INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **36**: 1114–1133 (2016)
Published online 19 July 2015 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/joc.4406

RMetS
Royal Meteorological Society

Reliability of climate models for China through the IPCC Third to Fifth Assessment Reports

Dabang Jiang,[a,b,c]* Zhiping Tian[a] and Xianmei Lang[a,b,d]

[a] *Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China*
[b] *CAS Center for Excellence in Tibetan Plateau Earth Sciences, Beijing, China*
[c] *Joint Center for Global Change Studies (JCGCS), Beijing, China*
[d] *Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters, Nanjing University of Information Science & Technology, China*

ABSTRACT: Based on observation and reanalysis data, 77 coupled global climate models (GCMs) participating in the Intergovernmental Panel on Climate Change (IPCC) Third (TAR), Fourth (AR4), and Fifth (AR5) Assessment Reports are evaluated in terms of their ability to simulate the mean state and year-to-year variability of surface air temperature at 2 m and precipitation over China and the climatological East Asian monsoon for the late decades of the 20th century. Results show that GCMs reliably reproduce the geographical distribution of the variables considered. Compared with observations, however, most GCMs have topography-related cold biases (although these are smaller than those found in previous studies), excessive precipitation, an underestimated southeast–northwest precipitation gradient, an overestimated magnitude and spatial variability of the interannual variability of temperature and precipitation, and an inadequate strength of the East Asian monsoon circulation. Pairwise comparison reveals that GCMs continue to improve from the TAR via the AR4 to the AR5 for temperature, but have little change for precipitation and the East Asian monsoon. The ability of GCMs varies with season and is affected to certain degree by their horizontal resolutions. Both the arithmetic mean and the median of multiple GCMs are little affected by filtering GCMs in terms of their ability, and the multi-model mean outperforms most of individual GCMs in every respect.

KEY WORDS global climate models; climatology; variability; evaluation; resolution

Received 2 December 2014; Revised 17 May 2015; Accepted 18 May 2015

1. Introduction

Global climate models (GCMs) are complex computer programs that solve a complete system of differential equations built on the fundamental laws of fluid dynamics, physics, chemistry, and biology. They have been widely applied to simulate climate at various spatial and temporal scales. Although GCMs are currently capable of reproducing many of the robust large-scale features of observed climate and climate change in the recent past – including, but not limited to, the evolution of the global mean temperature in the 20th century (Räisänen, 2007; Flato *et al.*, 2013) – they are inherently imperfect because of limited understanding of the real climate system, the non-linear nature of a number of model equations, and the parameterizations for processes, such as convection and cloud microphysics, that are too small-scale or complex to be explicitly resolved. It is thus of crucial importance to assess the reliability of GCMs from various aspects for better insights into climate simulations.

Growing attention has been paid to the evaluation of GCMs at the regional scale in recent years, because, as climate varies with region, it is local rather than global climate change that really matters to the human and natural ecosystem (e.g. Kumar *et al.*, 2014; Perez *et al.*, 2014). In general, GCMs perform somewhat worse in given regions than those for the globe, because their horizontal resolutions are often too coarse to resolve processes and features that are important at regional scales (Flato *et al.*, 2013). This is particularly true in China, a densely populated country where the climate has high spatial and temporal variation, and is influenced by a complex mixture of factors including the tropical and subtropical monsoon, Tibetan Plateau thermodynamics, and low- to high-latitude climate systems (Ding, 1994). Older generation GCMs have been shown to reasonably reproduce the geographical distribution of several key elements over China; however, GCM errors are also significant in some respects, including cold biases and excessive precipitation (Wang and Xiong, 2004; Jiang *et al.*, 2005; Xu *et al.*, 2007). Recently, state-of-the-art GCMs participating in the Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report (AR5) have been applied to simulate the past, present, and future climate within the framework of the Coupled Model Intercomparison Project Phase 5 (CMIP5, Taylor *et al.*, 2012). Their experiments have been applied to analyse the climate of China (e.g. Chen *et al.*, 2014; Sui *et al.*, 2014; Zhou *et al.*, 2014) and can be expected

* Correspondence to: D. Jiang, Climate Change Research Center, Institute of Atmospheric Physics, Chinese Academy of Sciences, P. O. Box 9804, Beijing 100029, China. E-mail: jiangdb@mail.iap.ac.cn

to be extensively applied in climate-related fields until the release of the next IPCC report. This raises the questions as to how the AR5 GCMs perform over China, and how they compare to their predecessors in this respect.

Several assessments of the AR5 GCMs in reproducing the mean state and changing trend of climate over China have been carried out (Xu and Xu, 2012; Guo *et al.*, 2013; Su *et al.*, 2013; Chen and Frauenfeld, 2014a, 2014b; Hua *et al.*, 2014). However, only about half, or less than half, of 40-plus AR5 GCMs have been investigated, hampering an objective estimation of the GCM ability. Moreover, most researchers only use the latest GCMs in the AR5, making it impossible to evaluate whether the current generation of GCMs represent an improvement over their earlier versions through a pairwise comparison of the different stages of simulations within the IPCC framework. In addition, the horizontal resolution has been proposed to be important in accurately simulating the East Asian precipitation based on numerical experiments of a regional climate model (Gao *et al.*, 2006), whereas it seems that no evident relationship has been detected between GCM skills and resolutions (Chen and Frauenfeld, 2014b; Song and Zhou, 2014). The multi-GCM mean, usually obtained by applying equal weights to each GCM or by calculating the median of GCMs, has been suggested as superior to individual GCMs (Jiang *et al.*, 2005; Xu *et al.*, 2007). Whether there is in fact a discernible influence, either from the horizontal resolution or the algorithm when calculating the mean of multiple GCMs, remains unresolved in the context of the IPCC simulations.

Besides the previously investigated mean state, another key aspect of climate is its year-to-year variability. It both directly and indirectly affects many climate processes, such as water resources and plant growth, that in turn feed back to the climate system, and it also leads to a range of disturbances, such as extreme climate events, that are more difficult for society to adapt to than an altered mean climate (Schär *et al.*, 2004; Huntingford *et al.*, 2013; Thornton *et al.*, 2014). Despite the importance of interannual variability, very few studies have attempted to elucidate how its observed spatial pattern and magnitude are reproduced by GCMs over China. Previously, Lu and Fu (2010) showed that there is a strong interannual variability of summer (June–July–August, JJA) precipitation over East Asia east of 100°E, and that 12 GCMs participating in the IPCC Fourth Assessment Report (AR4) demonstrated different skills in simulating the observed pattern. Whether the interannual variability of basic climate variables, namely temperature and precipitation, can be reliably reproduced by GCMs needs to be explicitly investigated.

Given this context, the purpose of this article is to present an evaluation of the ability of the GCMs participating in the IPCC Third (TAR), Fourth, and Fifth Assessment Reports from the perspective of climatology and year-to-year variability. The key questions to address are: (1) to what extent observed climatology and interannual climate variability over China are simulated by GCMs; (2) whether GCMs have improved from the IPCC TAR to the AR5; and (3) whether there is a detectable effect arising from the horizontal resolution of GCMs or from the algorithm to calculate the multi-GCM mean.

2. Data and methods

Model data are obtained from the following sources: the control simulations of seven GCMs forced by time-evolving equivalent CO_2 or greenhouse gases from the mid-20th century to 1990, then Special Report on Emissions Scenarios A2 emissions scenario to 2000 in the IPCC TAR; the 20th Century Climate in Coupled Models (20C3M) simulations of 23 GCMs forced by anthropogenic and natural forcings in the IPCC AR4; and the historical simulations of 47 GCMs for the past ∼1.5 centuries with all forcings in the IPCC AR5. According to the availability of the requisite data, 77, 76, 60, and 71 GCMs are applied for analysis of temperature, precipitation, and meridional winds at 10 m in winter (December–January–February, DJF) and at 850 hPa in summer, respectively. Basic information about GCMs and experiments are provided in Table 1. More details are available online at http://www.ipcc.ch/.

Temperature and precipitation data used for model evaluation are taken from the CN05.2 daily dataset, with a half-degree horizontal resolution, established through *in situ* data at 2416 stations over China by the National Climate Center of the China Meteorological Administration (Wu and Gao, 2013). Reanalysis data are obtained from the National Centers for Environmental Prediction and National Center for Atmospheric Research (NCEP–NCAR) reanalysis of monthly winds at 10 m and at 850 hPa, with a $2.5° \times 2.5°$ horizontal resolution (Kalnay *et al.*, 1996). Both kinds of data are hereafter referred to as observation for convenience.

Considering that some modelling groups provide multiple ensemble runs for the same experiment (shown in Table 1), all realizations are taken to avoid potential sampling biases and then averaged into a set of data model by model. Because the horizontal resolution of GCMs differs from one another, all GCM and NCEP–NCAR wind data are linearly interpolated to a half-degree horizontal resolution, corresponding to the grid mesh of the CN05.2 temperature and precipitation products. The multi-GCM mean is then calculated using both the ensemble mean with equal weight (hereafter referred to as the ensemble mean) and the median of the chosen GCMs. Based on the temporal coverage of observational and GCM data, the period 1961–2000 is chosen for analysis of temperature and precipitation and 1979–2000 for analysis of wind. This is because the CN05.2 data are from 1961 onwards, and because the NCEP–NCAR reanalysis data are more reliable after 1979 owing to the inclusion of satellite data.

3. Results

3.1. Temperature climatology over China

Based on 4470 grid points in China, the spatial correlation coefficient (SCC), the standard deviation, and the

Table 1. Basic information about the 77 GCMs and their experiments included in this study, and the data availability of wind at 10 m and 850 hPa.

Model ID		Country	Atmospheric resolution	Integration period	Experiment and ensemble size	Wind at 10 m	Wind at 850 hPa
Seven climate models in the IPCC TAR					Control		
01	CCSR/NIES	Japan	~5.6° × 5.5°, L20	1890–2000	1	–	Yes
02	CGCM2	Canada	3.75° × ~3.7°, L10	1900–2000	1	–	–
03	CSIRO-Mk2	Australia	~5.6° × 3.2°, L9	1881–2000	1	–	Yes
04	ECHAM4/OPYC3	Germany	~2.8° × 2.8°, L19	1860–2000	1	–	–
05	GFDL-R30	USA	3.75° × ~2.2°, L14	1961–2000	1	–	–
06	HadCM3	UK	3.75° × 2.5°, L19	1950–2000	1	–	–
07	NCAR-PCM	USA	~2.8° × 2.8°, L18	1900–2000	1	Yes	–
Twenty-three climate models in the IPCC AR4					20C3M		
08	BCCR-BCM2.0	Norway	~2.8° × 2.8°, L31	1850–2000	1	Yes	Yes
09	CCSM3	USA	~1.4° × 1.4°, L26	1870–2000	7	–	Yes
10	CGCM3.1(T47)	Canada	3.75° × ~3.7°, L31	1850–2000	1	Yes	Yes
11	CGCM3.1(T63)	Canada	1.875° × ~1.9°, L31	1850–2000	1	Yes	Yes
12	CNRM-CM3	France	~2.8° × 2.8°, L45	1860–2000	1	Yes	Yes
13	CSIRO-Mk3.0	Australia	1.875° × ~1.9°, L18	1871–2000	3	Yes	Yes
14	CSIRO-Mk3.5	Australia	1.875° × ~1.9°, L18	1871–2000	3	Yes	Yes
15	ECHAM5/MPI-OM	Germany	1.875° × ~1.9°, L32	1860–2000	4	Yes	Yes
16	FGOALS-g1.0	China	~2.8° × 3–6°, L9	1850–2000	3	Yes	Yes
17	GFDL-CM2.0	USA	2.5° × 2°, L24	1861–2000	3	Yes	Yes
18	GFDL-CM2.1	USA	2.5° × 2°, L24	1861–2000	3	Yes	Yes
19	GISS-AOM	USA	4° × 3°, L20	1850–2000	2	Yes	Yes
20	GISS-EH	USA	5° × 4°, L20	1880–2000	5	Yes	Yes
21	GISS-ER	USA	5° × 4°, L20	1880–2000	9	Yes	Yes
22	INGV-SXG	Italy	1.125° × ~1.1°, L19	1870–2000	1	–	Yes
23	INM-CM3.0	Russia	5° × 4°, L21	1871–2000	1	Yes	Yes
24	IPSL-CM4	France	3.75° × ~2.5°, L19	1860–2000	1	Yes	Yes
25	MIROC3.2(hires)	Japan	1.125° × ~1.1°, L56	1850–2000	1	Yes	Yes
26	MIROC3.2(medres)	Japan	~2.8° × 2.8°, L20	1850–2000	3	Yes	Yes
27	MRI-CGCM2.3.2	Japan	~2.8° × 2.8°, L30	1851–2000	5	Yes	Yes
28	PCM	USA	~2.8° × 2.8°, L18	1890–2000	4	–	Yes
29	UKMO-HadCM3	UK	3.75° × 2.5°, L19	1860–2000	2	Yes	Yes
30	UKMO-HadGEM1	UK	1.875° × 1.25°, L38	1860–2000	2	Yes	Yes
Forty-seven climate models in the IPCC AR5					Historical		
31	ACCESS1.0	Australia	1.875° × 1.25°, L38	1850–2005	1	Yes	Yes
32	ACCESS1.3	Australia	1.875° × 1.25°, L38	1850–2005	1	Yes	Yes
33	BCC-CSM1.1	China	~2.8° × 2.8°, L26	1850–2099	1	Yes	Yes
34	BCC-CSM1.1(m)	China	1.125° × ~1.1°, L26	1850–2005	1	Yes	Yes
35	BNU-ESM	China	~2.8° × 2.8°, L26	1850–2005	1	Yes	Yes
36	CanESM2	Canada	1.875° × ~1.9°, L35	1850–2005	5	Yes	Yes
37	CCSM4	USA	1.25° × ~0.9°, L26	1850–2005	6	–	Yes
38	CESM1(BGC)	USA	1.25° × ~0.9°, L26	1850–2005	1	–	Yes
39	CESM1(CAM5)	USA	1.25° × ~0.9°, L26	1850–2005	1	–	Yes
40	CESM1(FASTCHEM)	USA	1.25° × ~0.9°, L26	1850–2005	1	–	Yes
41	CESM1(WACCM)	USA	2.5° × ~1.9°, L23	1850–2005	1	–	Yes
42	CMCC-CESM	Italy	3.75° × ~3.7°, L39	1850–2005	1	Yes	Yes
43	CMCC-CM	Italy	0.75° × ~0.75°, L31	1850–2005	1	Yes	Yes
44	CMCC-CMS	Italy	1.875° × ~1.9°, L95	1850–2005	1	Yes	Yes
45	CNRM-CM5	France	~1.4° × 1.4°, L31	1850–2005	10	Yes	Yes
46	CNRM-CM5-2	France	~1.4° × 1.4°, L31	1850–2005	1	Yes	Yes
47	CSIRO-Mk3.6.0	Australia	1.875° × ~1.9°, L18	1850–2005	1	Yes	Yes
48	EC-EARTH	Europe	1.125° × ~1.1°, L62	1850–2005	1	–	–
49	FGOALS-g2	China	~2.8° × 3–6°, L26	1850–2005	1	–	Yes
50	FGOALS2-s	China	~2.8° × 1.7°, L26	1850–2005	3	Yes	Yes
51	FIO-ESM	China	~2.8° × 2.8°, L26	1850–2005	1	–	Yes
52	GFDL-CM2.1	USA	2.5° × 2°, L24	1861–2015	1	Yes	Yes
53	GFDL-CM3	USA	2.5° × 2°, L48	1860–2005	1	Yes	Yes
54	GFDL-ESM2G	USA	2.5° × 2°, L24	1861–2005	1	Yes	Yes
55	GFDL-ESM2M	USA	2.5° × 2°, L24	1861–2005	1	Yes	Yes
56	GISS-E2-H	USA	2.5° × 2°, L40	1850–2005	5	Yes	Yes
57	GISS-E2-H-CC	USA	2.5° × 2°, L40	1850–2005	1	Yes	Yes

