Clim Dyn (2015) 44:2947–2963
DOI 10.1007/s00382-015-2625-y

Non-stationary return levels of CMIP5 multi-model temperature extremes

Linyin Cheng[1] · **Thomas J. Phillips**[2] · **Amir AghaKouchak**[1]

Received: 18 January 2014 / Accepted: 16 April 2015 / Published online: 1 May 2015

Abstract The objective of this study is to evaluate to what extent the CMIP5 climate model simulations of the climate of the twentieth century can represent observed warm monthly temperature extremes under a changing environment. The biases and spatial patterns of 2-, 10-, 25-, 50- and 100-year return levels of the annual maxima of monthly mean temperature (hereafter, annual temperature maxima) from CMIP5 simulations are compared with those of Climatic Research Unit (CRU) observational data considered under a non-stationary assumption. The results show that CMIP5 climate models collectively underestimate the mean annual maxima over arid and semi-arid regions that are most subject to severe heat waves and droughts. Furthermore, the results indicate that most climate models tend to underestimate the historical annual temperature maxima over the United States and Greenland, while generally disagreeing in their simulations over cold regions. Return level analysis shows that with respect to the spatial patterns of the annual temperature maxima, there are good agreements between the CRU observations and most CMIP5 simulations. However, the magnitudes of the simulated annual temperature maxima differ substantially across individual models. Discrepancies are generally larger over higher latitudes and cold regions.

✉ Amir AghaKouchak
amir.a@uci.edu

Linyin Cheng
linyinc@uci.edu

[1] University of California, Irvine, E4130 Engineering Gateway, Irvine, CA 92697-2175, USA

[2] Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, CA 94550, USA

Keywords Temperature · Climate · CMIP5 · Extremes · Return level · Non-stationary

1 Introduction

During the period 1979–1992, on average nearly 400 people each year were killed in the United States by excessive heat (NOAA 1995; Kilbourne 1997). In fact, in this period over the United States, excessive heat accounted for more reported deaths annually than hurricanes, floods, tornadoes, and lightning combined (NOAA 1995). Furthermore, agriculture products such as wheat, rice, corn and maize can be significantly reduced by extreme high temperatures at key development stages (NOAA 1980; Hoerling et al. 2013). High temperatures also affect irrigation and evaporation (Sorooshian et al. 2014), drought development (Aghakouchak et al. 2014), energy production and consumption (Tarroja et al. 2014b) as well as greenhouse gas emissions associated with energy production (Tarroja et al. 2014a). Numerous studies indicate that temperature extremes are likely to intensify in the future under different plausible climate scenarios (Alexander et al. 2006; IPCC 2007).

Climate model simulations have been widely used to study extreme weather and climate across different spatial and temporal scales. Recently, international climate modeling groups have provided Coupled Model Intercomparison Project Phase 5 (CMIP5) historical and projected climate simulations (Taylor et al. 2012). The scope of CMIP5 also is broader than previous model intercomparison projects (e.g., CMIP3), with carbon emission-driven Earth system model (ESM) experiments now represented along with the typical concentration-driven atmosphere–ocean general circulation model (AOGCM) simulations (Meehl and Bony 2011). Thus, the multi-model gridded CMIP5

datasets provide an unprecedented opportunity to analyze and assess climate variability and change. However, model simulations are known to have high uncertainty (Kharin et al. 2013; Mehran et al. 2014; Liu et al. 2014; Sillmann et al. 2013), and different methods have been developed to assess their consistency against observations (Gleckler et al. 2008; AghaKouchak and Mehran 2013; Phillips and Gleckler 2006).

In a recent study, Kharin et al. (2013) argued that the global warm temperature extremes in the late twentieth century climate are reasonably simulated by the CMIP5 models (differences in CMIP5 models and reanalysis data were within a few degrees °C over most of the globe). Furthermore, the study showed that the spread amongst CMIP5 models for several temperature indices defined by the Expert Team on Climate Change Detection and Indices (ETCCDI; Zhang and Zwiers 2013) is reduced compared to CMIP3 models (see also Sillmann et al. 2013). However, the inter-model differences of warm temperature extremes are generally large over land with a standard deviation of around 4 °C (Kharin et al. 2013). Kharin et al. (2013) concluded that upward trends of warm extremes exceed those of cold extremes over tropical and subtropical land regions. Morak et al. (2013) showed that there is a significant increase in the trend in warm temperature extremes during both boreal cold and warm seasons over the second half of the twentieth century. Using CMIP5 simulations and observations, Hao et al. (2013) demonstrated that concurrent warm/dry and warm/wet extremes have increased substantially in the second half of the twentieth century.

Return periods and return levels (also known as return values) are often used to describe and assess risk of extremes (Cooley et al. 2007; AghaKouchak and Nasrollahi 2010; Katz 2010; Cooley 2013; Serinaldi 2014). In theory, the return period (T) of an event is the inverse of its probability of occurrence in any given year. That is, the n-year return level corresponds to an exceedance probability (by an annual extreme) of $1/n$. In the statistical literature, there are different definitions for return period and return level; for alternative definitions, the interested reader is referred to Bonnin et al. (2004), Mays (2010) and AghaKouchak et al. (2013).

