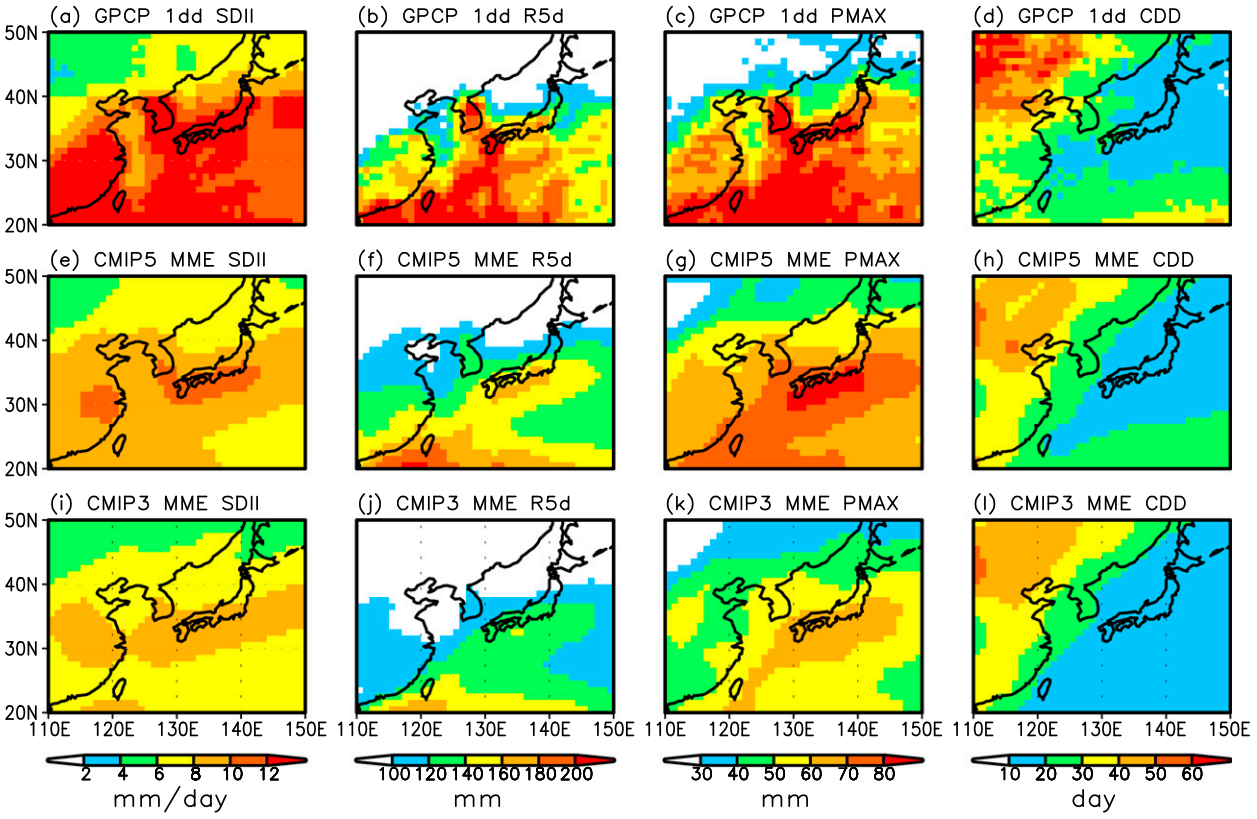

FIG. 9. Extreme precipitation events (Table 4). (a) SDII (mm day^{-1}) observations from GPCP 1DD data. (b)–(d) As in (a), but for R5d (mm), PMAX (mm), and CDD (day), respectively. (e)–(h) As in (a)–(d), but for CMIP5 MME. (i)–(l) As in (a)–(d) but for CMIP3 MME.

of the intense precipitation regions over East Asia, but the maximums over southern China, the Korean Peninsula, and Japan are reproduced to some degree. The CMIP3 MME (Fig. 9i) is similar to the CMIP5 MME (Fig. 9e), but the intensity simulated by the CMIP3 MME is smaller than that simulated by the CMIP5 MME. For the R5d precipitation index, the observations (Fig. 9b) also show the region of intense rainfall to the south of 40°N with maximums over the Korean Peninsula and over the sea to the south of 25°N. The CMIP5 MME (Fig. 9f) underestimates most of the intense precipitation regions over East Asia, but the maximum over Taiwan is reproduced to some extent precipitation intensity simulated by the CMIP3 MME (Fig. 9j) is smaller than seen in the CMIP5 MME (Fig. 9f). For the third precipitation index, PMAX, the CMIP5 MME (Fig. 9g) reproduces well the observed distribution of intense precipitation but with some underestimation. Similar to the SDII and R5d indices, the precipitation intensity simulated by the CMIP3 MME (Fig. 9k) is smaller than simulated by the CMIP5 MME (Fig. 9g).

The observed CDD (Fig. 9d) indicates that a drier climate prevails over northern China, which reflects the relatively lower precipitation intensity observed in this area (Figs. 9a–c). The CMIP5 MME (Fig. 9h) and CMIP3 MME (Fig. 9l) models capture well the observed distribution of CCD over East Asia, although the simulated CCD is underestimated in the northern part of China.