Table 1. Continued

Model ID		Country	Atmospheric resolution	Integration period	Experiment and ensemble size	Wind at 10 m	Wind at 850 hPa
58	GISS-E2-R	USA	2.5° × 2°, L40	1850–2005	1	Yes	Yes
59	GISS-E2-R-CC	USA	2.5° × 2°, L40	1850–2005	1	Yes	Yes
60	HadCM3	UK	3.75° × 2.5°, L19	1860–2005	1	Yes	Yes
61	HadGEM2-AO	South Korea	1.875° × 1.25°, L38	1860–2005	1	Yes	Yes
62	HadGEM2-CC	UK	1.875° × 1.25°, L60	1860–2005	1	Yes	Yes
63	HadGEM2-ES	UK	1.875° × 1.25°, L38	1860–2005	4	Yes	Yes
64	INM-CM4	Russia	2° × 1.5°, L21	1850–2005	1	Yes	Yes
65	IPSL-CM5A-LR	France	3.75° × ~1.9°, L39	1850–2005	4	Yes	Yes
66	IPSL-CM5A-MR	France	2.5° × ~1.3°, L39	1850–2005	1	Yes	Yes
67	IPSL-CM5B-LR	France	3.75° × ~1.9°, L39	1850–2005	1	Yes	Yes
68	MIROC4h	Japan	~0.56° × 0.56°, L56	1850–2005	1	Yes	Yes
69	MIROC5	Japan	~1.4° × 1.4°, L40	1850–2005	3	Yes	Yes
70	MIROC-ESM	Japan	~2.8° × 2.8°, L80	1850–2005	3	Yes	Yes
71	MIROC-ESM-CHEM	Japan	~2.8° × 2.8°, L80	1850–2005	1	Yes	Yes
72	MPI-ESM-LR	Germany	1.875° × ~1.9°, L47	1850–2005	3	Yes	Yes
73	MPI-ESM-MR	Germany	1.875° × ~1.9°, L95	1850–2005	1	Yes	Yes
74	MPI-ESM-P	Germany	1.875° × ~1.9°, L47	1850–2005	1	Yes	Yes
75	MRI-CGCM3	Japan	1.125° × ~1.1°, L48	1850–2005	3	Yes	Yes
76	NorESM1-M	Norway	2.5° × ~1.9°, L26	1850–2005	3	Yes	Yes
77	NorESM1-ME	Norway	2.5° × ~1.9°, L26	1850–2005	1	Yes	Yes

centred root-mean-square error (CRMSE) of each simulation against the observed climatology of annual and seasonal temperatures for the period 1961–2000 are calculated individually. The Taylor diagrams (Figure 1, Taylor, 2001) show that SCCs range from 0.55 to 0.98, indicative of a good agreement between simulated and observed distributions of annual and seasonal temperatures. Normalized standard deviations are 0.80–1.19, 0.78–1.14, 0.87–1.35, 0.64–1.24, and 0.84–1.15 for the annual, winter, spring (March–April–May, MAM), summer, and autumn (September–October–November, SON) temperatures, respectively. That is, most GCMs reliably reproduce the spatial variability of annual and seasonal temperatures, but overestimate the variability in spring. Normalized CRMSEs are 0.27–0.68 for the year, 0.28–0.60 for winter, 0.32–0.96 for spring, 0.26–0.84 for summer, and 0.24–0.65 for autumn. Taken together, the GCMs reliably simulate the annual and seasonal temperatures, with relatively little variation between the individual models, and GCMs have an overall better performance in winter and autumn than in the other seasons, owing to a better reproducibility of both geographical distribution and spatial variability.

In general, the normalized CRMSEs are the smallest for the AR5 GCMs, but the largest for the TAR GCMs (Figure 1). That means the AR5 GCMs perform best, while the TAR GCMs perform worst. Furthermore, the seven TAR GCMs are compared to their successors in the AR4 and AR5. It is noteworthy that if there are different versions of AR4 or AR5 GCMs for one specific TAR GCM, only the high-resolution version is taken. Similarly, 17 pairs of high-resolution AR4 and AR5 GCMs are chosen from different climate modelling groups and are then compared with each other. It is found that GCMs have an obvious improvement from the TAR to the AR4 and AR5, excluding the HadCM3 series (Table S1, Supporting Information), as the CRMSEs of the individual TAR GCMs are systematically larger than those of their AR4 and AR5 counterparts for the annual and seasonal temperatures. Meanwhile, 12 of 17 AR5 GCMs outperform their AR4 versions, while the remaining five AR5 GCMs have similar statistics as that of their AR4 predecessors (Table S1). Thus, the ability of GCMs in reproducing the annual and seasonal temperatures continues to improve from the TAR via AR4 to AR5, particularly from the TAR to AR4.

The original horizontal resolution of the 77 GCMs ranges from approximately 0.56° × 0.56° to 5.6° × 5.5°. Here we classify all 77 GCMs into three groups: 13 low-resolution GCMs with the grid area above 3° × 3°, 31 mid-resolution GCMs with the grid area smaller than 3° × 3° and larger than 2° × 2°, and 33 high-resolution GCMs with the grid area below 2° × 2°. It is noteworthy that no objective threshold truly defines the low-, mid-, and high-resolution GCMs, and that this classification aims to assess the overall effect of model resolution, although the fact that the low-resolution group of models is much smaller than the others may play a role. For the annual and seasonal temperatures, the normalized CRMSE averages are 0.32–0.40, 0.35–0.51, and 0.39–0.58 for the 33 high-, 31 mid-, and 13 low-resolution individual GCMs, respectively. Moreover, Figure 2(a) illustrates how the normalized CRMSEs of GCMs against observation grow with the area of original grid mesh, and hence the ability of GCMs in reproducing the annual and seasonal temperatures over China is enhanced when the horizontal resolution becomes finer, a trend that is statistically significant at the 99% confidence level.

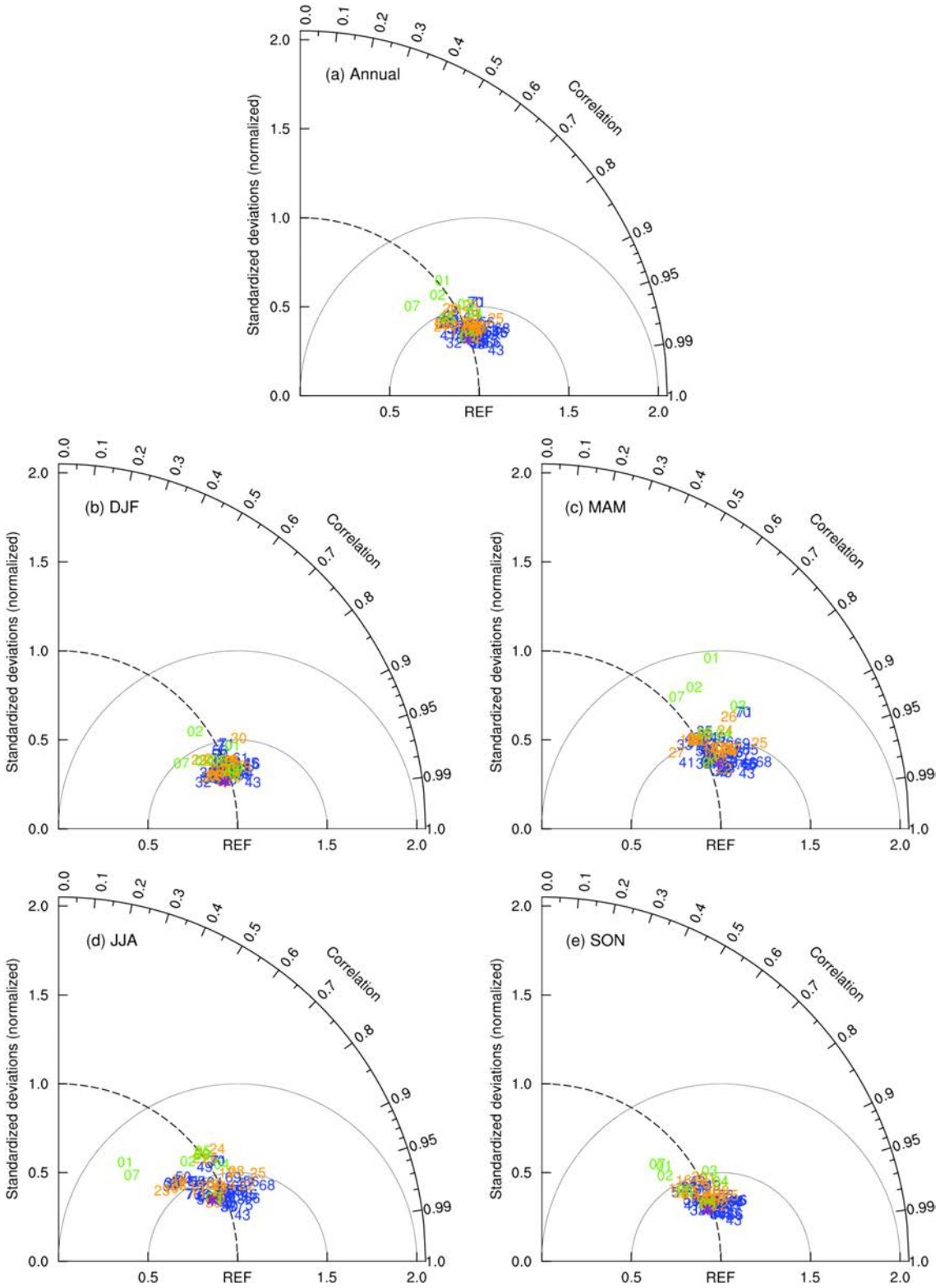

Figure 1. Taylor diagrams displaying normalized pattern statistics of climatological (a) annual, (b) DJF, (c) MAM, (d) JJA, and (e) SON temperatures over China between 77 GCMs and observation for the period 1961–2000. The radial co-ordinate gives the standard deviation normalized by the observed value, and the angular co-ordinate gives the correlation with observation. The normalized CRMSE between a GCM and observation (marked as REF) is their distance apart. Numbers indicate GCMs listed in Table 1. Colour coding is green for TAR, orange for AR4, and blue for AR5 GCMs. Red and purple asterisks indicate the ensemble mean and the median of the 77 GCMs, respectively.