In recent years, extreme value theory (EVT) has been widely used for analysis of climate extremes and their return levels (Zwiers and Kharin 1998; Clarke 2002; Katz et al. 2002; Kharin and Zwiers 2005; Parey et al. 2010; Kunkel 2013; Serinaldi and Kilsby 2015; Cheng et al. 2014b; Cooley 2013). Fisher and Tippett (1928) introduced the concept of asymptotic theory in extreme value distributions and laid the foundation for a generalized approach to extreme value analysis. Gnedenko (1943) mathematically proved that three families of extreme value distributions—namely Weibull, Gumbel and Fréchet—can represent the limiting distributions of extremes in random variables. The generalized extreme value (GEV) distribution is essentially a combination of these three distribution families, and has been applied in a variety of studies (Gumbel 1942; Smith 2001; Katz 2013).

Numerous studies indicate that climatic extremes (e.g., hot days, heavy precipitation) have increased significantly, particularly in the second half of the twentieth century (Karl and Knight 1998; Easterling et al. 2000; Vose et al. 2005; Hansen et al. 2010; Villarini et al. 2011; Hao et al. 2013; Field et al. 2012; Wehner 2013). In addition to their number, the frequency of extremes has changed in the past, and is likely to change in the future (Milly et al. 2008; Easterling et al. 2000; IPCC 2007). It is evident that ignoring time-varying (non-stationary) behavior of extremes could lead to underestimation of extremes and considerable damage to human life and society (McMichael et al. 2003; Cheng and AghaKouchak 2014). Therefore, it is necessary to assess non-stationarity in the CMIP5 climate models simulations, and to document the extent to which the model-simulated patterns are consistent with observations.

In most previous climate model evaluation studies, the first or second order statistics of models are compared against observations. The primary objective of this paper is to evaluate to what extent the CMIP5 model simulations of the historical climate of the period 1901–2005 can represent observed warm monthly temperature extremes under the non-stationary assumption. In this study, the GEV distribution is used to investigate the return levels of annual monthly temperature maxima considering a changing climate. The return levels of temperature maxima estimated from the CMIP5 climate simulations are compared with those of Climatic Research Unit (CRU) temperature observations.

The remainder of the paper is organized as follows. In Sect. 2, the data sets and study area are discussed. Section 3 presents the methodology for non-stationary extreme value analysis. The results including representation of annual maxima in their return levels are provided in Sect. 4. Section 5 summarizes the main results and offers concluding remarks.

2 Study area and data resources

Monthly observations of temperature provided by the Climatic Research Unit (CRU, New et al. 2000; Mitchell and Jones 2005), available in a 0.5° spatial resolution, are used as reference data. In this study, 41 CMIP5 historical monthly temperature simulations from 1901 to 2005 are subjected to extreme value analysis, and a subset of 17 of these simulations are investigated in more detail. These data sets represent the most extensive and

Table 1 List of 17 climate models/experiments whose simulations are displayed in Figs. 1, 2, 3, 4, 5, 6 and 7 and their related institutions and countries

Model/experiment	Institutions	Countries
BCC-CSM1-1_esm	Beijing Climate Center, China Meteorological Administration	China
MIROC-ESM	Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute, The University of Tokyo	Japan
	National Institute for Environmental Studies	
NorESM1-M	Norwegian Climate Centre	Norway
IPSL-CM5A-LR	Institut Pierre-Simon Laplace	France
GFDL-CM3	NOAA Geophysical Fluid Dynamics Laboratory	USA
CCSM4	National Center for Atmospheric Research	USA
GISS-E2-H	NASA Goddard Institute for Space Studies	USA
INMCM4_esm	Institute for Numerical Mathematics	Russia
HadGEM2-ES_esm	Met Office Hadley Centre	UK
CSIRO-ACCESS1-0	Commonwealth Scientific and Industrial Research Organisation, and Bureau of Meteorology	Australia
MRI-ESM1_esm	Meteorological Research Institute	Japan
MPI-ESM-P	Max Planck Institute for Meteorology	Germany
CanESM2	Canadian Centre for Climate Modelling and Analysis	Canada
FGOALS-g2	Institute of Atmospheric Physics	China
	Chinese Academy of Sciences and Tsinghua University	
CESM1-CAM5	National Science Foundation, Department of Energy, and National Center for Atmospheric Research	USA
CNRM-CM5	Centre National de Recherches Meteorologiques	France
	Centre Europeen de Recherche et Formation Avancees en Calcul Scientifique	
CESM1-WACCM	National Science Foundation, Department of Energy, and National Center for Atmospheric Research	USA

The suffix "_esm" designates an historical climate run of an ESM with prescribed emissions

ambitious multi-model simulations that contribute to the World Climate Research Programme's CMIP multi-model dataset (Meehl and Bony 2011; Taylor et al. 2012). For this extreme value analysis, the CMIP5 model simulations and CRU observations all are regridded to a common 2 × 2-degree resolution. This regridding entailed use of bilinear interpolation, with special attention given to appropriate use of model-specific land fraction masks so as to minimize data distortions along coastlines. The selected models (Table 1) include physical climate models (without a prognostic global carbon cycle), as well as ESMs (with the designation "ESM" appearing in the model title). The former are run in a standard "historical climate" configuration with prescribed historically increasing CO_2 concentrations and the latter are run with CO_2 emissions (fluxes corresponding to the prescribed historically increasing CO_2 concentrations—designated as "_esm" experiments in Table 1).