Figure 10 compares the MME and AVM of the CMIP5 and CMIP3 models using the RMSE. As units and magnitudes differ among the four indices, the RMSEs were normalized using the average of the observations over the whole domain (20°–45°N, 110°–150°E) with respect to each index. For example, the average of observed SDII over the whole domain was 10.6 mm day^{-1} and the RMSE of the CMIP5 MME was 3.3 mm day^{-1}. Therefore, the normalized RMSE becomes (3.3/10.6) × 100 = 30.7%. Figure 10 clearly illustrates the advantage of the CMIP5 models over the CMIP3 models for all four indices and also for the MME and AVM. These advantages are statistically significant except for the AVM of PMAX. As in Figs. 6 and 8, the errors associated with the MME are smaller than those of the AVM, suggesting the advantage of the MME over the AVM.

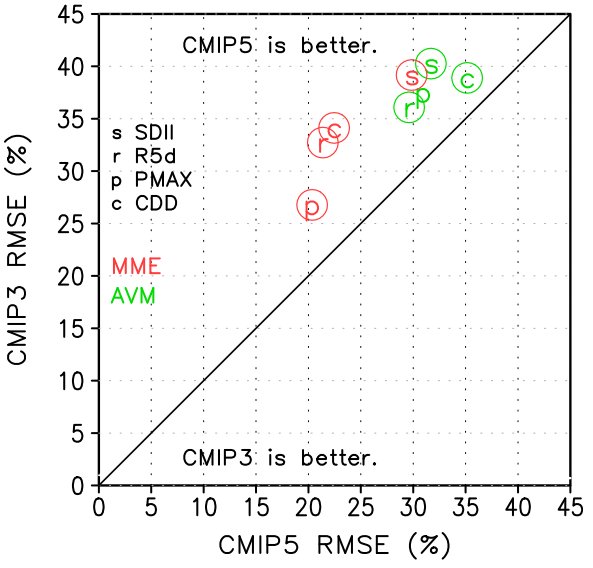

FIG. 10. Comparison of RSME (%) between the CMIP5 models and CMIP3 models for extreme precipitation events (Figs. 9e–l). The RMSE were calculated against the GPCP 1DD data (Figs. 9a–d) and normalized against the observed average over the whole domain with respect to each index. The MME and AVM are indicated in red and green, respectively. Circles indicate a significance level exceeding 90% based on the bootstrap method (see the appendix).

Kusunoki and Arakawa (2012) indicated that those CMIP3 models that show higher reproducibility of the precipitation climatology also tend to show higher reproducibility of precipitation intensity (SDII) over East Asia for the rainy season from June to July. We examined whether this relationship also holds for the annual statistics of precipitation intensity in the CMIP5 and CMIP3 models. Table 5 summarizes the correlation coefficients between the skill of simulating the annual precipitation and skill of precipitation intensity as a function of precipitation intensity and skill measure. In the case of the CMIP5 models, the correlation coefficients are all positive. In particular, the highest correlation coefficient of 0.598 is seen for the relationship between R5d and *R*. Scatter diagram in this particular case is depicted in Fig. 11. Most points fall below the diagonal line. This suggests that simulating precipitation intensity is much more difficult than simulating annual precipitation. Similarly, the correlation coefficients are also all positive for the CMIP3 model, but the level of statistical significance is lower than for the CMIP5 models. This can be partly attributed to the smaller sample size (15). In summary, our analysis provides robust evidences for relationship that models with higher reproducibility of precipitation climatology tend to show higher reproducibility of precipitation intensity.

TABLE 5. Correlation coefficient between skill of annual precipitation and skill of precipitation intensity.

Precipitation intensity index	Skill measure: Spatial correlation coefficient *R*	Skill measure: RMSE	Skill measure: Taylor skill score *S*
CMIP5 models (sample size 31)			
SDII	0.386*	0.024	0.235
R5d	0.598**	0.482**	0.401*
PMAX	0.577**	0.435*	0.279
CMIP3 models (sample size 15)			
SDII	0.329	0.276	0.277
R5d	0.556*	0.421	0.526*
PMAX	0.687**	0.378	0.517*

* Above the 95% statistical significance level.
** Above the 99% statistical significance level.

7. Why are CMIP5 models better than CMIP3 models?

a. *Horizontal resolution*

Many factors have contributed to the improvement of climate model. Especially, the higher horizontal resolution of the atmosphere enhances model performance for numerous phenomena including extreme precipitation, diurnal variations of precipitation, tropical cyclone intensity, and so on. [For details, see Box 9.3 in Stocker et al. (2013).] For example, Fig. 12 demonstrates the dependence of model skill on the grid spacing for all 46 models from CMIP5 and CMIP3. The skill measure is the spatial correlation coefficient over the East Asia

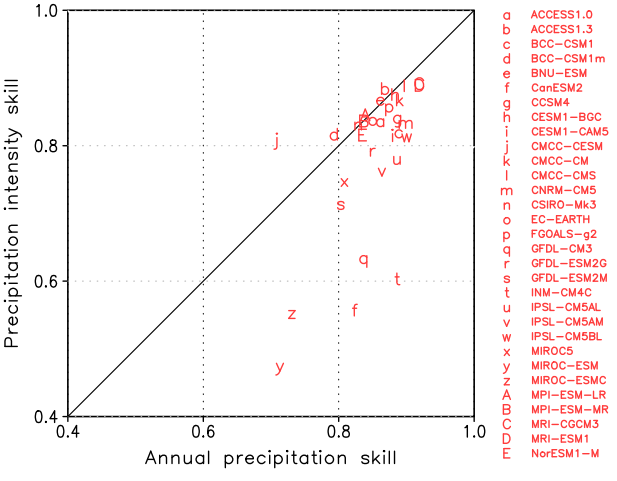

FIG. 11. Relationship between reproducibility of annual precipitation and precipitation intensity (R5d) for the CMIP5 models. The skill measure is the spatial correlation coefficient between the GPCP 1DD observation and simulations over the East Asia region (20°–50°N, 110°–150°E). The correlation coefficient between annual precipitation skill and R5d skill is +0.598, which exceeds the 99% significance level based on Student's *t* test.