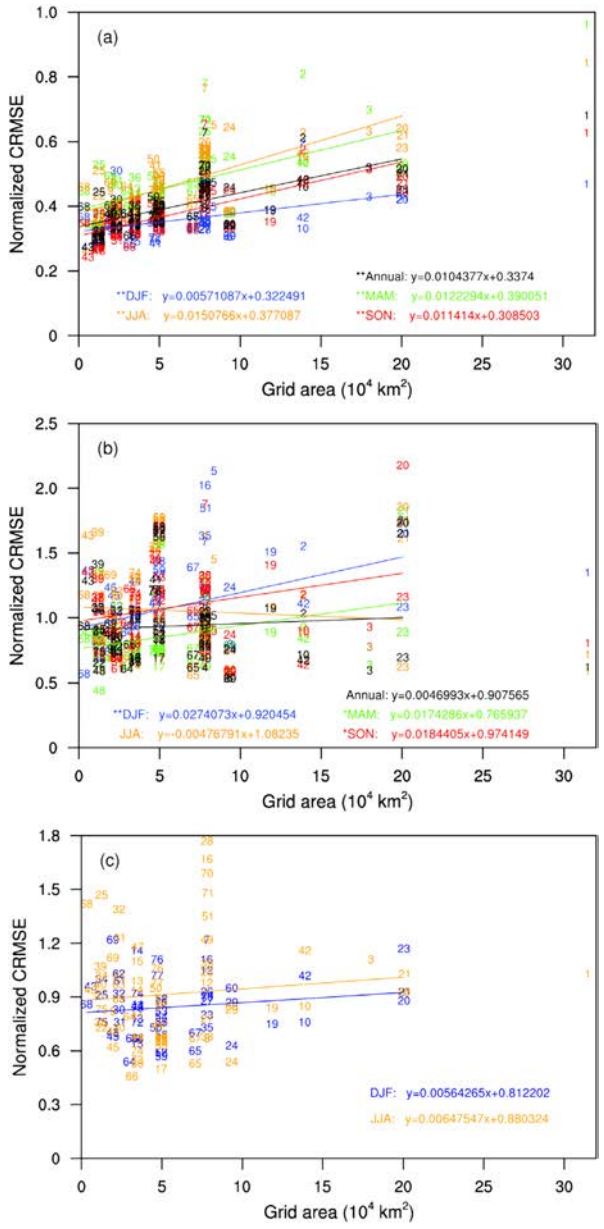

Figure 2. The vertical axis is the normalized CRMSE of GCMs against observation for (a) temperature and (b) precipitation over China for the period 1961–2000 and (c) the East Asian monsoon for the period 1979–2000; the horizontal axis is the original grid mesh area averaged over China for the GCMs. Numbers correspond to the GCMs as listed in Table 1. Black, blue, green, orange, and red indicate the annual, DJF, MAM, JJA, and SON, respectively. Straight lines represent the least-square linear fitting with equal weight for all GCMs except the number 1 because that GCM has too coarse resolution. The fitting equations are shown in the lower corner of the panels, in which * (**) indicates that the linear trend is statistically significant at the 90% (99%) confidence level.

Given that there are no obvious outliers in the Taylor diagrams (Figure 1), all 77 GCMs are used to calculate the ensemble mean and the median of GCMs. It is found that the multi-model mean algorithms have very little effect, as they give almost the same statistics (Figure 1). Comparatively, no single GCM is best for all the aspects considered. The multi-GCM mean outperforms most of individual GCMs for the annual and seasonal temperatures over China, as already found in global studies examining the mean climate (e.g. Lambert and Boer, 2001; Gleckler *et al.*, 2008). This superiority is argued to be because of the inclusion of a large number of diverse GCMs, which tends to reduce the effect of natural internal climate variability and cancel offsetting errors (Pierce *et al.*, 2009). Another possible explanation is that the GCM solutions scatter more or less evenly about the truth, and the errors behave partly like random noise that can be efficiently removed by averaging (Reichler and Kim, 2008; Knutti *et al.*, 2010).

On a large scale, the observed annual and seasonal temperatures generally decrease from the south to the north, and feature a large extent of low values over the Tibetan Plateau because of topographic effects (Figure 3). Such a geographical distribution is well reproduced by individual GCMs and their means. However, most GCMs still underestimate national average temperatures (Figure 3), supporting previous evaluations using part of the present GCMs (Jiang *et al.*, 2005; Xu *et al.*, 2007; Xu and Xu, 2012; Guo *et al.*, 2013; Chen and Frauenfeld, 2014a). Cold biases average −0.93 °C for the year, −1.12 °C for winter, −0.15 °C for spring, −0.38 °C for summer, and −2.09 °C for autumn in terms of the median of the 77 GCMs. It is noteworthy that, except in autumn, they are weaker than the recent estimate of 0.81–2.37 °C for the 20th century and its second half by 22 AR4 and 20 AR5 GCMs (Chen and Frauenfeld, 2014a), owing largely to the different choice of period and GCMs. These cold biases require further analyses of the radiation energy budget and associated processes, and larger biases in cold than in warm seasons imply that GCMs may incorrectly represent snow–albedo feedbacks. In western China, the annual and seasonal temperatures are notably underestimated over the Tibetan Plateau, consistent with cold biases over the eastern Tibetan Plateau found for 24 AR5 GCMs (Su *et al.*, 2013), and in the Tarim Basin; however, they are overestimated along the Aerhchin and Qilian mountains and the Tien Shan (Figure 3). These biases are obviously related to regional topography and hence to the treatment of complex terrain in GCMs. More than half of eastern China has cold biases for the year, winter, summer, and autumn, but warm biases for spring. Temperature is generally lower than that observed in North China, southeastern Northeast China, and the Sichuan Basin, but higher in northwestern Northeast China except for autumn. Moreover, individual GCMs agree on most of the above biases, as the consistency, defined as the percentage of the number of GCMs sharing the same sign of the median bias at each grid, averages 73–83% for the annual and seasonal temperatures over the country (right panels of Figure 3).

3.2. Interannual variability of temperature over China

The ability of GCMs in reproducing the interannual variability of temperature is much weaker than for time mean temperature over China as manifested by lower SCCs and greater normalized CRMSEs (Figures 1 and 4). SCCs range from 0.02 to 0.88 for the year, −0.36 to 0.83 for winter, −0.07 to 0.85 for spring, −0.17 to 0.71 for summer, and 0.01 to 0.82 for autumn, with several outliers with

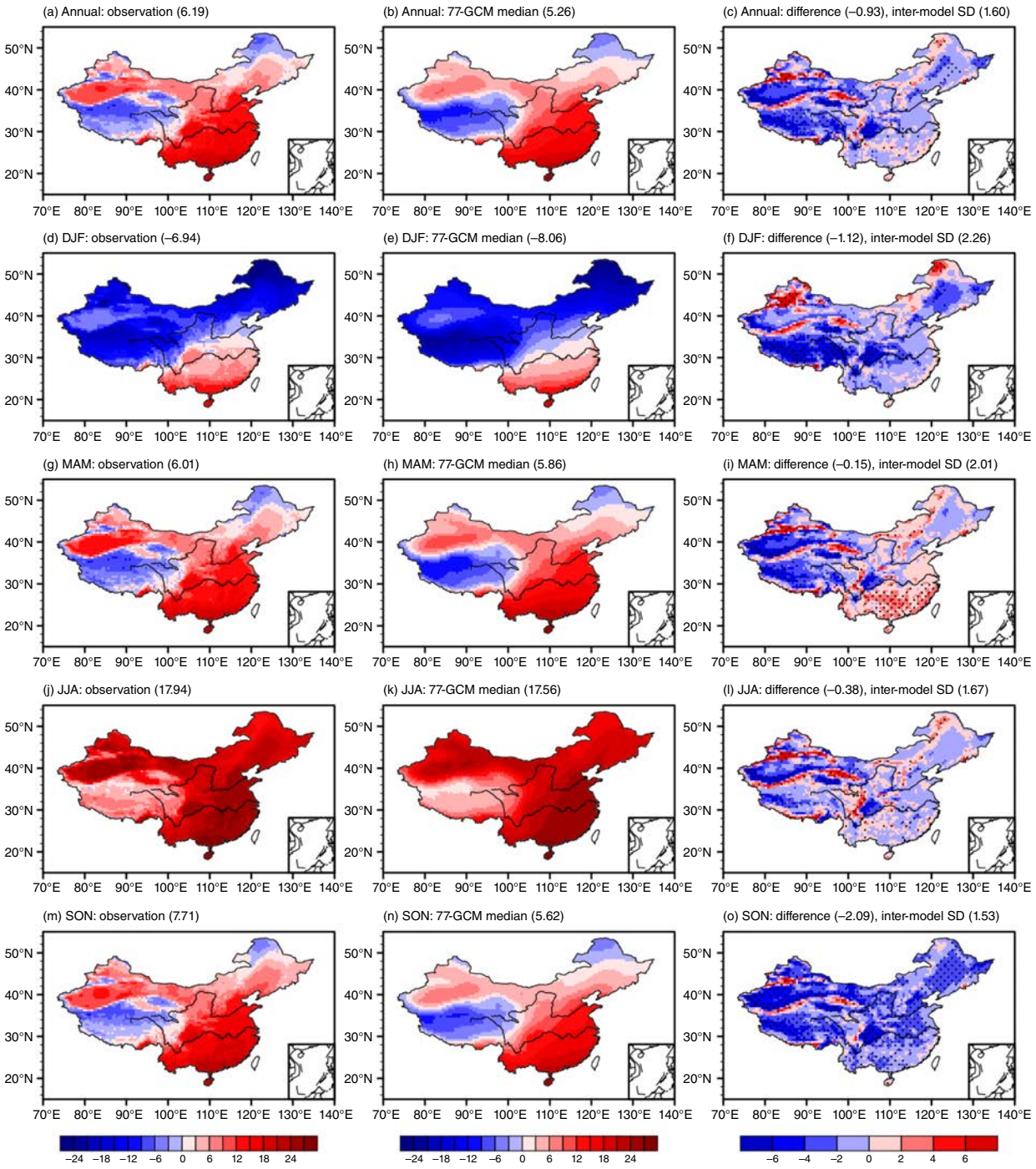

Figure 3. Climatological annual and seasonal temperatures (units: °C) over China for the period 1961–2000 as obtained from observation (left column), the median of the 77 GCMs (middle column), and the difference between the median and observation (right column). The regional average value in China and the inter-model standard deviation of the difference averaged over the country (right column, inter-model SD, units: °C) are given in parentheses. The two solid black lines indicate the Yellow River valley in the north and the Yangtze River valley in the south. The dotted areas in the right panels represent regions where at least 80% of the GCMs share the same sign of bias.

negative or very small SCCs. Normalized standard deviations are 0.59–2.43, 0.61–2.45, 0.56–2.62, 0.88–3.61, and 0.68–1.94 for the year, winter, spring, summer, and autumn, respectively. Normalized CRMSEs are 0.56–2.34 for the year, 0.65–2.42 for winter, 0.61–2.54 for spring, 0.96–3.26 for summer, and 0.78–1.89 for autumn. As such, most GCMs reliably reproduce the geographical distribution of the interannual variability of annual and seasonal temperatures, as their SCCs are relatively high. Most GCMs overestimate the spatial variability of the interannual variability of annual and seasonal temperatures, particularly in summer, as their standard deviations are greater than observed. This derives from the stronger large-scale geographical gradient of interannual variability in the models than observed (Figure S1, Supporting Information). GCMs have an overall worse performance