This study focuses on global land areas (excluding Antarctica) for which the CRU observations are provided. From CRU observations and CMIP5 simulations, pixel-based annual monthly temperature maxima (hereafter, annual temperature maxima) are extracted for estimation of extreme temperature return levels. It should be noted that the Hadley Centre has adopted an unconventional time model for all their CMIP5 output data, with an endpoint in November rather than December of 2005, and thus the HadGEM2 outputs include 1 month fewer data than those of the other models. This issue will not affect the results in the Northern Hemisphere since the annual maxima of monthly data do not often occur in December. However, it might slightly impact the analyses in the Southern Hemisphere.

3 Methodology

The cumulative distribution function (CDF) of the GEV can be written as (Coles 2001):

$$\Psi(x) = exp\left\{-\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{\frac{-1}{\xi}}\right\} \qquad \left(1+\xi\left(\frac{x-\mu}{\sigma}\right)>0\right) \tag{1}$$

where $\Psi(x)$ refers to the limiting CDF of the block maxima of independently and identically distributed random variables (Leadbetter et al. 1983). To avoid serial dependence

in the data, in this study, only one value per year (annual maxima) is sampled for extreme value analysis.

The GEV distribution models different characteristics of extremes using three parameters $\theta = (\mu, \sigma, \xi)$: location parameter ($\mu$), scale parameter ($\sigma$), and shape parameter ($\xi$). The location parameter indicates where the extremes distribution is centered, whereas the scale parameter specifies the deviations around μ. The shape parameter describes the tail behavior of the GEV distribution such that $\xi = 0$, $\xi < 0$ and $\xi > 0$ represent the Gumbel, Weibull and Fréchet families, respectively (Coles 2001). The GEV CDF ($\Psi(x)$) is defined for $1 + \xi(\frac{x-\mu}{\sigma}) > 0$; elsewhere, $\Psi(x)$ is either 0 or 1 (Smith 2001).

The stationary form of the GEV has been studied and applied in numerous studies (Leadbetter et al. 1983; Coles 2001; Schlather 2002; Li et al. 2005; Reiss and Thomas 2007; Papalexiou and Koutsoyiannis 2013). Under stationarity, the model parameters are time invariant (Renard et al. 2013). In practice, many environmental time series vary systematically in response to climatic change, and hence may exhibit non-stationary behavior (Zwiers and Kharin 1998; Solomon et al. 2007; Rootzén and Katz 2013). In a non-stationary world, the GEV parameters are time-dependent, and thus the extremal properties of the GEV would vary with time (Meehl et al. 2000; Cheng et al. 2014a). In most studies, the non-stationarity is accounted for by assuming the location parameter is a linear function of time, while keeping the scale and shape parameters constant (Gilleland and Katz 2006; Cooley 2009; Katz 2010; Renard et al. 2013):

$$\mu(t) = \mu_1 t + \mu_0 \tag{2}$$

where (μ_1, μ_0) are regression parameters and t is the time. Technically, a more complex nonlinear model can also be used for non-stationary analysis. However, a non-linear model would be more sensitive to the length of the record, and cannot be justified without an evidence of a nonlinear trend. Presence of a nonlinear trend can be assessed using a wide variety of methods (e.g., Phillips 2007; Morton and Henderson 2008; Granger et al. 1997). Furthermore, including a trend in the scale parameter and/or shape parameter is straightforward. However, given a lack of evidence that they have changed substantially in the past across the globe, we have not included them in this paper. Given that the objective of this study is to evaluate return levels of the historical CMIP5 model simulations with ground-based observations, the most stable form of non-stationary extreme value analysis is used where the location parameter changes over time.

Considering a linear trend ($\mu(t) = \mu_1 t + \mu_0$) imposes a stringent criteria of a certain type of non-stationarity on the modeled outputs. For this reason, outputs should be actively checked through statistical model selection criteria such as the Bayes factor (discussed below). Imposition of a linear model means that when a linear trend is found to be statistically significant, then the model is expected to simulate outputs with a similar trend. When analyzing historical data, which is the case in this study, presence of trends and their significance can be easily evaluated. However, this assumption may not be valid for future simulations, especially beyond a certain point since the observed historical trend may not continue into the future.

In this study, a Bayesian framework is employed to estimate the GEV parameters. This Bayesian framework integrates the information brought by a prior distribution $p(\theta)$ and the observation vector $\overrightarrow{y}$ into the posterior distribution of the GEV parameters. For non-stationary GEV parameter estimation, the Bayes formula can be written as (Winkler 1973; Renard et al. 2006; Cheng et al. 2014a, b):

$$p(\theta|\overrightarrow{x}) \propto p(\overrightarrow{x}|\theta)p(\theta) = \prod_{t=1}^{N_t} p(x_t|\theta)p(\theta) \tag{3}$$

$$\begin{aligned} &p(\beta|\overrightarrow{x}, y) \propto p(\overrightarrow{x}|\beta, y)p(\beta|y) \\ &p(\overrightarrow{x}|\beta, y) = \prod_{t=1}^{N_t} p(x_t|\beta, y(t)) = \prod_{t=1}^{N_t} p(x_t|\mu(t), \sigma, \xi) \end{aligned} \tag{4}$$