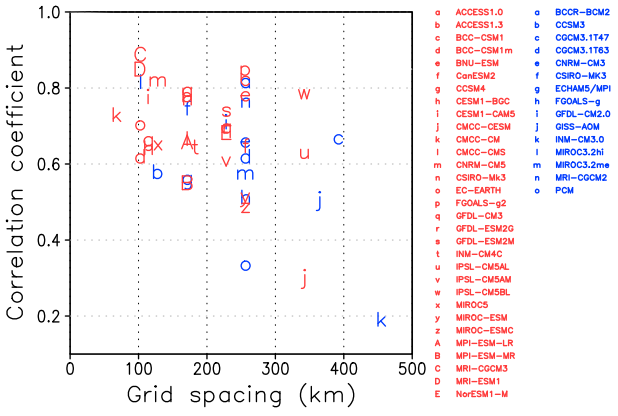

FIG. 12. Dependence of model skill on the grid spacing of the CMIP5 (red) and CMIP3 (blue) models. The skill measure is the spatial correlation coefficient R between the GPCP 1DD observation and simulations over the East Asia region (20°–50°N, 110°–150°E) for spring (April–May) precipitation. The correlation coefficient between grid spacing and $R = -0.485$, which exceeds the 99% significance level based on Student's t test.

region for precipitation in spring. The advantage of models with a higher horizontal resolution is evident from the statistically significant relationship between grid spacing and model skill.

Kusunoki and Arakawa (2012) suggested that those CMIP3 models with a higher horizontal resolution tend to have relatively higher skill for precipitation distribution over East Asia in the rainy season from June to July, but the relationship is not statistically significant and they only used the SDII. We extended their investigation to other extreme indices, as well as to other seasons and the seasonal march of the rainy season for the CMIP3 and CMIP5 models. Table 6 summarizes the dependence of model skill on the grid size of the model. A smaller grid size corresponds to a higher horizontal resolution. Therefore, if higher horizontal resolution models tend to have higher skill, the correlation coefficient between R and grid size becomes negative. The situation is the same for the Taylor skill score S. Conversely, in the case of RMSE, the correlation coefficient becomes positive.

All of the correlation coefficients associated with R are negative, and the correlation coefficients associated with S are also negative, except for winter. In contrast, the correlation coefficients associated with the RMSE are positive, except for winter. The signs of the correlation coefficients for R, RMSE, and S suggest the advantage of a higher horizontal resolution. Moreover, most of the correlation coefficients are statistically significant. As for extreme events, the advantage of a high horizontal resolution is more evident than for the seasonal average and seasonal march, with a higher level of

TABLE 6. Correlation coefficient between the skill and grid size of the 15 CMIP3 and 31 CMIP5 models (total 46). Seasonal average: geographical distribution (20°–50°N, 110°–150°E) of seasonal average precipitation. Seasonal march: time–latitude cross section of precipitation averaged for longitude over China, the Korean Peninsula, and Japan defined in Fig. 7. Extreme events: geographical distribution (20°–50°N, 110°–150°E) of indices. Model skill was calculated against GPCP 1DD data. Grid sizes of models are based on the values listed in Tables 1 and 2.

Term	Skill measure		
	Spatial correlation coefficient R	RMSE	Taylor skill score S
Seasonal average			
Spring (Apr–May)	−0.485[c]	0.316[b]	−0.365[b]
Summer (Jun–Aug)	−0.257[a]	0.155	−0.167
Autumn (Sep–Nov)	−0.566[c]	0.442[c]	−0.520[c]
Winter (Dec–Feb)	−0.267[a]	−0.004	0.191
Annual (Jan–Dec)	−0.444[c]	0.206	−0.525[c]
Seasonal march			
China	−0.243	0.009	−0.205
Korean Peninsula	−0.285[a]	0.249[a]	−0.226
Japan	−0.463[c]	0.349[b]	−0.445[c]
Extreme events			
SDII	−0.224	0.453[c]	−0.487[c]
R5d	−0.300[b]	0.466[c]	−0.546[c]
PMAX	−0.469[c]	0.341[b]	−0.461[c]
CDD	−0.451[c]	0.520[c]	−0.422[c]

[a] Above the 90% statistical significance level.
[b] Above the 95% statistical significance level.
[c] Above the 99% statistical significance level.

statistical significance. As intense rainfall events are often concentrated over a small horizontal scale, models with a higher horizontal resolution tend to reproduce more realistic extreme rainfall events. Moreover, intense rainfall events are common in mountainous regions because of the orographic effect. Models with a higher horizontal resolution can represent orography more realistically, which may lead to the better reproduction of strong orographic rainfall.

The results in Table 6 confirm the advantage of an increased horizontal resolution for the CMIP5 and CMIP3 models. The average grid size of the CMIP5 models (193 km) is smaller than that of the CMIP3 models (254 km). The difference in grid sizes is statistically significant at the 95% level using Student's t test and 98% level using the bootstrap method (see the appendix). The advantage of the CMIP5 models over the CMIP3 models seen when reproducing geographical precipitation distributions (section 4), the seasonal march of the rainy zone (section 5), and extreme precipitation events (section 6) can be partly attributed to the higher horizontal resolution of the CMIP5 models. Watterson et al. (2014) demonstrated that the improvement in the CMIP5 models over the CMIP3

models relates to the improvement in the horizontal resolution of the CMIP5 models. Our results are consistent with their conclusion.