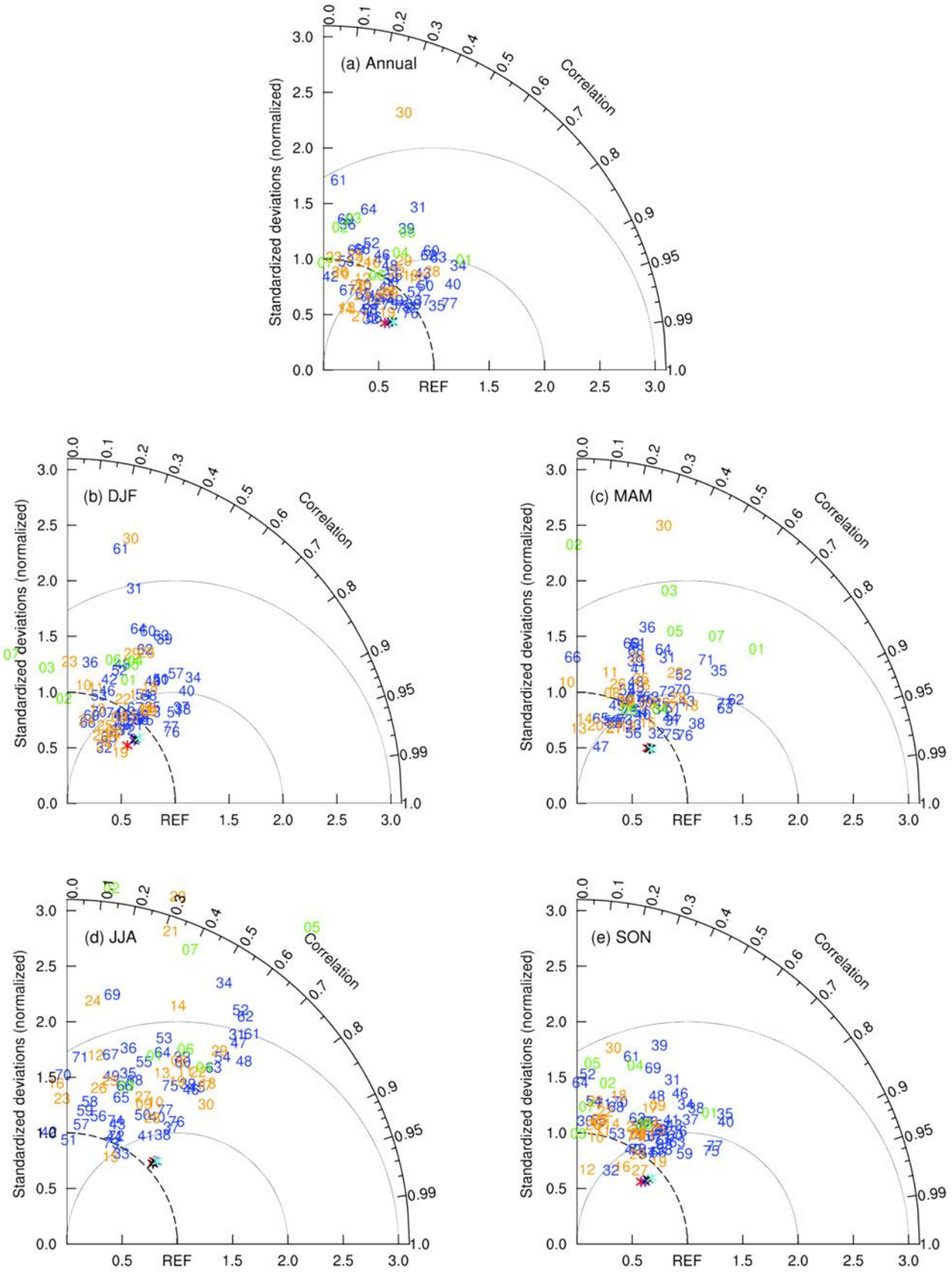

Figure 4. Taylor diagrams displaying normalized pattern statistics of the interannual variability of (a) annual, (b) DJF, (c) MAM, (d) JJA, and (e) SON temperatures over China for the period 1961–2000 between 77 GCMs and observation. Red, purple, black, and cyan asterisks indicate the ensemble mean and the median of the 77 GCMs as well as the ensemble mean and the median of the 57 reliable GCMs, respectively. Other aspects are the same as Figure 1.

for summer than that for the year and the other seasons because of a worse reproducibility of the spatial pattern and variability of the interannual variability.

The TAR GCMs are worse than their successors in the AR4 and AR5 as a whole. Excluding the HadCM3 and CCSR/NIES series, all the other five TAR GCMs have greater normalized CRMSEs than their AR4 and AR5 counterparts (Table S2). Meanwhile, 9 (5) of 17 AR5 GCMs generally perform better (worse) than their AR4 predecessors, and the remaining three pairs of AR4 and AR5 GCMs are similar in skills (Table S2). Therefore, the ability of GCMs in reproducing the interannual variability of annual and seasonal temperatures over China is enhanced from the TAR to the AR4, while changing little from the AR4 to the AR5.

When viewed from model resolution, the 33 high-, 31 mid-, and 13 low-resolution individual GCMs have average SCCs of 0.48–0.58, 0.39–0.60, 0.33–0.49, and normalized CRMSEs of 0.97–1.52, 0.88–1.57, and 1.08–1.71 for the interannual variability of annual and seasonal temperatures, respectively. Figure 5(a) further illustrates that when the area of the original grid mesh becomes large, normalized CRMSEs of GCMs against observation grow statistically significantly for spring and summer, but change little for the year, winter, and autumn. Hence, except for spring and summer, the ability of GCMs in reproducing the interannual variability of annual and seasonal temperatures is not systematically affected by the horizontal resolution.

Considering that there is a large scatter in the performance of individual GCMs (Figure 4), a positive SCC and a normalized CRMSE below 2.00 are set to identify relatively reliable GCMs. Twenty (four TAR, eight AR4, and eight AR5) GCMs are excluded accordingly. The means of the remaining 57 GCMs, obtained by both the ensemble mean and the median, give almost the same evaluation statistics as those of all 77 GCMs (Figure 4). This indicates little influence of the algorithm to calculate the multi-GCM mean and of the filtering of GCMs in terms of their ability. In addition, the multi-GCM mean outperforms all (most) individual GCMs for the interannual variability of summer and autumn (annual, winter, and spring) temperatures (Figure 4).

As illustrated in Figure 6, the observed interannual variability of annual and seasonal temperatures over China generally increases from the south to the north, and is characterized by an increased trend from summer via autumn and spring to winter. Both the major features are reliably simulated by GCMs, as is the magnitude of interannual variability. For example, the median of the 57 reliable GCMs matches well with the observation on both annual and seasonal scales (Figure 6). For the whole country, the simulated and observed interannual variabilities of annual temperature are almost the same, and the GCM–observation discrepancy mainly lies in an exaggerated interannual variability of seasonal temperature, with an average of 0.18 °C for winter, 0.17 °C for spring, 0.12 °C for summer, and 0.14 °C for autumn. Regionally, the simulated interannual variability of annual temperature

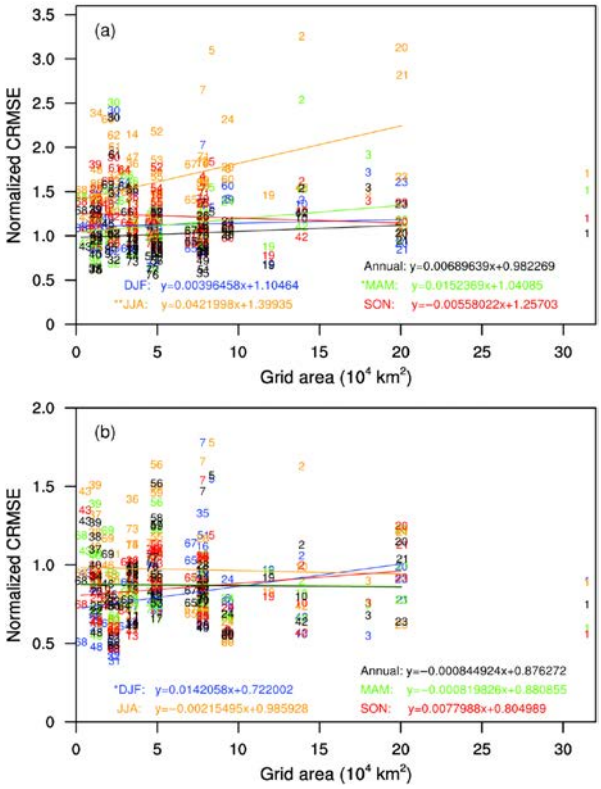

Figure 5. The vertical axis shows the normalized CRMSE of the GCMs against observation for the interannual variability of annual and seasonal (a) temperatures and (b) precipitation over China for the period 1961–2000; the horizontal axis is the original grid mesh area averaged over China for GCMs. Numbers correspond to the GCMs as listed in Table 1. Black, blue, green, orange, and red indicate the annual, DJF, MAM, JJA, and SON, respectively. Straight lines represent the least-square linear fitting with equal weight for all GCMs except the number 1 because that GCM has too coarse resolution. The fitting equations are shown in the lower corner of the panels, in which * (**) indicates that the linear trend is statistically significant at the 90% (99%) confidence level.

is too small in northern Xinjiang, most of Inner Mongolia, northern Northeast China, and parts of Southeast China and eastern coastal areas. In winter, the interannual variability is obviously excessive over the southern Tibetan Plateau northeastwards to Northeast China, but too small in northwestern Xinjiang, northern central Inner Mongolia, and part of South China. The simulated interannual variability is slightly greater than observed in most of China for the other three seasons. The inter-GCM consistency for the median bias averages 75, 78, 79, and 80% for winter, spring, summer, and autumn, respectively.

3.3. Precipitation climatology over China

In terms of the climatological mean state, the ability of GCMs in reproducing precipitation is much weaker than for temperature over China (Figures 1 and 7), as found in global-scale assessments (Räisänen, 2007; Flato *et al.*, 2013). For the annual mean (Figure 7(a)), SCCs are 0.35–0.86; normalized standard deviations are 0.58–2.15; and normalized CRMSEs are 0.54–1.73 for all 76 GCMs. Comparatively, SCCs are more dispersive among GCMs

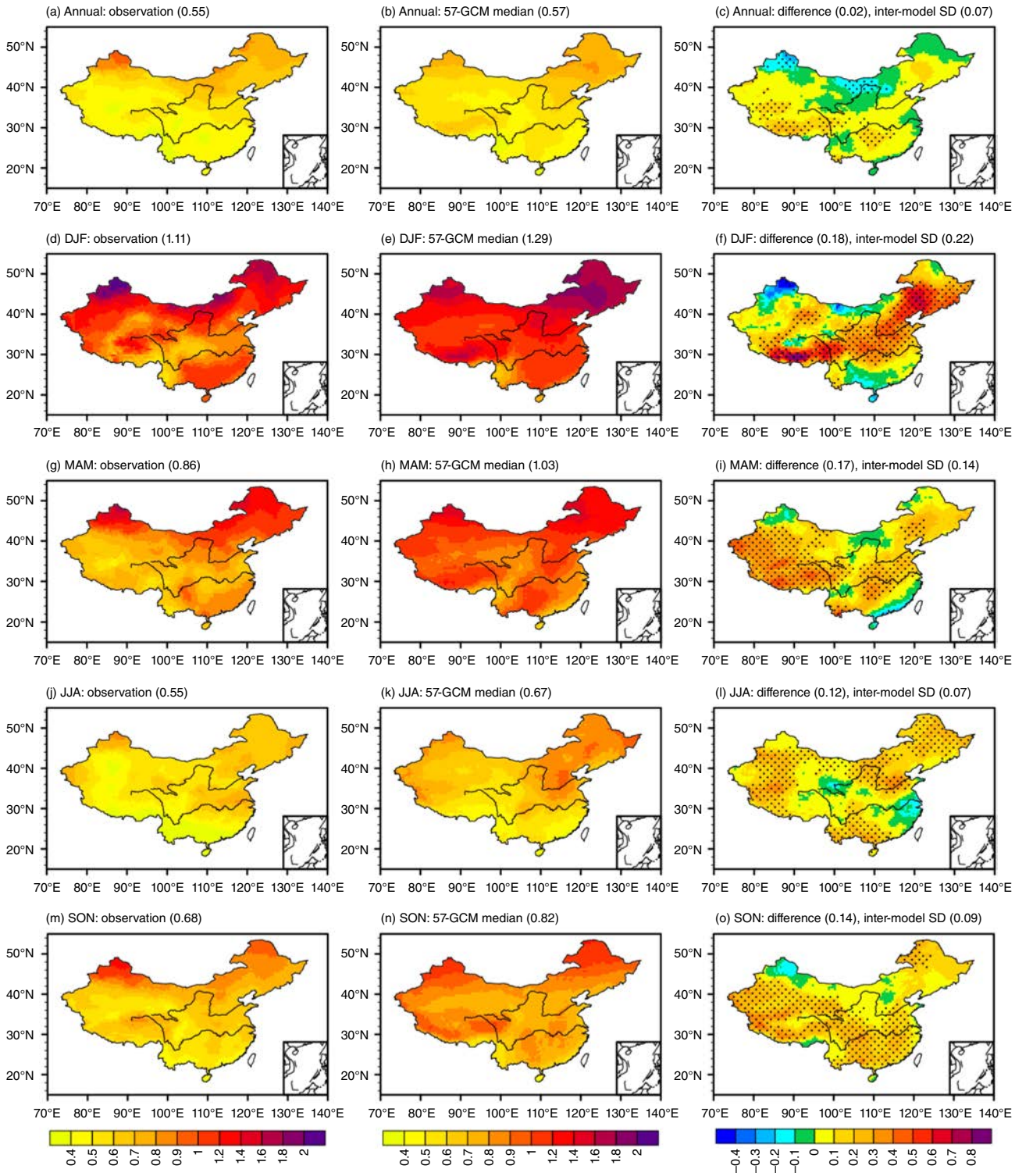

Figure 6. Interannual variability of annual and seasonal temperatures (units: °C) over China for the period 1961–2000 obtained from observation (left column), the median of the 57 reliable GCMs (middle column), and the difference between the median and observation (right column). The regional average value in China and the inter-model standard deviation of the difference averaged over the country (right column, inter-model SD) are given in parentheses. The dotted areas in the right panels represent regions where at least 80% of the GCMs share the same sign of bias.