where $\beta = (\mu_1, \mu_0, \sigma, \xi)$, and $y(t)$ is the covariate value under non-stationarity. The parameters of the GEV distribution including $\mu_0, \mu_1, \sigma, \xi$ are derived by generating random realizations from the posterior distributions of the model parameters using the differential evolution Markov Chain (DE-MC; ter Braak 2004; Vrugt et al. 2009). In fact, the DE-MC, with a Metropolis–Hastings (Renard et al. 2013) step to update each parameter, is used to derive a sample from the posterior distributions of the GEV model parameters whose initial states are determined by a maximum likelihood method (Coles 2001). A statistical method known as the criterion $\widehat{R}$ (Gelman and Shirley 2011) is adopted for convergence checking. A detailed explanation of the parameter estimation approach used in this study is provided in Renard et al. (2013). Compared to the conventional Markov Chain Monte Carlo (MCMC) algorithm, the DE-MC exhibits faster calculation and convergence speed (ter Braak 2006). In this approach, the parameters are estimated based on the posterior of the MCMC samples for each pixel separately.

The fitted GEV distribution is then used to derive the return levels of annual temperature maxima of CMIP5 climate simulations and CRU reference observations. The median of the estimated location parameter, obtained from the DE-MC procedure, is used as the final location parameter:

$$\widetilde{\mu}_m = median(\mu(1), \mu(2), \ldots, \mu(t)) \tag{5}$$

While here the median of the DE-MC samples is selected from simulations, more conservative quantiles

(90th, 95th percentile) can also be used, depending on the objective of the study (Cheng et al. 2014a).

It should be noted that the DE-MC Bayesian approach provides posterior distributions of the model parameters. For this reason, upper and lower bounds (or other quantiles) of uncertainty can be derived using the upper and lower bounds of the parameters' posterior distributions.

In this study, *return levels* in model simulations and observations are evaluated. The return level is defined as a function of the return period T explained above (Cooley 2013):

$$T = \frac{1}{1-p} \tag{6}$$

where p is the non-exceedance occurrence probability. The non-exceedance probability can be obtained using either a non-parametric approach (Makkonen 2006) or a parametric distribution function such as the GEV. Here, the return levels of temperature extremes are computed for five return periods: $T = 2, T = 10, T = 25, T = 50$ and $T = 100$.