However, Song and Zhou (2014a) reported that the reproducibility of precipitation and the 850-hPa wind over East Asia in summer by the atmospheric models of CMIP5 and CMIP3 does not depend upon the horizontal resolution of the models. In addition, Song and Zhou (2014b) did not find any advantage of higher horizontal resolution in the AOGCMs of CMIP5 with respect to precipitation and the 850-hPa wind over East Asia in summer. The differences between their results and our study can be partly attributed to differences in target models, target region, target season, and skill measure.

b. *Large-scale circulation*

The WPSH plays an important role in the climatology and variability of climate over East Asia. Here, we investigated the reproducibility of the WPSH by CMIP5 and CMIP3 models in conjunction with the reproducibility of precipitation. Figure 13 compares observed 850-hPa geopotential height distributions with those simulated by the models for seasonal and annual averages. The biases of the models are also shown. The observed WPSH (above 1500 m) in spring (Fig. 13a) spreads over Taiwan and to the south of Japan. The CMIP5 MME (Fig. 13f) and CMIP3 MME (Fig. 13p) reproduce the observed WPSH reasonably well, but with a slight overestimation of height (Figs. 13k,u). The observed WPSH in summer (Fig. 13b) extends to the south of Japan with the ridge at about 30°N. Anticlockwise transport of moisture from the humid tropics associated with anticyclonic circulation around the WPSH is primarily responsible for the formation and maintenance of the rainy season over East Asia. The CMIP5 MME (Fig. 13g) reproduces the observed WPSH reasonably well, but the models have a positive bias over Japan and a negative bias to the south of Japan (Fig. 13l), which indicates a northward bias in the location of the WPSH. This bias is consistent with analysis of the AGCMs of CMIP5 by He and Zhou (2014) and Song and Zhou (2014a). The CMIP3 MME (Fig. 11q) also reproduces the observed WPSH reasonably well, but the models have a northward bias in the location of the WPSH (Fig. 13v) with a larger positive bias over Japan compared with CMIP5 (Fig. 13j). This northward bias in position is also consistent with analysis of the AGCMs of CMIP3 by Song and Zhou (2014a).

Positive biases of height over Japan are also found in autumn (Figs. 13m,w) and the annual average (Fig. 13y). Positive biases around Japan by CMIP5 models are smaller than those by CMIP3 models. The erroneous northward shift of the WPSH is consistent with the previous studies for atmospheric models of CMIP5 and CMIP3 (Song and Zhou 2014a; He and Zhou 2014).

Figure 14 compares the RMSEs of the CMIP5 MME with those of the CMIP3 MME for all seasons. The RMSEs of the MME are smaller than those of the AVM for all seasons. The advantage of the CMIP5 MME over the CMIP3 MME is also evident. This suggests that higher reproducibility of the WPSH and its accompanying circulation by the CMIP5 models is closely related to the higher reproducibility of precipitation when compared with the CMIP3 models. In summer, the statistically significant advantage of CMIP5 is evident only for the AVM (green number 2 in a circle).

Figure 15 shows the relationship between the reproducibility of the 850-hPa geopotential height and precipitation in summer for the CMIP5 models over East Asia. The correlation coefficient between height skill and precipitation skill is +0.577, which is statistically significant at the 95% level. If we use the spatial correlation coefficient R and the RMSE as a skill measure, the correlation coefficients between height skill and precipitation skill are positive, but the statistical significance levels are below 90%. Most points are plotted below the diagonal line, indicating that simulating precipitation is much more difficult than simulating the 850-hPa geopotential height. The high correlation coefficient between the skill of height and precipitation means that a model that more accurately simulates the WPSH tends to have a higher reproducibility for precipitation. This is consistent with analysis of the AGCMs of CMIP5 and CMIP3 by Song and Zhou (2014a), in that a better simulation of the summer monsoon circulation is generally associated with a better simulation of the precipitation distribution.

Similarly, the CMIP3 models also show positive correlation coefficient of +0.546 with a statistical significance level above 95%. Precipitation amounts and the frequency of intense rainfall events over East Asia in summer are much larger than those in other seasons, mainly because the WPSH transports large amounts of moisture from the tropics to East Asia in summer. In seasons other than summer, the relationship depicted in Fig. 15 is weak and not statistically significant, suggesting that circulation systems other than the WPSH control the precipitation over East Asia.

According to the assumption that precipitation in summer over East Asia is strongly affected by the main part of the WPSH at lower latitudes, we have tried the same calculations as Fig. 15 but limited the target region to the 850-hPa geopotential height in the subtropics. For the regions 20°–25°N, 20°–30°N, and 20°–35°N for the same longitude range of 110°–150°E, the correlation coefficients between height skill and precipitation skill were 0.38 (not statistically significant), 0.44 (not statistically significant),