for the individual seasons, and both normalized standard deviations and CRMSEs become larger in winter and autumn (Figure 7(b)–(e)). In general, GCMs reliably simulate the geographical distribution of annual and seasonal precipitation, as most SCCs are at relatively high levels. Except in spring, the majority of normalized standard deviations are larger than one, and hence GCMs overestimate the spatial variability of precipitation, owing to the stronger large-scale gradient of precipitation in the models than observed (Figure S2). The GCM biases come mainly from unrealistic simulations of the spatial pattern for spring, and from both the spatial pattern and variability

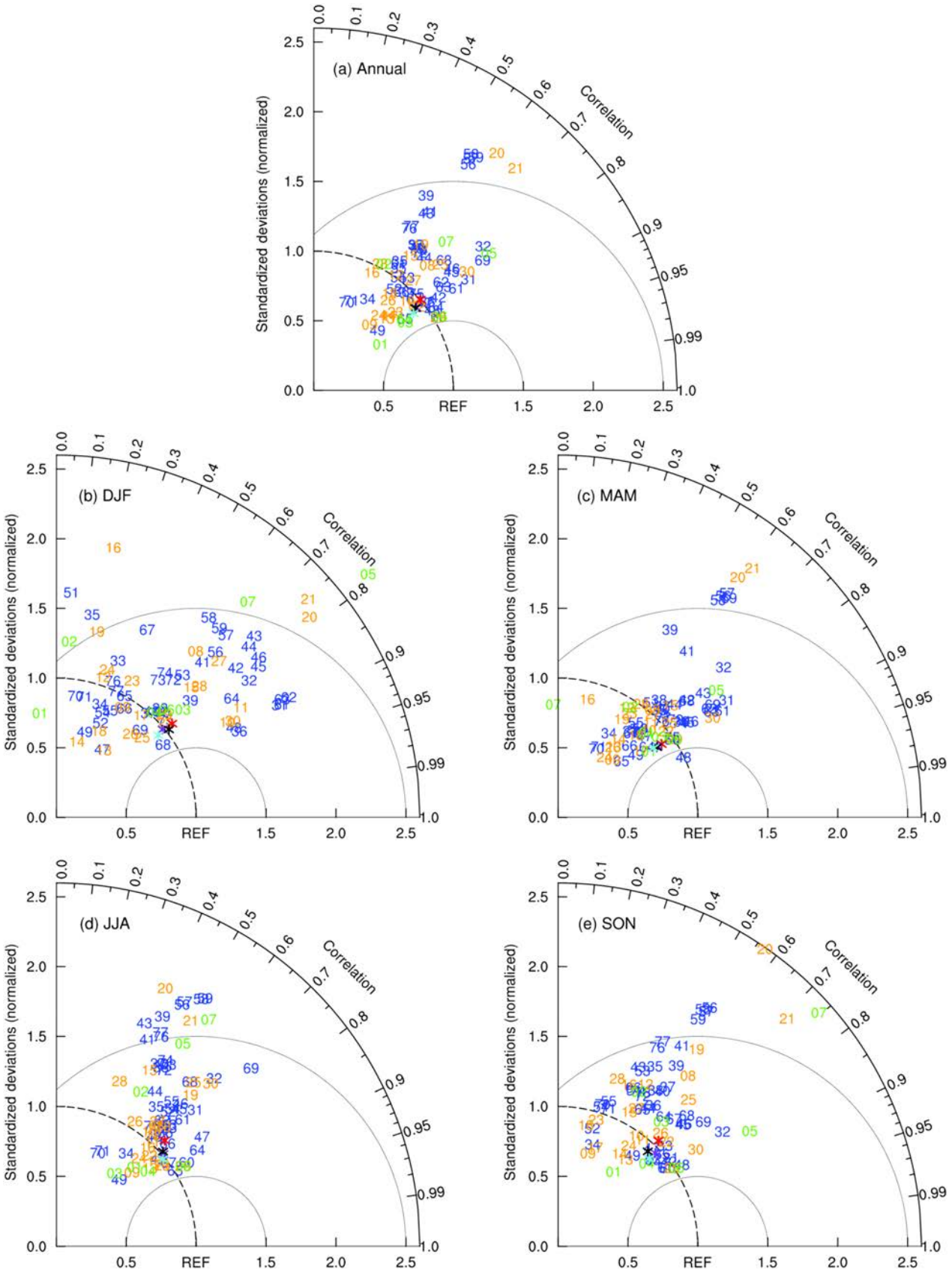

Figure 7. Taylor diagrams displaying normalized pattern statistics of climatological (a) annual, (b) DJF, (c) MAM, (d) JJA, and (e) SON precipitation over China for the period 1961–2000 between 76 GCMs and observation. Red, purple, black, and cyan asterisks indicate the ensemble mean and the median of the 76 GCMs, as well as the ensemble mean and the median of the 57 reliable GCMs, respectively. Other aspects are the same as Figure 1.

for the annual mean and the other seasons. The ability of GCMs decreases from spring via summer and autumn to winter.

Compared to the seven TAR GCMs, their successors in the AR4 and AR5 generally perform better than CGCM2, GFDL-R30, and NCAR-PCM, worse than CCSR/NIES, CSIRO-Mk2, and ECHAM4/OPYC3, and similarly to HadCM3 (Table S3). The median of the seven TAR GCMs has a normalized CRMSE of 0.58–0.75, which is comparable to both 0.64–0.83 and 0.62–0.89 obtained, respectively, from the medians of their AR4 and AR5 successors. On the other hand, 5 (2) of 17 AR5 GCMs perform better (worse) than their counterparts in the AR4, and the remaining ten pairs of AR4 and AR5 GCMs have similar skills overall (Table S3). As such, the ability of GCMs in reproducing the annual and seasonal precipitation over China has not been improved from the TAR via AR4 to AR5 when considering these seven series of GCMs, and this is also the case for the 17 common pairs of GCMs between AR4 and AR5. The latter differs from a recent study by Chen and Frauenfeld (2014b) showing that the AR5 GCMs agree better with precipitation observation over China than the AR4 GCMs, in which they directly compare 20 AR5 GCMs with 22 AR4 GCMs without considering whether or not GCMs develop from the same modelling group, and hence performing a fair inter-GCM comparison.

Averaged across the 33 high-, 30 mid-, and 13 low-resolution individual GCMs, SCCs are 0.62–0.76, 0.47–0.65, and 0.59–0.72, and normalized CRMSEs are 0.78–0.99, 0.82–1.22, and 0.70–1.11 for the annual and seasonal precipitation, respectively. Furthermore, Figure 2(b) illustrates that when the area of the original grid mesh increases, the normalized CRMSEs of the GCMs against observation grow statistically significantly for winter, spring, and autumn, but vary little for the year and summer. In other words, the ability of GCMs in reproducing winter, spring, and autumn precipitation over China is affected by model resolution. This agrees partly with Gao *et al.* (2006) who found that the horizontal resolution plays an important role in accurately simulating the East Asian precipitation based on numerical experiments undertaken by a regional climate model with various horizontal resolutions, and is in line with the result that no evident relationship is seen between atmospheric model skills and resolutions for the East Asian summer precipitation (Song and Zhou, 2014).

Based on the Taylor diagrams (Figure 7), 57 GCMs with positive SCCs and normalized CRMSEs below 1.50 are regarded as relatively reliable GCMs to simulate the annual and seasonal precipitation over China. Both the ensemble mean and the median of the 57 GCMs give very close values for the annual and seasonal precipitation (Figure 7), indicative of little effect of the algorithms on calculating the multi-GCM mean. Moreover, those values are also very close to the statistics of all 76 GCMs (Figure 7), indicative of little effect of screening GCMs in terms of their ability before the mean is calculated. Additionally, the multi-GCM mean outperforms most of individual GCMs for both the annual and seasonal precipitation (Figure 7).

Large-scale annual and seasonal precipitation over China is generally reduced from the southeast to the northwest (Figure 8). Precipitation is characterized by an obvious seasonality and its annual sum is determined first by summer precipitation, due mainly to monsoon rainfall, and second by spring and autumn precipitation as detailed by Sui *et al.* (2013). These features are well simulated by most individual GCMs. For example, the geographical distribution of annual and seasonal precipitation as derived from the median of the 57 reliable GCMs agrees well with observation (Figure 8). Compared to the observation, however, precipitation is overestimated by GCMs in most of China except for Southeast China – a trend already visible in subsets of the present GCMs (e.g. Jiang *et al.*, 2005; Xu *et al.*, 2007; Xu and Xu, 2012; Chen and Frauenfeld, 2014b) – and by a national average of 27, 67, 48, 11, and 22% for annually, and in winter, spring, summer, and autumn, respectively. Convective and stratiform precipitation and the underlying convective and microphysical parameterization schemes, as well as atmospheric moisture convergence and surface evaporation, need to be explicitly investigated to understand the model biases in future studies. Such an overestimation may also be partly explained by the fact that the observed precipitation has not been corrected for gauge undercatch, which can be substantial, particularly in winter (e.g. Adam and Lettenmaier, 2003). Additionally, there is no statistically significant relationship between the country-averaged precipitation and temperature biases, except for a negative correlation in summer (larger cold biases in models that more strongly overestimate precipitation) among the 52 commonly reliable GCMs for both climatological temperature and precipitation. This does not support the suggestion that the excessive evaporative cooling because of overestimated precipitation leads to cold biases in current GCMs (Mueller and Seneviratne, 2014), except summer when a negative temperature–precipitation correlation is physically most expected, because the surface energy budget in midlatitudes tends to be more sensitive to variations in soil moisture and cloudiness in summer than in the other seasons. On the whole, GCMs underestimate the southeast–northwest precipitation gradient over the country. Annual and seasonal precipitation is excessive in arid, semi-arid, and semi-humid regions in the northwest, but too small in humid regions in the southeast. The consistency of the 57 GCMs for the median bias averages 75% for summer and 85–92% for the annual mean and the other seasons.

3.4. Interannual variability of precipitation over China

GCMs can reliably reproduce the geographical distribution of the interannual variability of precipitation over China, as SCCs between individual GCMs and observation range from 0.19 to 0.92 on the annual and seasonal scales (Figure 9). Normalized standard deviations range from 0.26 to 2.16 and are greater than one for more than half of 75 GCMs (excluding ECHAM4/OPYC3 where annual precipitation data are not available) except in winter. The spatial variability of the interannual variability of

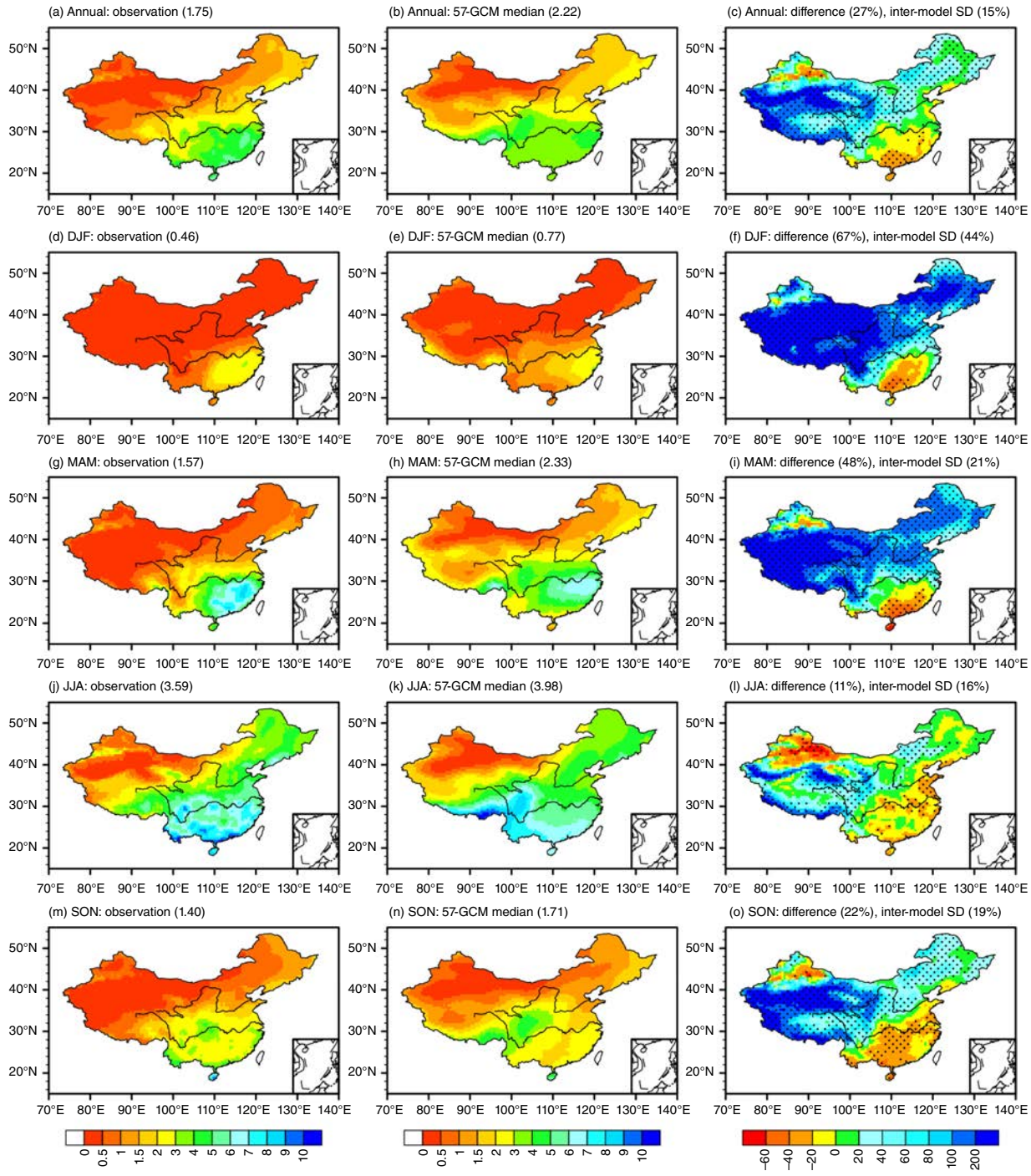

Figure 8. Climatological annual and seasonal precipitation over China for the period 1961–2000 as obtained from observation (left column, units: mm day^{-1}), the median of the 57 reliable GCMs (middle column, units: mm day^{-1}), and the difference in percentage between the median and observation (right column). The regional average value in China and the inter-model standard deviation of the difference in percentage averaged over the country (right column, inter-model SD) are given in parentheses. The dotted areas in the right panels represent regions where at least 80% of the GCMs share the same sign of bias.

annual, spring, summer, and autumn precipitation is therefore overestimated by most GCMs. The majority of normalized CRMSEs lie in the range 0.50–1.50, and they mainly result from the unrealistic simulations of the spatial variability of interannual precipitation variability. In general, GCM biases vary little between the seasons. The GCMs simulate the interannual variability of precipitation similarly to its climatology, worse than the climatological temperature, and better than the interannual temperature variability (Figures 1, 4, 7, and 9).