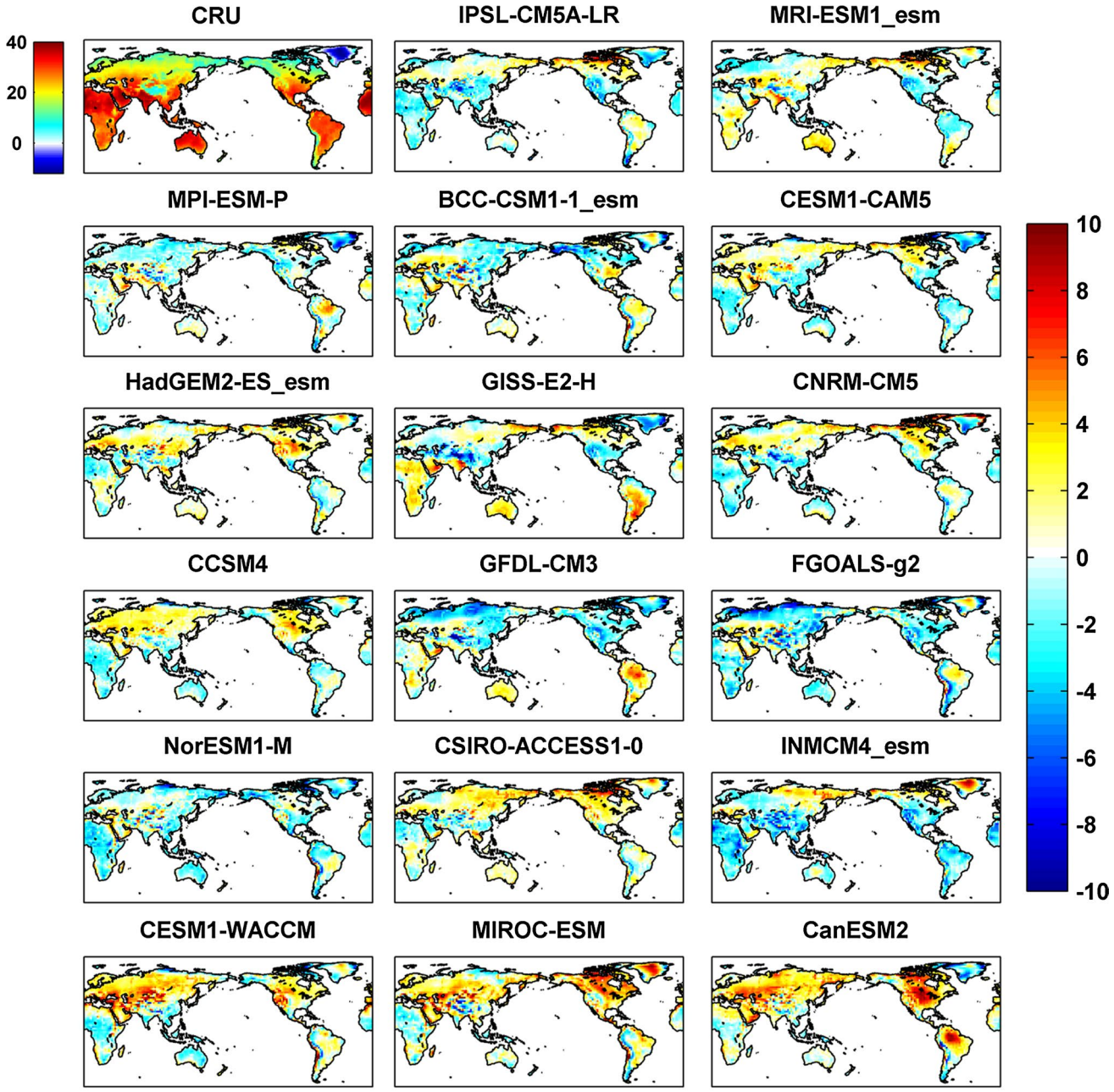

Fig. 1 Mean annual temperature maxima (in °C) based on 1901–2005 Climatic Research Unit (CRU) observations (*upper left panels*), and the differences between selected CMIP5 climate simulations and CRU reference data (*remaining panels*)

The presence of non-stationarity is tested by means of the commonly used Mann–Kendall trend test (Mann 1945; Kendall 1976) at the $\alpha = 0.05$ significance level. If the null hypothesis of no trend cannot be rejected, we assume a stationary climate. The stationary GEV-based p-return level (p-quantile curve, q_p) can be derived as:

$$q_p = \left(\left(-\frac{1}{\log p} \right)^{\xi} - 1 \right) \times \frac{\sigma}{\xi} + \mu, \; for \quad \xi \neq 0 \tag{7}$$

If the null hypothesis of no trend is rejected (data exhibits non-stationary behavior), we assume a non-stationary condition with respect to temperature extremes. The non-stationary GEV-based return level is then estimated as:

$$q_p = \left(\left(-\frac{1}{\log \widetilde{p}} \right)^{\xi} - 1 \right) \times \frac{\sigma}{\xi} + \widetilde{\mu}_m, \; for \quad \xi \neq 0 \tag{8}$$

where $\widetilde{\mu}_m$ is defined in Eq. 5 and $\widetilde{p}$ is the corresponding probability. It should be noted that the Mann–Kendall trend test is only used to avoid implementing a non-stationary model on a time series that does not exhibit a statistically significant change in extremes. However, the methodology is general and can be applied to the data, regardless of the underlying trend. To further evaluate the fit of the non-stationary model, the Bayes factor (Kass and Raftery 1995) is implemented to check the null hypothesis of no trend. The Bayes factor evaluates the null-hypothesis of no trend against the alternative using the posterior distributions of sampled parameters (Kass and Raftery 1995).

4 Results

4.1 Representation of annual temperature maxima

In the first step, the annual maxima of CMIP5 temperature simulations, determined from monthly time series, are compared with those of the CRU observations. Figure 1 (top left) displays the mean annual temperature maxima from 1901 to 2005 as represented by CRU observations. The rest of the panels in Fig. 1 demonstrate the differences in the mean annual temperature maxima between CMIP5 climate simulations and CRU observations (CMIP5 model-observation). In this figure, positive (negative) values indicate overestimation (underestimation) of the annual temperature maxima. Figure 1 shows the results for the 17 CMIP5 models listed in Table 1. One can see that the climate models individually display different patterns of overestimation and underestimation. The discrepancies between the model simulations and observations are primarily within ±1–3 °C. However, local errors for some regions may be as high as ±10 °C (see also the empirical cumulative distribution of the mean error in Fig. 2).

Figure 1 shows that over the United States most models (excepting HadGEM2-ES_esm, CCSM4, CSIRO-ACCESS1-0, CESM1-WACCM, MIROC-ESM and CanESM2) tend to underestimate the mean annual temperature maxima by 1–3 °C. Here, CanESM2 instead substantially overestimates the mean annual temperature maxima. Over Australia, on the other hand, several models (e.g. CSIRO-ACCESS1-0, HadGEM2-ES_esm, MPI-ESM-P and CanESM2) demonstrate little or no bias.