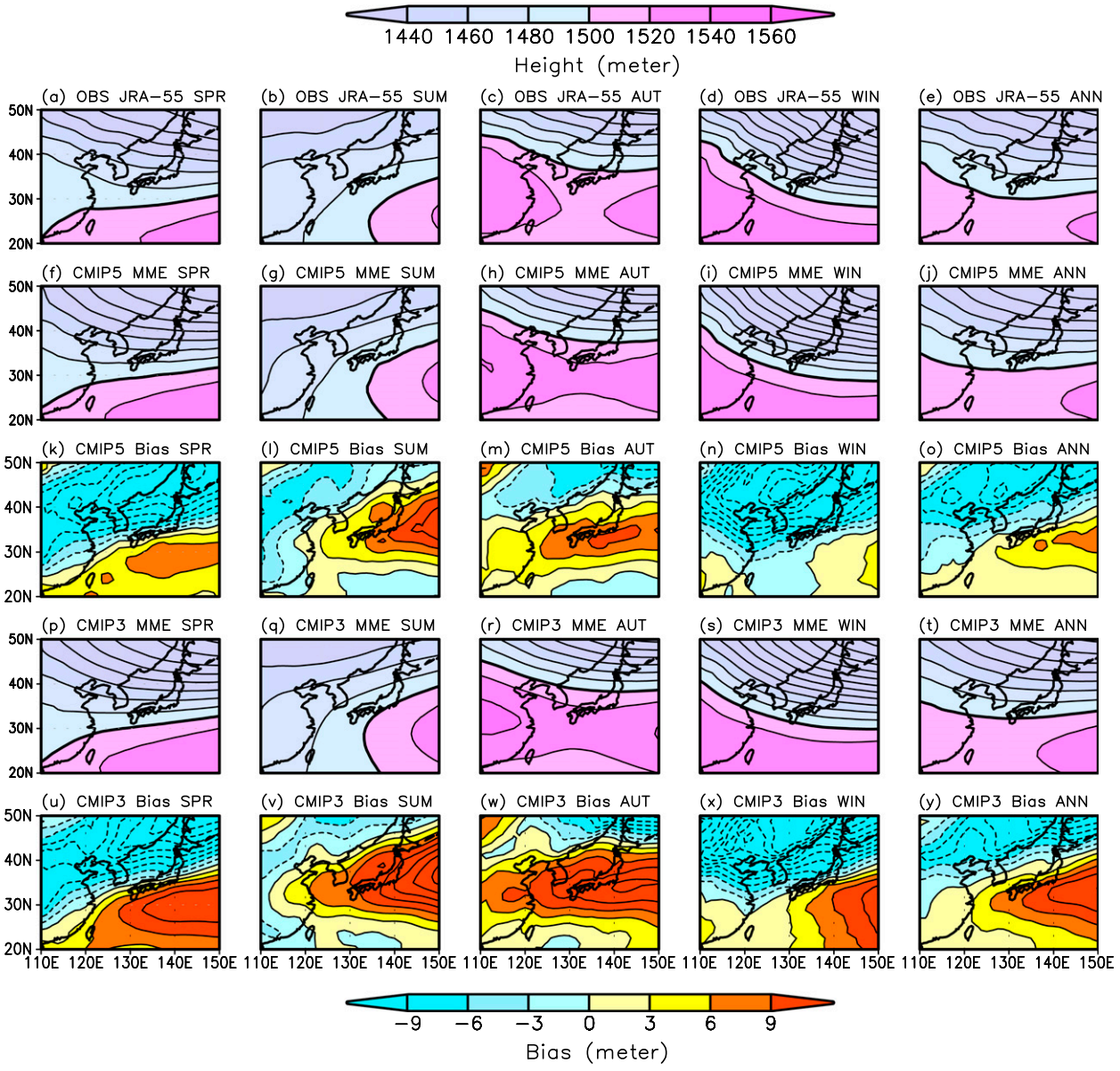

FIG. 13. Geopotential height at 850 hPa for 1981–2000 climatology and bias. (a) the JRA-55 (Ebita et al. 2011) observations for spring (April–May). Contour interval is 20 m. Thick contour denotes 1500 m. (b)–(e) As in (a), but for summer (June–August), autumn (September–November), winter (December–February), and the annual mean (January–December), respectively. (f)–(j) As in (a)–(e), but for CMIP5 MME. (k)–(o) As in (a)–(e), but for the bias of CMIP5 MME; Contour interval is 3 m. (p)–(t) As in (a)–(e), but for CMIP3 MME. (u)–(y) As in (k)–(o), but for the bias of CMIP3 MME.

and 0.45 (significant above the 90% level), respectively. All of these correlation coefficients are lower than those in Fig. 15. This suggests that the reproducibility of precipitation in summer over East Asia (20°–50°N, 110°–150°E) is highly connected with the reproducibility of large-scale circulation at the 850-hPa level over the same region rather than that of the WPSH in the subtropics.

In summary, the advantage of the CMIP5 models over the CMIP3 models can be partly attributed to the improvement in the reproducibility of the WPSH, especially in summer.

c. *Other factors*

1) FLUX ADJUSTMENT

The adoption of flux adjustment into a model implies that the model has large errors in some of the fluxes between atmosphere and ocean. Flux adjustment was

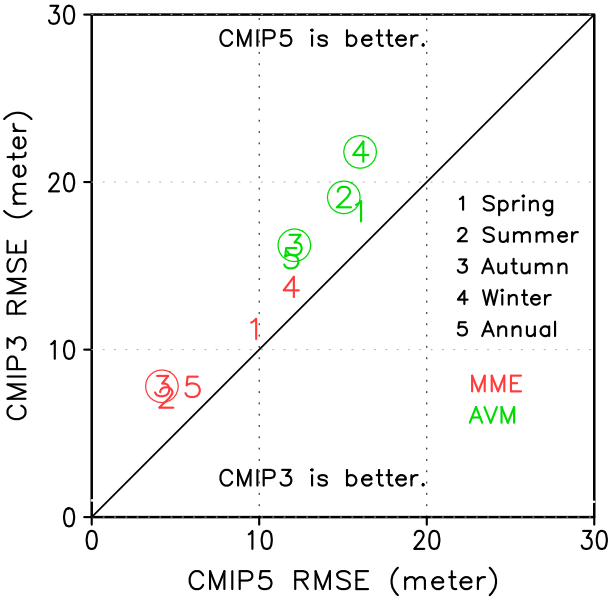

FIG. 14. Comparison of RMSE (m) between the CMIP5 and CMIP3 models for the 850-hPa geopotential height (Fig. 13). The RMSEs were calculated against the JRA-55 data. The MME and AVM are indicated in red and green, respectively. Circles indicate the significance level exceeding 90% based on the bootstrap method (see the appendix).