Compared to the six TAR GCMs (excluding ECHAM4/OPYC3), their AR4 and AR5 counterparts generally perform better than CGCM2, CSIRO-Mk2, GFDL-R30, and NCAR-PCM, worse than CCSR/NIES, and similarly to HadCM3 for the interannual variability of annual and seasonal precipitation (Table S4). Meanwhile,

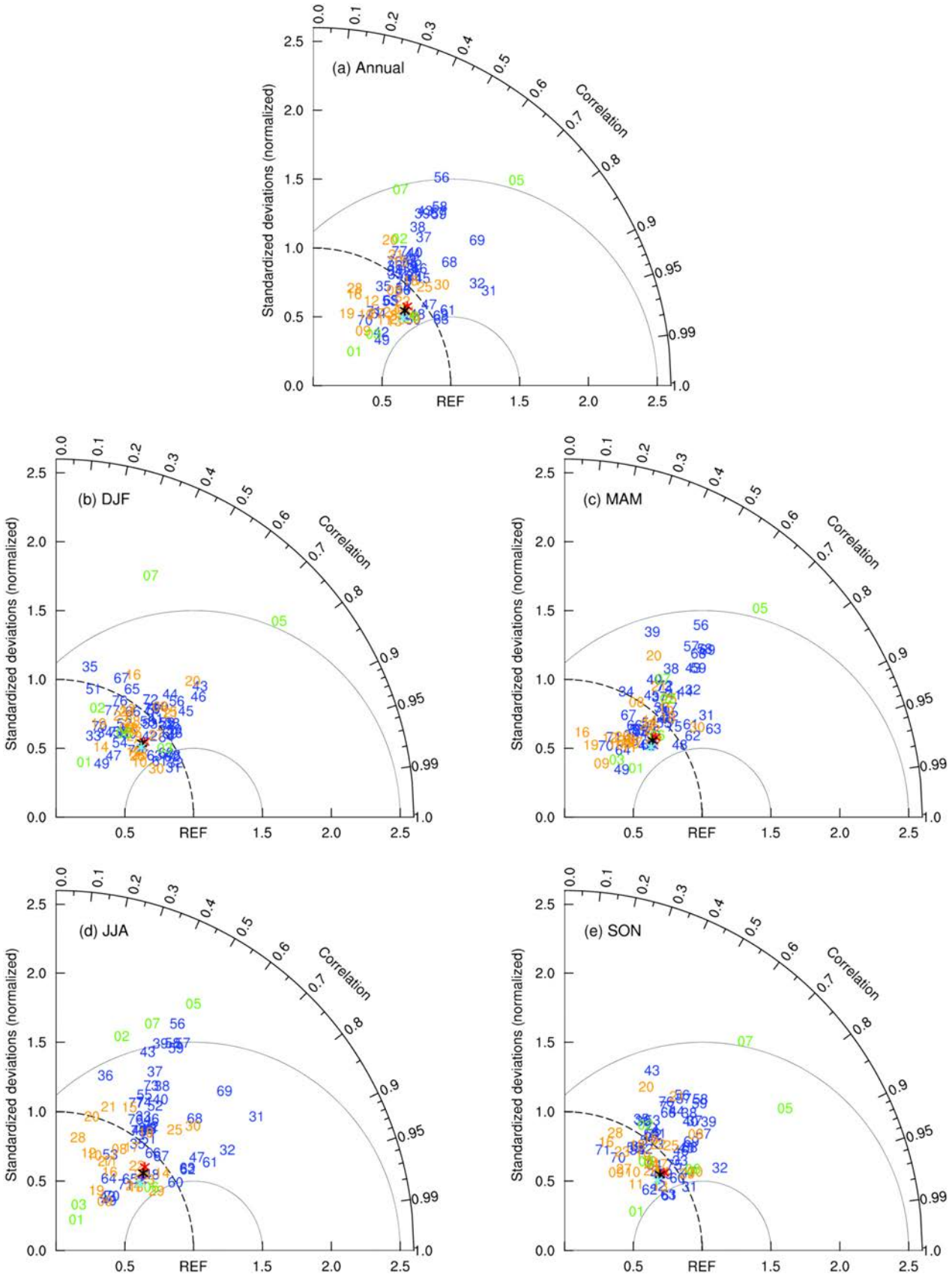

Figure 9. Taylor diagrams displaying normalized pattern statistics of the interannual variability of (a) annual, (b) DJF, (c) MAM, (d) JJA, and (e) SON precipitation over China for the period 1961–2000 between the 75 GCMs and observation. Red, purple, black, and cyan asterisks indicate the ensemble mean and the median of the 75 GCMs as well as the ensemble mean and the median of the 69 reliable GCMs, respectively. Other aspects are the same as Figure 1.

8 (4) out of 17 AR5 GCMs underperform (outperform) their counterparts in the AR4, and the remaining five pairs of AR4 and AR5 GCMs have similar skills (Table S4). Therefore, the ability of GCMs in reproducing the interannual variability of annual and seasonal precipitation over China increases from the TAR to AR4 and AR5 in terms of three generations of the six GCMs, and reduces from the AR4 to AR5 in terms of 17 pairs of GCMs to certain extent.

The 33 high-, 29 mid-, and 13 low-resolution individual GCMs give average SCCs of 0.61–0.77, 0.55–0.65, and 0.56–0.76, normalized standard deviations of 0.95–1.17, 0.88–1.06, and 0.80–0.89, and normalized CRMSEs of 0.65–0.95, 0.82–0.92, and 0.69–0.83 for the interannual variability of annual and seasonal precipitation, respectively. There are no obvious differences between the three groups. More specifically, when the area of the original grid mesh increases, normalized CRMSEs of GCMs against observation vary little, except for winter when an increased trend is statistically significant at the 90% confidence level (Figure 5(b)). Therefore, the ability of GCMs in reproducing the interannual variability of annual and seasonal (except winter) precipitation over China is not affected by model resolution.

Considering the large scatter for individual GCMs in simulating the interannual variability of annual and seasonal precipitation (Figure 9), a normalized CRMSE below 1.50 is set to identify relatively reliable GCMs. Out of 75 GCMs, 6 GCMs fail to meet this requirement and are filtered out accordingly. The evaluation statistics of both the ensemble mean and the median of the remaining 69 GCMs for the interannual variability of annual and seasonal precipitation are very close to each other, and also close to those of all 75 GCMs (Figure 9), indicating that the effect of both the algorithms to calculate the multi-GCM mean and the screening of GCMs in terms of their ability is quite limited. In addition, the multi-GCM mean outperforms most of individual GCMs.

Like climatological annual and seasonal distributions of precipitation, the interannual precipitation variability decreases from the southeast to the northwest over the country and generally has the largest values in summer, the second largest values in spring and autumn, and the smallest values in winter (Figure 10). The median of the 69 reliable GCMs, for example, agrees well with observation in terms of the geographical distribution on both annual and seasonal scales (Figure 10). The seasonality of interannual precipitation variability is also reproduced reasonably well. Quantitatively, however, GCMs notably overestimate the interannual variability of precipitation in most of China, with a high inter-GCM consistency. This is particularly obvious in arid areas in western China and at the western and southernmost parts of the Tibetan Plateau annually, and in summer and autumn, in Northeast China and western China excluding northwestern Xinjiang for winter, and in most of western China and eastern Inner Mongolia for spring. By contrast, a deficit occurs in most of Southeast China. Taking the country as a whole, the interannual precipitation variability is overestimated by 22, 40, 39, 15, and 29% for the year, winter, spring, summer, and autumn, respectively. It is noteworthy that the GCMs underestimate the relative precipitation variability (except in summer), as characterized by the coefficient of variation (Figure S3), because the relative overestimate in the standard deviation is smaller than that in the mean precipitation.

3.5. East Asian winter and summer monsoons

The eastern part of China is within the East Asian monsoon region. In winter, cold high-pressure systems dominate the East Asian continent, while warm low-pressure systems dominate the adjacent oceans, owing to the difference in thermal capacity between land and ocean. The opposite situation generally holds true in summer. Accordingly in the low troposphere, cold and dry northerly winds blow in winter, and warm and wet southerly winds blow in summer. Given that the monsoon is most significant in the near-surface atmosphere over coastal East Asia in winter and at 850 hPa over eastern China in summer (Ding, 1994), meridional winds at 10 m over 2400 grid points within the regions of 25°–40°N and 120°–140°E plus 10°–25°N and 110°–130°E are used to measure the East Asian winter monsoon (e.g. Chen *et al.*, 2000), and meridional winds at 850 hPa over 1125 grid points within the region of 20°–40°N and 105°–120°E are used to measure the East Asian summer monsoon (e.g. Jiang and Tian, 2013).

For the climatology of winter meridional winds at 10 m in the target region (Figure 11(a)), SCCs are 0.33–0.86 for all the 60 available GCMs (Table 1); normalized standard deviations are 0.50–1.68; and normalized CRMSEs are 0.54–1.22 across all GCMs and are larger than one for 10 GCMs. In summer, there is a large spread in the ability of all the 71 available GCMs to simulate meridional winds at 850 hPa in the target region (Figure 11(b)). SCCs range from −0.44 to 0.93 and are negative for PCM, CMCC-CESM, and CSIRO-Mk3.6.0; normalized standard deviations are 0.35–2.18; and normalized CRMSEs are 0.46–1.77 across all GCMs and are larger than one for 23 GCMs. On the whole, the ability of GCMs in reproducing the winter monsoon is less variable and stronger than that for the summer monsoon.