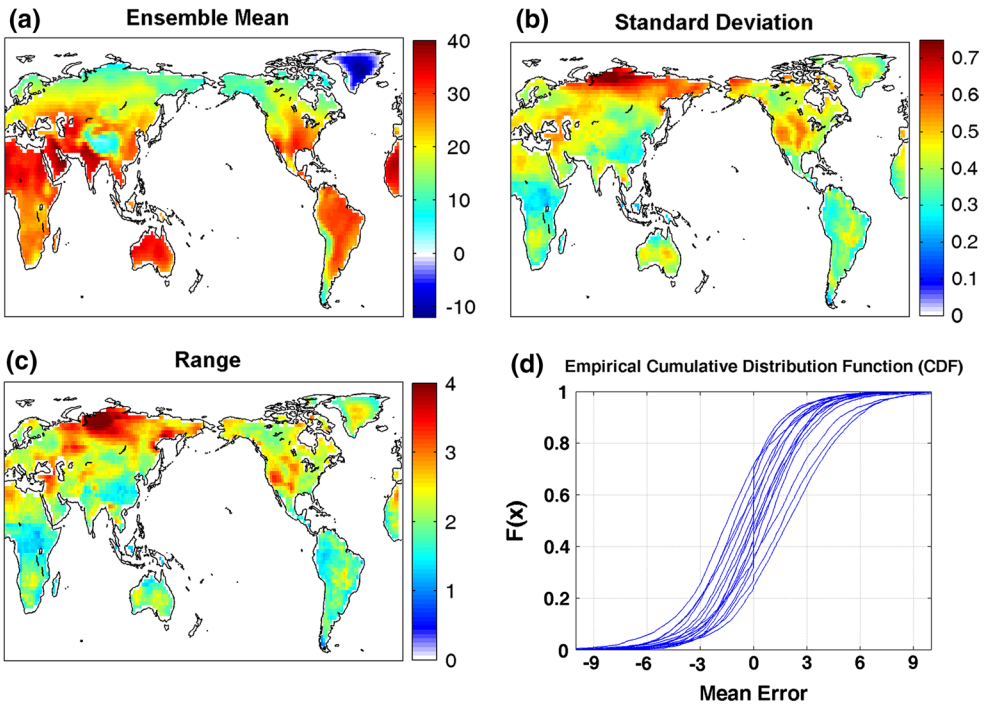

Fig. 2 **a** Ensemble mean, **b** inter-model standard deviation, and **c** range of the annual temperature maxima in CMIP5 simulations. **d** The empirical cumulative distribution (CDF) of the mean error of simulations relative to observations

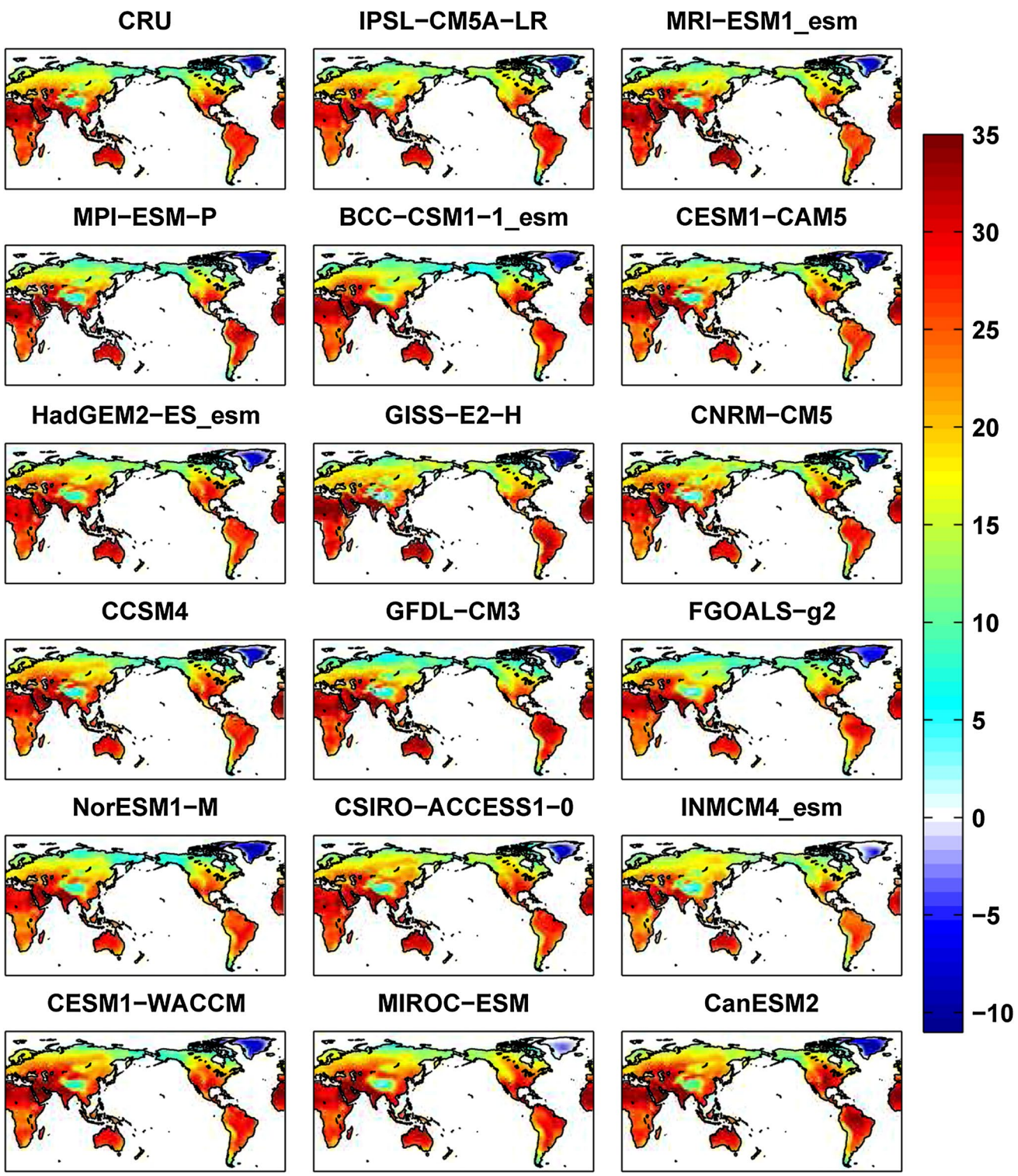

Fig. 3 2-Year return level (in °C) of the annual temperature maxima based on the CRU observations (*upper left panel*), and on selected CMIP5 climate model simulations (*remaining panels*)

Over Amazonia, the mean annual temperature maxima are mostly underestimated, except in a few models (e.g., GFDL-CM3, CanESM2) where they are overestimated.