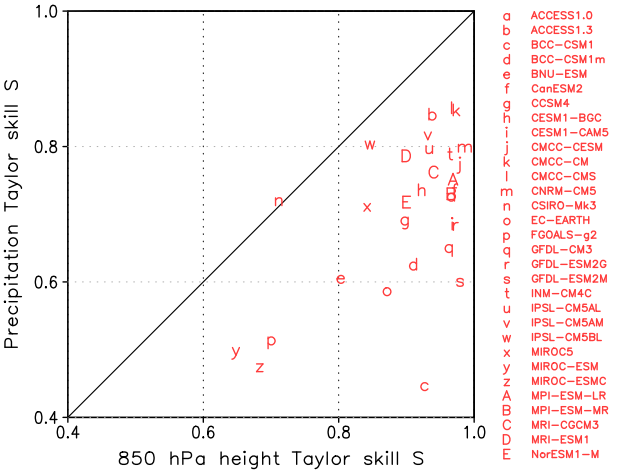

FIG. 15. Relationship between reproducibility of 850-hPa geopotential height and precipitation in summer (June–August) for the CMIP5 models. The horizontal axis is the Taylor skill score for the 850-hPa geopotential height against the JRA-55 data. The vertical axis is the Taylor skill score of precipitation against the GPCP 1DD data. The target area is the East Asia region (20°–50°N, 110°–150°E). The correlation coefficient between height skill and precipitation skill is +0.577, which is statistically significant at the 95% level.

used in some of the CMIP3 models, but in none of the CMIP5 models. It seems reasonable to suggest that the elimination of flux adjustment from the CMIP5 models is convincing evidence of the reduction of various biases and the improvement of physical processes in the CMIP5 models. Consequently, the improvement of physical processes in the CMIP5 models may partly contribute to the improvement in the reproducibility of precipitation over East Asia.

2) SST BIAS

The spurious extension of the cold tongue in sea surface temperature (SST) near the equator commonly found in the CMIP3 models was reduced by 30% in the CMIP5 models. See Fig. 9.14 in Stocker et al. (2013) for details. This improvement in the reproducibility of SST over the tropical Pacific region might lead further to improvement in the reproducibility of the large-scale circulation such as the Hadley and Walker circulations. On the other hand, Song and Zhou (2014b) reported a cold SST bias over the eastern Pacific region in the CMIP5 AOGCMs. Nevertheless, they found that the AOGCMs tend to better simulates precipitation and low-level circulation over East Asia in summer than did the AGCMs, which are forced with observed SSTs. They attributed this paradoxical advantage of the AOGCM to the suppression of evaporation and precipitation over the WPSH due to the cold SST bias. This strengthens the WPSH, which leads to the better representation of precipitation and low-level circulation. The cold SST bias in AOGCM is an example of lucky accident. These SST biases of the AOGCMs over the WPSH can be significantly reduced by the proper treatment of the air–sea coupling processes (Song and Zhou 2014b), as well as of the SST biases of the regional atmosphere–ocean coupled model (Zou and Zhou 2013).

3) CUMULUS CONVECTION

The cumulus (deep) convection scheme greatly affects the reproducibility of precipitation by climate models. Kusunoki and Arakawa (2012) reported that some CMIP3 models using the Arakawa–Schubert scheme tend to show higher skill with respect to precipitation intensity. However, they indicated that the advantage of the Arakawa–Schubert scheme over other convection schemes could not be separated from the skill dependency of horizontal resolution. There is a similar difficulty in separating the deep convection scheme effect from the horizontal resolution effect for the simulated summertime precipitation over India in Sperber and Palmer (1996). It is by no means easy to classify the deep convection schemes used in the CMIP5 models into groups because of the increased complexity of the schemes and the lack of an established standard for the clustering schemes. Also, separation of the deep convection scheme effect from the horizontal-resolution

effect remains challenging for the CMIP5 models. Although we cannot quantify the contribution from the improvement of the deep convection scheme, we believe that the improvements in the CMIP5 models described in this paper can be partly attributed to the improvement of the deep convection schemes.

4) JET STREAM

The observed northward migration of the rainy zone over East Asia is characterized by the corresponding northward migration of the jet stream in the upper troposphere. To clarify the source of the improvement in the CMIP5 models when simulating the seasonal march of the rainy zone, we investigated the reproducibility of the zonal component of the 200-hPa wind (U200) by the CMIP5 and CMIP3 models. Contrary to our expectations, the seasonal march of U200 in summer simulated by the CMIP5 models is worse than that simulated by the CMIP3 models in terms of the RMSE, bias, and Taylor skill score (figure not shown). In particular, the CMIP5 models greatly underestimate the magnitude of U200, whereas the CMIP3 models show almost no bias. These results indicate that the performance with respect to precipitation is primarily affected by the WPSH (Fig. 15), whereas the role of the upper jet stream highlighted in previous studies such as Zhang et al. (2006) deserves further study.