Figure 11 illustrates that the normalized CRMSE of the only TAR GCM is greater than those of the AR4 and AR5 GCMs for the winter monsoon, and the only two TAR GCMs are worse than most AR4 and AR5 GCMs for the summer monsoon. Furthermore, 14 and 17 AR4 GCMs are compared to their successors in the AR5 for winter and summer monsoons, respectively (Table S5). Eight (six) AR5 GCMs perform better (worse) than their AR4 predecessors for the winter monsoon. In summer, ten (five) AR5 GCMs outperform (underperform) their AR4 predecessors, and the remaining two pairs of GCMs have comparable skills. Therefore, the AR5 GCMs are superior to their AR4 counterparts more often than vice versa for the East Asian monsoon. This is compatible with the idea that the AR5 GCMs simulate more realistic global monsoon climatology than the AR4 GCMs (Flato *et al.*, 2013), but differs from an obvious improvement from the AR4

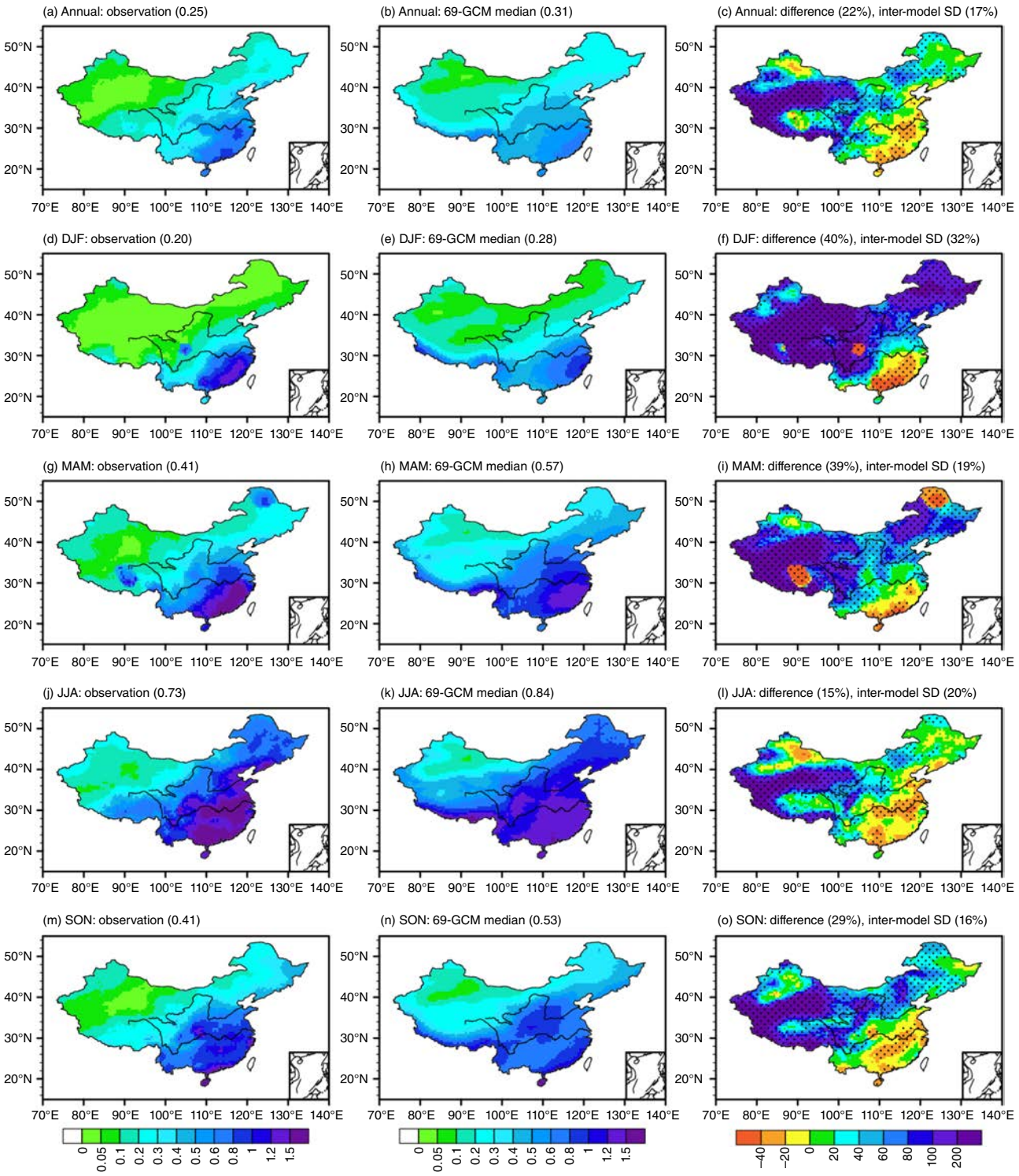

Figure 10. Interannual variability of annual and seasonal precipitation over China for the period 1961–2000 as obtained from observation (left column, units: mm day^{-1}), the median of the 69 reliable GCMs (middle column, units: mm day^{-1}), and the difference in percentage between the median and observation (right column). The regional average value in China and the inter-model standard deviation of the difference in percentage averaged over the country (right column, inter-model SD) are given in parentheses. The dotted areas in the right panels represent regions where at least 80% of the GCMs share the same sign of bias.

to AR5 in simulating the East Asian summer monsoon as obtained through a direct comparison, rather than the present pairwise comparison, of the whole atmospheric (Song and Zhou, 2014) or coupled (Sperber *et al.*, 2013) GCMs between the AR4 and AR5.

When viewed from the horizontal resolution, Figure 2(c) shows that the normalized CRMSEs statistically insignificantly grow with the grid area of GCMs for both winter and summer monsoons. This means the ability of GCMs in reproducing the East Asian monsoon is not systematically

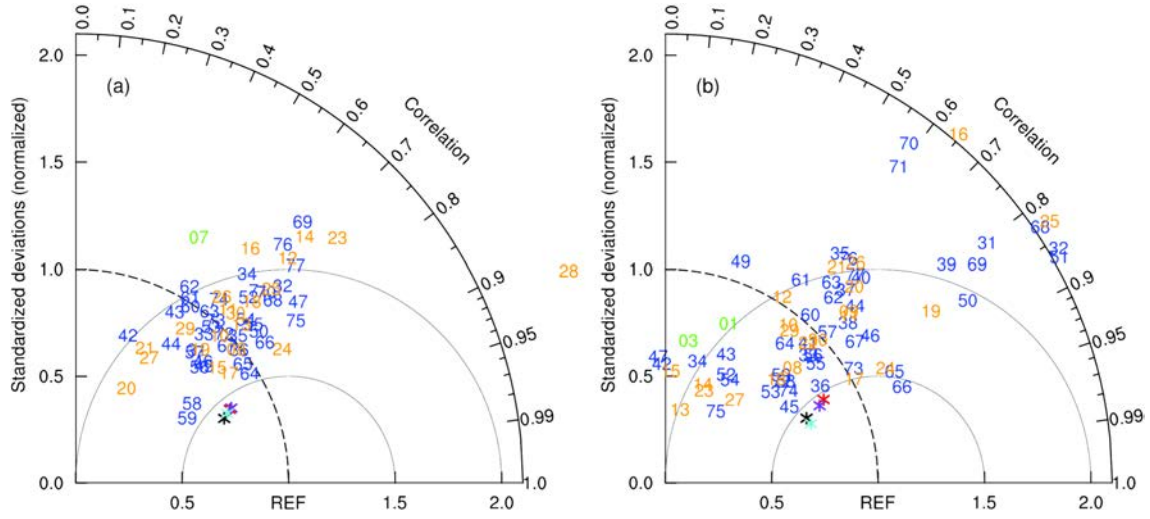

Figure 11. Taylor diagrams displaying normalized pattern statistics of climatological meridional winds at (a) 10 m within the regions of 25°–40°N and 120°–140°E, plus 10°–25°N and 110°–130°E between the 60 GCMs and observation in winter, and (b) at 850 hPa within the region of 20°–40°N and 105°–120°E between the 71 GCMs and observation in summer for the period 1979–2000. Red, purple, black, and cyan asterisks indicate the ensemble mean and the median of the 60 and 71 GCMs as well as the ensemble mean and the median of the 50 and 48 reliable GCMs, respectively. Other aspects are the same as Figure 1.

affected by model resolution. As for the algorithm to calculate the multi-GCM mean, the ensemble mean and the median of all GCMs analysed give the same SCC of 0.90 and normalized CRMSE of 0.44 for the winter monsoon, as well as the same SCC of 0.89 and normalized CRMSEs of 0.47 and 0.45, respectively, for the summer monsoon (Figure 11). Furthermore, a normalized CRMSE below 1.00 is set to identify GCMs reliably reproducing the East Asian monsoon. As a result, ten (23) GCMs are rejected for the winter (summer) monsoon. For the ensemble mean (the median) of the remaining GCMs, the SCC is 0.92 (0.91), and the normalized CRMSE is 0.43 (0.44) for the winter monsoon; while the SCC is 0.91 (0.93), and the normalized CRMSE is 0.45 (0.42) for the summer monsoon (Figure 11). Altogether, the algorithm to calculate the multi-GCM mean has little influence, as does the screening of individual GCMs according to their ability. Additionally, the multi-GCM mean outperforms all individual GCMs for the East Asian monsoon.

In winter, the observed surface winds in the northern target region are divided into two branches (Figure 12(a)). One turns eastward to the subtropical western North Pacific, and the other turns southwestward and flows along coastal East Asia and onto the South China Sea. These features are well reproduced by GCMs, as clearly seen for example in the median of the 50 reliable GCMs (Figure 12(b)). This is consistent with the recent analysis of 18 AR5 GCMs showing that the key tropospheric components of the East Asian winter monsoon can be reasonably simulated in terms of climatology (Gong *et al.*, 2014). Compared to the observation, however, GCMs generally underestimate the strength of the East Asian winter monsoon as manifested by southerly wind anomalies (Figure 12(c)). In summer, the southwest monsoon from the Bay of Bengal, the cross-equatorial air flow over South Asia, and the southeast monsoon from the western North Pacific converge and form the southwesterly monsoon winds in East Asia (Figure 12(d)). GCMs reliably reproduce these three large-scale components of the East Asian summer monsoon circulation in the lower troposphere (Figure 12(e)). However, the simulated subtropical high over the western North Pacific and the cross-equatorial flow are weaker than in the observation, and in turn northerly wind anomalies occur in southern East Asia and the adjacent ocean (Figure 12(f)). The summer monsoon circulation is therefore inadequate in southern East Asia. This is consistent with a recent evaluation of the Asian summer monsoon based on the AR4 and AR5 GCMs that suggests the northerly error over southern China and the South China Sea is related to a large bias in the simulation of the western North Pacific subtropical high, and is indicative of lower moisture content of air and reduced rainfall along the Meiyu rainfall front (Figure 8(l)) (Sperber *et al.*, 2013). In addition, when considering all 17 GCMs that reproduce both excessive summer precipitation averaged across eastern China (east of 105°E) and a stronger-than-observation East Asian summer monsoon, measured by regionally averaged meridional wind at 850 hPa within the region of 20°–40°N and 105°–120°E, the overestimated magnitude correlates statistically significantly, and positively with each other (Figure S4).

4. Conclusion

This study evaluates the performance of 77 IPCC TAR, AR4, and AR5 GCMs in simulating the mean state and year-to-year variability of climate over China and the East Asian monsoon for the late decades of the 20th century, with respect to observation and reanalysis data. The primary conclusions are as follows.

Apart from a reliable simulation of the geographical distribution of annual and seasonal temperatures

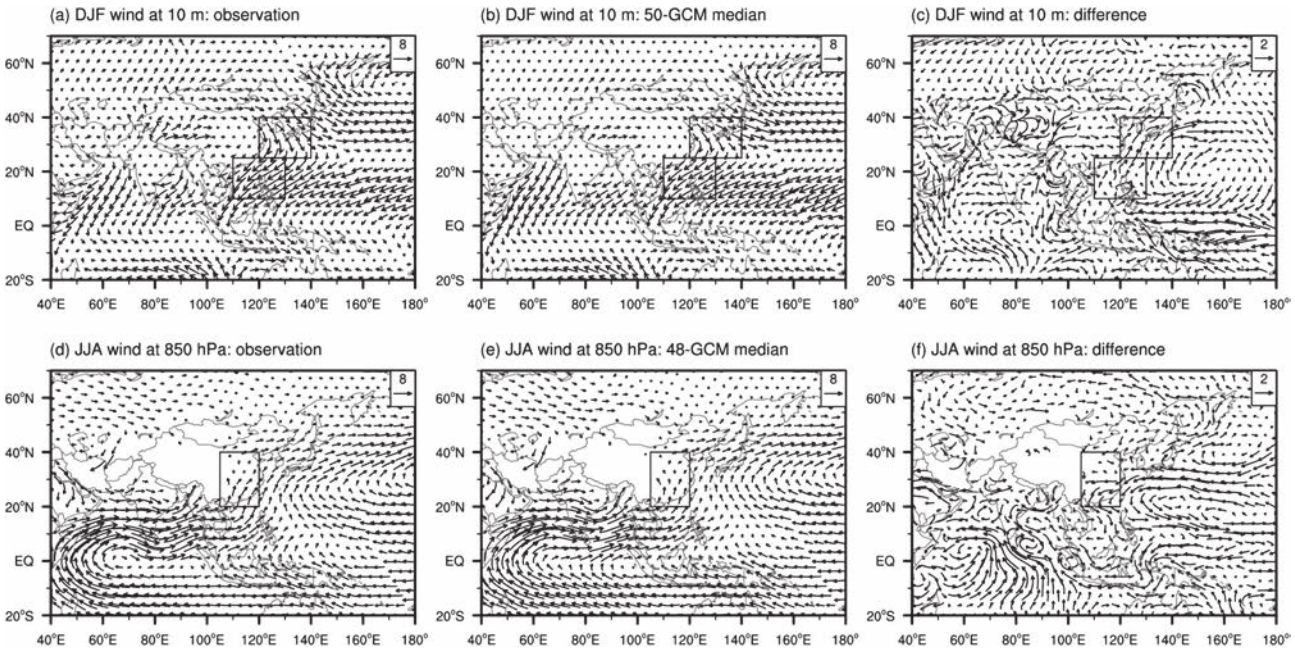

Figure 12. Climatological winter winds at 10 m (top row, units: m s^{-1}) and summer winds at 850 hPa (bottom row, units: m s^{-1}) for observation (left column), the median of the 50 and 48 reliable GCMs (middle column), and the difference between the median and observation (right column) for the period 1979–2000. The rectangles in the upper panels show the regions of 25°–40°N and 120°–140°E plus 10°–25°N and 110°–130°E, and the rectangles in the bottom panels shows the region of 20°–40°N and 105°–120°E.

and precipitation that has been reported before, it is revealed that most GCMs overestimate the spatial variability of annual and spring temperatures, and of precipitation (except in spring). Most GCMs still have cold biases, particularly in winter and autumn, which are obviously related to regional topography, but are much smaller than previously estimated, except in autumn. GCMs overestimate precipitation across most of China, and underestimate the southeast–northwest precipitation gradient over the country. The ability of GCMs generally decreases from spring via summer and autumn to winter for precipitation. GCMs continue to improve from the TAR via AR4 to AR5 for temperature, but there is little or no systematic improvement for precipitation. In general, the higher the horizontal resolution of GCMs is, the stronger their abilities become.