The results indicate that model simulations particularly diverge from one another over cold regions (e.g., northern Russia, Canada) except for Greenland, where most

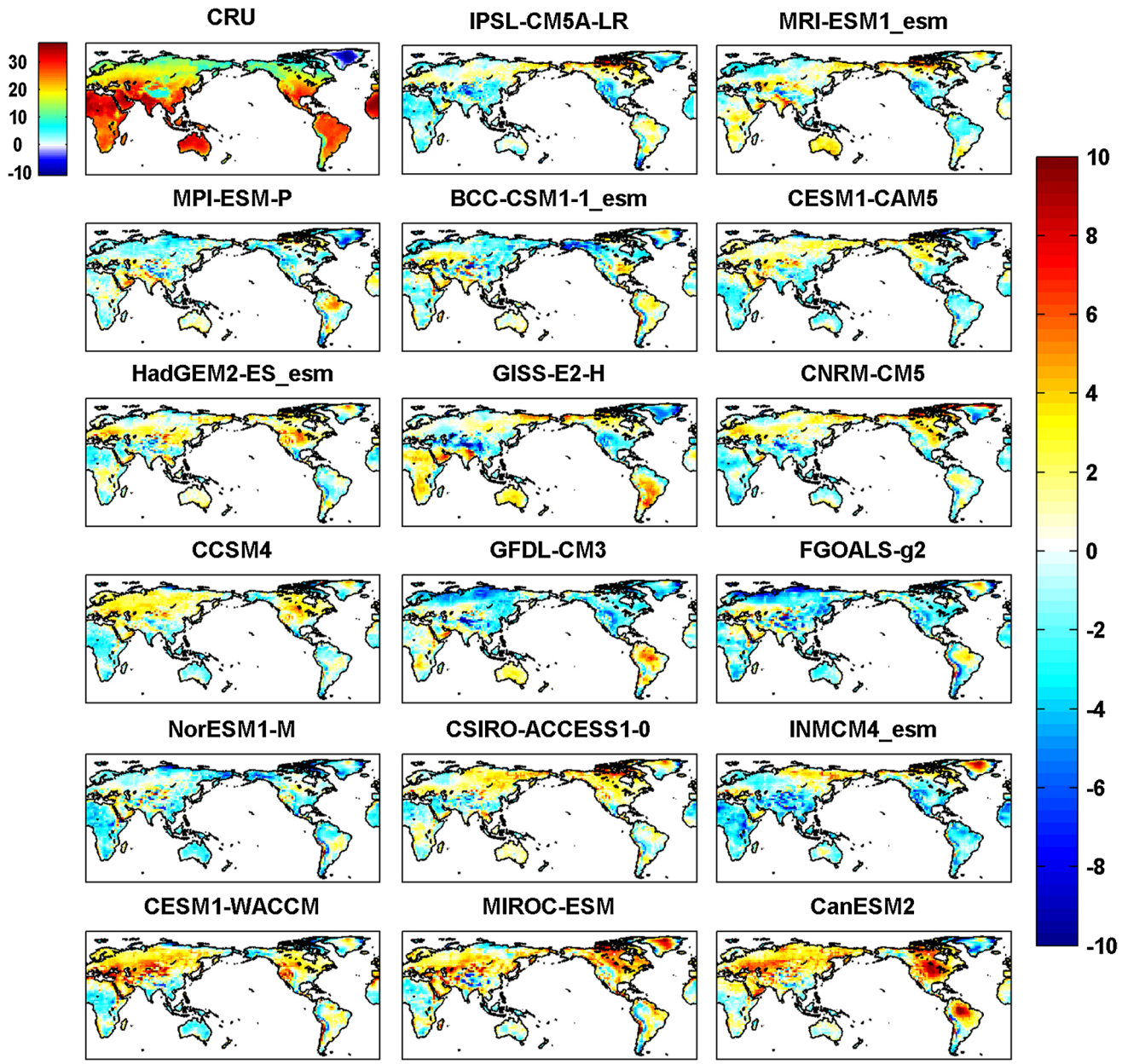

Fig. 4 2-Year return level (in °C) of the annual temperature maxima based on the CRU observations (*upper left panel*), and the 2-year return-level differences between selected CMIP5 climate simulations and CRU reference data (CMIP5–CRU; *remaining panels*)

models (but not MIROC-ESM and INMCM4_esm) underestimate the mean annual temperature maxima. Such a consistent underestimation could substantially impact model-based analyses of changes in ice-sheets, and snow/glacier melt studies. Krabill et al. (2004) reported that Greenland is losing coastal ice sheets quite rapidly (see also Ren et al. 2011; Kjær et al. 2012). CMIP5 models' underestimation of annual maxima climatology implies that the ice loss rate in Greenland might be greater than that reported in model-based studies. Similar to the modeling results by Alley et al. (2005) and Reeh (1989), rapid ice-marginal changes may indicate greater ice-sheet sensitivity to warming than has been acknowledged previously. However, over other cold regions that are at most risk of accelerated ice melt (e.g. Alaska, Northern Canada, and Siberia), most models tend to overestimate the mean annual temperature maxima relative to the CRU reference data (Fig. 1).