8. Conclusions

The results can be summarized as follows:

1) The CMIP5 and CMIP3 models underestimate precipitation amount over East Asia in the warm season, but overestimate it in the cold season. The models have some difficulty in simulating the seasonal march of the rainy season over China, the Korean Peninsula, and Japan. The models underestimate the precipitation intensity over East Asia.
2) The CMIP5 models show higher reproducibility of precipitation over East Asia than the CMIP3 models with respect to the geographical distribution of precipitation throughout the year, the seasonal march of the rainy season, and extreme precipitation events.
3) Models with higher reproducibility of annual precipitation tend to show higher reproducibility of precipitation intensity in the case of the CMIP5 and CMIP3 models.
4) Correlation analysis using all of the CMIP5 and CMIP3 models revealed that models with a higher horizontal resolution tend to perform better than lower-resolution models. The advantage of the CMIP5 models over the CMIP3 models can be attributed to their higher horizontal resolution.
5) The advantage of the CMIP5 models over the CMIP3 models can be partly attributed to the improvement in the reproducibility of the west Pacific subtropical high, particularly in summer.

Acknowledgments. This study was supported by the Environment Research and Technology Development Fund (A-1201) of the Ministry of the Environment, Japan. We also acknowledge the ''Data Integration and Analysis System (DIAS)'' Fund for National Key Technology from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. This work was partly conducted under the framework of ''the Development of Basic Technology for Risk Information on Climate Change'' supported by the MEXT. We acknowledge the international modeling groups for providing model data for our analysis, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) for collecting and archiving the CMIP5 and CMIP3 multimodel data, and the Joint Scientific Committee (JSC)/CLIVAR Working Group on Coupled Modeling (WGCM). The data archive at the Lawrence Livermore National Laboratory (LLNL) is supported by the Office of Science, U.S. Department of Energy. We also thank the three anonymous reviewers whose valuable comments and suggestions greatly improved the manuscript.

APPENDIX

Bootstrap Method

The statistical significance of differences between the skill of the CMIP5 and CMIP3 models with respect to the average of the multimodel ensemble (MME) and the average skill of individual models (AVM) in Figs. 6, 8, 10, and 14 was evaluated using the bootstrap method (Wilks 2011). The calculations for MME consist of the following steps:

1) Generate a resample of the 31 CMIP5 datasets with replacement (permitting duplication) randomly taken from the combined dataset of 46 models (31 CMIP5 models and 15 CMIP3 models).
2) Calculate the MME average for this sample. In the case of Fig. 6, the target variable is monthly precipitation and the target region is the two-dimensional field over East Asia (20°–50°N, 110°–150°E).
3) Calculate the skill score such as the spatial correlation coefficient R, RMSE, and Taylor skill score

S for the MME average of this resampled CMIP5 data.

4) Generate a resample of the 15 CMIP3 datasets with replacement randomly taken from the combined dataset of the 46 models.
5) Calculate a MME average for this resampled CMIP3 data.
6) Calculate the skill scores for the MME average of this resampled CMIP3 data.
7) Calculate the difference between the CMIP5 skill and the CMIP3 skill. For the RMSE, denote this difference as RMSE5 − RMSE3.
8) Repeat steps 1–7 for 10 000 times.
9) Sort 10 000 samples of RMSE5 − RMSE3 according to their values.
10) Count the number *N* of resampled values greater than the original value of RMSE5 − RMSE3. A large negative value of the original RMSE5 − RMSE3 indicates the advantage of the CMIP5 models over the CMIP3 models. The ratio of *N* to the total sample size of 10 000 is used to estimate as the statistical significance level. For example, $N = 9000$ corresponds to the 90% level.

The procedure is much simpler for the AVM than the MME because the resampled data are just taken from the skill scores of all 46 individual models.

REFERENCES

Adler, R. F., and Coauthors, 2003: The version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–present). *J. Hydrometeor.,* **4,** 1147–1167, doi:10.1175/1525-7541(2003)004<1147:TVGPCP>2.0.CO;2.

Ebita, A., and Coauthors, 2011: The Japanese 55-Year Reanalysis "JRA-55": An interim report. *SOLA,* **7,** 149–152, doi:10.2151/sola.2011-038.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.,* **113,** D06104, doi:10.1029/2007JD008972.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting. Part I: Basic concept. *Tellus,* **57A,** 219–233, doi:10.1111/j.1600-0870.2005.00103.x.

He, C., and T. Zhou, 2014: The two interannual variability modes of the western North Pacific subtropical high simulated by 28 CMIP5–AMIP models. *Climate Dyn.,* **43,** 2455–2469, doi:10.1007/s00382-014-2068-x.

Huffman, G. J., R. F. Adler, M. M. Morrissey, D. T. Bolvin, S. Curtis, R. Joyce, B. McGavock, and J. Susskind, 2001: Global precipitation at one-degree daily resolution from multisatellite observations. *J. Hydrometeor.,* **2,** 36–50, doi:10.1175/1525-7541(2001)002<0036:GPAODD>2.0.CO;2.

——, and Coauthors, 2007: The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *J. Hydrometeor.,* **8,** 38–55, doi:10.1175/JHM560.1.

Kitoh, A., 2004: Effects of mountain uplift on East Asian summer climate investigated by a coupled atmosphere–ocean GCM. *J. Climate,* **17,** 783–802, doi:10.1175/1520-0442(2004)017<0783:EOMUOE>2.0.CO;2.