The ability of GCMs in reproducing the interannual variability of temperature (precipitation) is much weaker than for (similar to) its climatology. Most GCMs reliably reproduce the geographical distribution of interannual variability, but overestimate its magnitude and spatial variability. The skill of GCMs improves from the TAR to AR4, and remains stable from the AR4 to AR5 for temperature; while it improves from the TAR to AR4, and decreases from the AR4 to AR5 for precipitation. The ability of GCMs in reproducing the interannual variability of spring and summer temperatures and winter precipitation enhances when their resolutions increase.

The large-scale features of the East Asian monsoon can be well simulated by GCMs, although the strength of monsoon circulation is underestimated over East Asia in winter and over southern East Asia in summer. The ability of GCMs is less variable and stronger for the winter than that for the summer monsoon. The AR5 GCMs are superior to their AR4 counterparts more often than vice versa. The influence of model resolution is not perceptible. Finally, both the arithmetic mean and the median of multiple GCMs with and without filtering GCMs in terms of their ability have similar skills, and outperform most of individual GCMs in all the aspects considered.

Acknowledgements

We sincerely thank the two anonymous reviewers for their insightful comments. We also thank the climate modelling groups (listed in Table 1) for producing and sharing their model output. This research was supported by the National Basic Research Program of China (2012CB955401) and the National Natural Science Foundation of China (41421004 and 41375084).

Supporting Information

The following supporting information is available as part of the online article:

Figure S1. Interannual variability of annual and seasonal temperatures (units: °C) over China for the period 1961–2000 obtained from observation (left column), NorESM1-ME (Model ID: 77, middle column), and the difference between NorESM1-ME and observation (right column). The regional average value in China is given in parentheses.

Figure S2. Climatological annual and seasonal precipitation over China for the period 1961–2000 as obtained from observation (left column, units: mm day^{-1}), ACCESS1.3 (Model ID: 32, middle column, units: mm day^{-1}), and the difference in percentage between ACCESS1.3 and observation (right column, units: %). The regional average value in China is given in parentheses.

Figure S3. The coefficient of variation for the annual and seasonal precipitation over China for the period 1961–2000 as obtained from observation (left column), the median of the 56 reliable GCMs (middle column), and the difference between the median and observation (right column). The regional average value in China and the inter-model standard deviation of the difference averaged over the country (right column, inter-model SD) is given in parentheses. The dotted areas in the right panels represent regions where at least 80% of the GCMs share the same sign of bias.

Figure S4. Bias for winter wind at 10 m averaged within the regions of 25°–40°N and 120°–140°E plus 10°–25°N and 110°–130°E (blue) and summer wind at 850 hPa within the region of 20°–40°N and 105°–120°E (orange) versus precipitation averaged within the region 105°E east of China between 35 and 36 reliable GCMs (excluding TAR GCMs) and observation, respectively. Numbers indicate GCMs listed in Table 1. Lines represent the linear fitting for GCMs, with blue line for both positive bias in wind and precipitation for 27 GCMs, and orange lines for positive wind bias as well as positive (17 GCMs) and negative (19 GCMs) precipitation biases respectively. The fitting equations are shown in the figure, and the linear trend significant at the 99% confidence level is shown as **.

Table S1. Normalized CRMSEs for climatological annual and seasonal temperatures over China in the seven TAR GCMs and their high-resolution version successors in the AR4 and AR5, as well as in the 17 high-resolution AR4 GCMs and their successors in the AR5.

Table S2. Normalized CRMSEs for the interannual variability of annual and seasonal temperatures over China in the seven TAR GCMs and their high-resolution version successors in the AR4 and AR5, as well as in the 17 high-resolution AR4 GCMs and their successors in the AR5.

Table S3. Normalized CRMSEs for climatological annual and seasonal precipitation over China in the seven TAR GCMs and their high-resolution version successors in the AR4 and AR5, as well as in the 17 high-resolution AR4 GCMs and their successors in the AR5.

Table S4. Normalized CRMSEs for the interannual variability of annual and seasonal precipitation over China in the six TAR GCMs and their high-resolution version successors in the AR4 and AR5, as well as in the 17 high-resolution AR4 GCMs and their successors in the AR5.

Table S5. Normalized CRMSEs for climatological East Asian winter (EAWM) and summer (EASM) monsoons in the 14 and 17 high-resolution AR4 GCMs and their successors in the AR5, respectively.

References

Adam JC, Lettenmaier DP. 2003. Adjustment of global gridded precipitation for systematic bias. *J. Geophys. Res.* **108**(D9): 4257, doi: 10.1029/2002JD002499.

Chen L, Frauenfeld OW. 2014a. Surface air temperature changes over the twentieth and twenty-first centuries in China simulated by 20 CMIP5 models. *J. Clim.* **27**: 3920–3937, doi: 10.1175/JCLI-D-13-00465.1.

Chen L, Frauenfeld OW. 2014b. A comprehensive evaluation of precipitation simulations over China based on CMIP5 multimodel ensemble projections. *J. Geophys. Res.* **119**: 5767–5786, doi: 10.1002/2013JD021190.

Chen W, Graf H, Huang RH. 2000. The interannual variability of East Asian winter monsoon and its relation to the summer monsoon. *Adv. Atmos. Sci.* **17**: 48–60.

Chen HP, Sun JQ, Chen XL. 2014. Projection and uncertainty analysis of global precipitation-related extremes using CMIP5 models. *Int. J. Climatol.* **34**: 2730–2748, doi: 10.1002/joc.3871.

Ding YH. 1994. *Monsoons over China*. Springer: Dordrecht, the Netherlands, 1–419, doi: 10.1007/978-94-015-8302-2.

Flato G, Marotzke J, Abiodun B, Braconnot P, Chou SC, Collins W, Cox P, Driouech F, Emori S, Eyring V, Forest C, Gleckler P, Guilyardi E, Jakob C, Kattsov V, Reason C, Rummukainen M. 2013. Evaluation of climate models. In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, Stocker TF, Qin D, Plattner GK, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, Midgley PM (eds). Cambridge University Press: Cambridge, UK and New York, NY, 741–866, doi: 10.1017/CBO9781107415324.020.

Gao XJ, Xu Y, Zhao ZC, Pal JS, Giorgi F. 2006. On the role of resolution and topography in the simulation of East Asia precipitation. *Theor. Appl. Climatol.* **86**: 173–185, doi: 10.1007/s00704-005-0214-4.

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

Gong H, Wang L, Chen W, Wu R, Wei K, Cui X. 2014. The climatology and interannual variability of the East Asian winter monsoon in CMIP5 models. *J. Clim.* **27**: 1659–1678, doi: 10.1175/JCLI-D-13-00039.1.

Guo Y, Dong WJ, Ren FM, Zhao ZC, Huang JB. 2013. Surface air temperature simulations over China with CMIP5 and CMIP3. *Adv. Clim. Change Res.* **4**: 145–152, doi: 10.3724/SP.J.1248.2013.145.

Hua WJ, Chen HS, Sun SL. 2014. Uncertainty in land surface temperature simulation over China by CMIP3/CMIP5 models. *Theor. Appl. Climatol.* **117**: 463–474, doi: 10.1007/s00704-013-1020-z.

Huntingford C, Jones PD, Livina VN, Lenton TM, Cox PM. 2013. No increase in global temperature variability despite changing regional patterns. *Nature* **500**: 327–330, doi: 10.1038/nature12310.

Jiang D, Tian Z. 2013. East Asian monsoon change for the 21st century: results of CMIP3 and CMIP5 models. *Chin. Sci. Bull.* **58**: 1427–1435, doi: 10.1007/s11434-012-5533-0.

Jiang D, Wang HJ, Lang X. 2005. Evaluation of East Asian climatology as simulated by seven coupled models. *Adv. Atmos. Sci.* **22**: 479–495.

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Leetmaa A, Reynolds R, Jenne R, Joseph D. 1996. The NCEP/NCAR reanalysis project. *Bull. Am. Meteorol. Soc.* **77**: 437–472.

Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA. 2010. Challenges in combining projections from multiple climate models. *J. Clim.* **23**: 2739–2758, doi: 10.1175/2009JCLI3361.1.

Kumar D, Kodra E, Ganguly AR. 2014. Regional and seasonal intercomparison of CMIP3 and CMIP5 climate model ensembles for temperature and precipitation. *Clim. Dyn.* **43**: 2491–2518, doi: 10.1007/s00382-014-2070-3.

Lambert SJ, Boer GJ. 2001. CMIP1 evaluation and intercomparison of coupled climate models. *Clim. Dyn.* **17**: 83–106.

Lu RY, Fu YH. 2010. Intensification of East Asian summer rainfall interannual variability in the twenty-first century simulated by 12 CMIP3 coupled models. *J. Clim.* **23**: 3316–3331, doi: 10.1175/2009JCLI3130.1.

Mueller B, Seneviratne SI. 2014. Systematic land climate and evapotranspiration biases in CMIP5 simulations. *Geophys. Res. Lett.* **41**: 128–134, doi: 10.1002/2013GL058055.

Perez J, Menendez M, Mendez FJ, Losada IJ. 2014. Evaluating the performance of CMIP3 and CMIP5 global climate models over the

north-east Atlantic region. *Clim. Dyn.* **43**: 2663–2680, doi: 10.1007/s00382-014-2078-8.

Pierce DW, Barnett TP, Santer BD, Gleckler PJ. 2009. Selecting global climate models for regional climate change studies. *Proc. Natl. Acad. Sci.* **106**: 8441–8446.

Räisänen J. 2007. How reliable are climate models? *Tellus* **59A**: 2–29, doi: 10.1111/j.1600-0870.2006.00211.x.

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

Schär C, Vidale PL, Lüthi D, Frei C, Häberli C, Liniger MA, Appenzeller C. 2004. The role of increasing temperature variability in European summer heatwaves. *Nature* **427**: 332–336, doi: 10.1038/nature02300.

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

Sperber KR, Annamalai H, Kang IS, Kitoh A, Moise A, Turner A, Wang B, Zhou T. 2013. The Asian summer monsoon: an intercomparison of CMIP5 vs. CMIP3 simulations of the late 20th century. *Clim. Dyn.* **41**: 2711–2744, doi: 10.1007/s00382-012-1607-6.

Su FG, Duan XL, Chen DL, Hao ZC, Cuo L. 2013. Evaluation of the global climate models in the CMIP5 over the Tibetan Plateau. *J. Clim.* **26**: 3187–3208, doi: 10.1175/JCLI-D-12-00321.1.

Sui Y, Jiang D, Tian Z. 2013. Latest update of the climatology and changes in the seasonal distribution of precipitation over China. *Theor. Appl. Climatol.* **113**: 599–610, doi: 10.1007/s00704-012-0810-z.

Sui Y, Lang X, Jiang D. 2014. Time of emergence of climate signals over China under the RCP4.5 scenario. *Clim. Change* **125**: 265–276, doi: 10.1007/s10584-014-1151-y.

Taylor KE. 2001. Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.* **106**(D7): 7183–7192.

Taylor KE, Stouffer RJ, Meehl GA. 2012. An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.* **93**: 485–498, doi: 10.1175/BAMS-D-11-00094.1.

Thornton PK, Ericksen PJ, Herrero M, Challinor AJ. 2014. Climate variability and vulnerability to climate change: a review. *Glob. Change Biol.* **20**: 3313–3328, doi: 10.1111/gcb.12581.

Wang SY, Xiong Z. 2004. The preliminary analysis of 5 coupled ocean–atmosphere global climate models simulations of regional climate in Asia. *Clim. Environ. Res.* **9**: 240–250.

Wu J, Gao XJ. 2013. A gridded daily observation dataset over China region and comparison with the other datasets. *Chin. J. Geophys.* **56**: 1102–1111, doi: 10.6038/cjg20130406.

Xu Y, Xu CH. 2012. Preliminary assessment of simulations of climate changes over China by CMIP5 multi-models. *Atmos.Oceanic Sci. Lett.* **5**: 489–494.

Xu CH, Shen XY, Xu Y. 2007. An analysis of climate change in East Asia by using the IPCC AR4 simulations. *Adv. Clim. Change Res.* **3**: 287–292.

Zhou BT, Wen QH, Xu Y, Song LC, Zhang XB. 2014. Projected changes in temperature and precipitation extremes in China by the CMIP5 multimodel ensembles. *J. Clim.* **27**: 6591–6611, doi: 10.1175/JCLI-D-13-00761.1.