——, and S. Kusunoki, 2008: East Asian summer monsoon simulation by a 20-km mesh AGCM. *Climate Dyn.,* **31,** 389–401, doi:10.1007/s00382-007-0285-2.

Kusunoki, S., and O. Arakawa, 2012: Change in the precipitation intensity of the East Asian summer monsoon projected by CMIP3 models. *Climate Dyn.,* **38,** 2055–2072, doi:10.1007/s00382-011-1234-7.

——, M. Sugi, A. Kitoh, C. Kobayashi, and K. Takano, 2001: Atmospheric seasonal predictability experiments by the JMA AGCM. *J. Meteor. Soc. Japan,* **79,** 1183–1206, doi:10.2151/jmsj.79.1183.

——, J. Yoshimura, H. Yoshimura, A. Noda, K. Oouchi, and R. Mizuta, 2006: Change of Baiu rain band in global warming projection by an atmospheric general circulation model with a 20-km grid size. *J. Meteor. Soc. Japan,* **84,** 581–611, doi:10.2151/jmsj.84.581.

Lambert, S. J., and G. J. Boer, 2001: CMIP1 evaluation and intercomparison of coupled climate models. *Climate Dyn.,* **17,** 83–106, doi:10.1007/PL00013736.

Meehl, G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer, and K. E. Taylor, 2007: The WCRP CMIP3 multimodel dataset: A new era in climate change research. *Bull. Amer. Meteor. Soc.,* **88,** 1383–1394, doi:10.1175/BAMS-88-9-1383.

Ogata, T., and Coauthors, 2014: Projected future changes in the Asian monsoon: A comparison of CMIP3 and CMIP5 model results. *J. Meteor. Soc. Japan,* **92,** 207–225, doi:10.2151/jmsj.2014-302.

Palmer, T. N., and Coauthors, 2004: Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER). *Bull. Amer. Meteor. Soc.,* **85,** 853–872, doi:10.1175/BAMS-85-6-853.

Reichler, T., and J. Kim, 2008: How well do coupled models simulate today's climate? *Bull. Amer. Meteor. Soc.,* **89,** 303–311, doi:10.1175/BAMS-89-3-303.

Song, F., and T. Zhou, 2014a: Interannual variability of East Asian summer monsoon simulated by CMIP3 and CMIP5 AGCMs: Skill dependence on Indian Ocean–western Pacific anticyclone teleconnection. *J. Climate,* **27,** 1679–1697, doi:10.1175/JCLI-D-13-00248.1.

——, and ——, 2014b: The climatology and interannual variability of East Asian summer monsoon in CMIP5 coupled models: Does air–sea coupling improve the simulations? *J. Climate,* **27,** 8761–8777, doi:10.1175/JCLI-D-14-00396.1.

Sperber, K. R., and T. N. Palmer, 1996: Interannual tropical rainfall variability in general circulation model simulations associated with the Atmospheric Model Intercomparison Project. *J. Climate,* **9,** 2727–2750, doi:10.1175/1520-0442(1996)009<2727:ITRVIG>2.0.CO;2.

——, H. Annamalai, I.-S. Kang, A. Kitoh, A. Moise, A. G. Turner, B. Wang, and T. Zhou, 2013: The Asian summer monsoon: An intercomparison of CMIP5 vs. CMIP3 simulations of the late 20th century. *Climate Dyn.,* **41,** 2711–2744, doi:10.1007/s00382-012-1607-6.

Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K. Averyt, M. Tignor, and H. L. Miller Jr., Eds., 2007: *Climate Change 2007: The Physical Science Basis.* Cambridge University Press, 996 pp.

Stocker, T. F., and Coauthors, 2013: *Climate Change 2013: The Physical Science Basis.* Cambridge University Press, 1535 pp.

[Available online at www.climatechange2013.org/images/report/WG1AR5_ALL_FINAL.pdf.]

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.,* **106,** 7183–7192, doi:10.1029/2000JD900719.

——, R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.,* **93,** 485–498, doi:10.1175/BAMS-D-11-00094.1.

Wang, B., and LinHo, 2002: Rainy season of the Asian–Pacific summer monsoon. *J. Climate,* **15,** 386–398, doi:10.1175/1520-0442(2002)015<0386:RSOTAP>2.0.CO;2.

Watterson, I. G., J. Bathols, and C. Heady, 2014: What influences the skill of climate models over the continents? *Bull. Amer. Meteor. Soc.,* **95,** 689–700, doi:10.1175/BAMS-D-12-00136.1.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences.* International Geophysics Series, Vol. 100, Academic Press, 704 pp.

Xie, P., and P. Arkin, 1997: Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *Bull. Amer. Meteor. Soc.,* **78,** 2539–2558, doi:10.1175/1520-0477(1997)078<2539:GPAYMA>2.0.CO;2.

Yatagai, A., O. Arakawa, K. Kamiguchi, H. Kawamoto, M. I. Nodzu, and A. Hamada, 2009: A 44-year daily gridded precipitation dataset for Asia based on a dense network of rain gauges. *SOLA,* **5,** 137–140, doi:10.2151/sola.2009-035.

Zhang, Y., X. Kuang, W. Guo, and T. Zhou, 2006: Seasonal evolution of the upper-tropospheric westerly jet core over East Asia. *Geophys. Res. Lett.,* **33,** L11708, doi:10.1029/2006GL026377.

Zou, L., and T. Zhou, 2013: Can a regional ocean–atmosphere coupled model improve the simulation of the interannual variability of the western North Pacific summer monsoon? *J. Climate,* **26,** 2353–2367, doi:10.1175/JCLI-D-11-00722.1.