It is also noteworthy that the model simulations collectively underestimate the mean annual maxima over arid

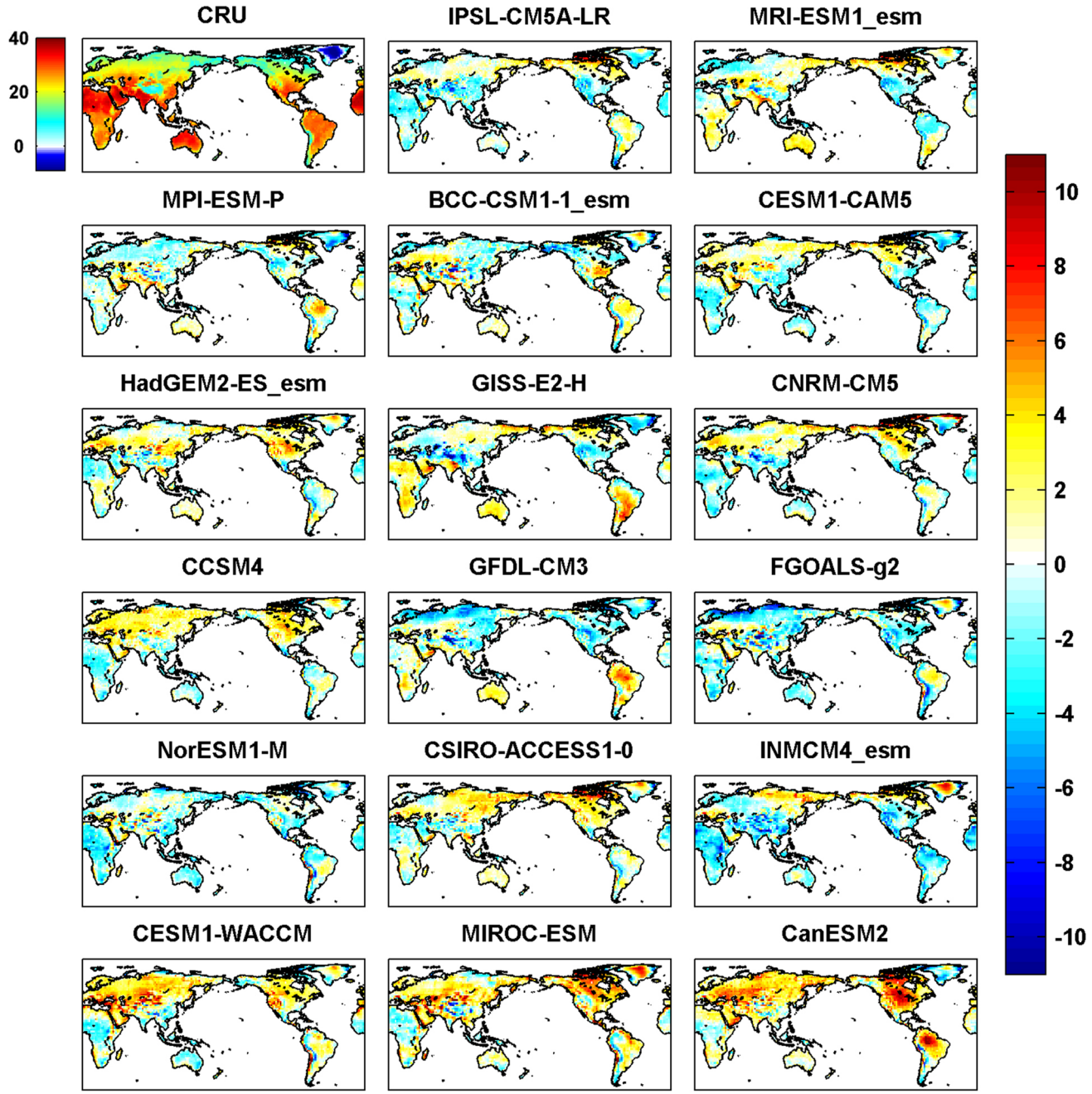

Fig. 5 25-Year return level (in °C) of the annual temperature maxima from CRU observations (*upper left panel*), and the 25-year return-level differences between selected CMIP5 climate simulations and CRU reference data (CMIP5–CRU; *remaining panels*)

and semi-arid regions (e.g., Sahara, southwestern US), that are most subject to severe heat waves. Considering the magnitudes of deviations from the CRU, there is a better agreement between CMIP5 simulations and observations in such subtropical regions than in high-latitude cold regions. This is consistent with the findings reported in Kharin et al. (2007) based on CMIP3 climate model simulations.

Figure 2 displays the ensemble mean (a), inter-model standard deviation (b), and range (c) of the annual temperature maxima in CMIP5 simulations, as well as the empirical CDF of the mean error relative to observations (d). The figure shows that the inter-model variability and range of simulations are greater over Siberia, the western United States, and parts of the Middle East and Sahara compared to other regions.

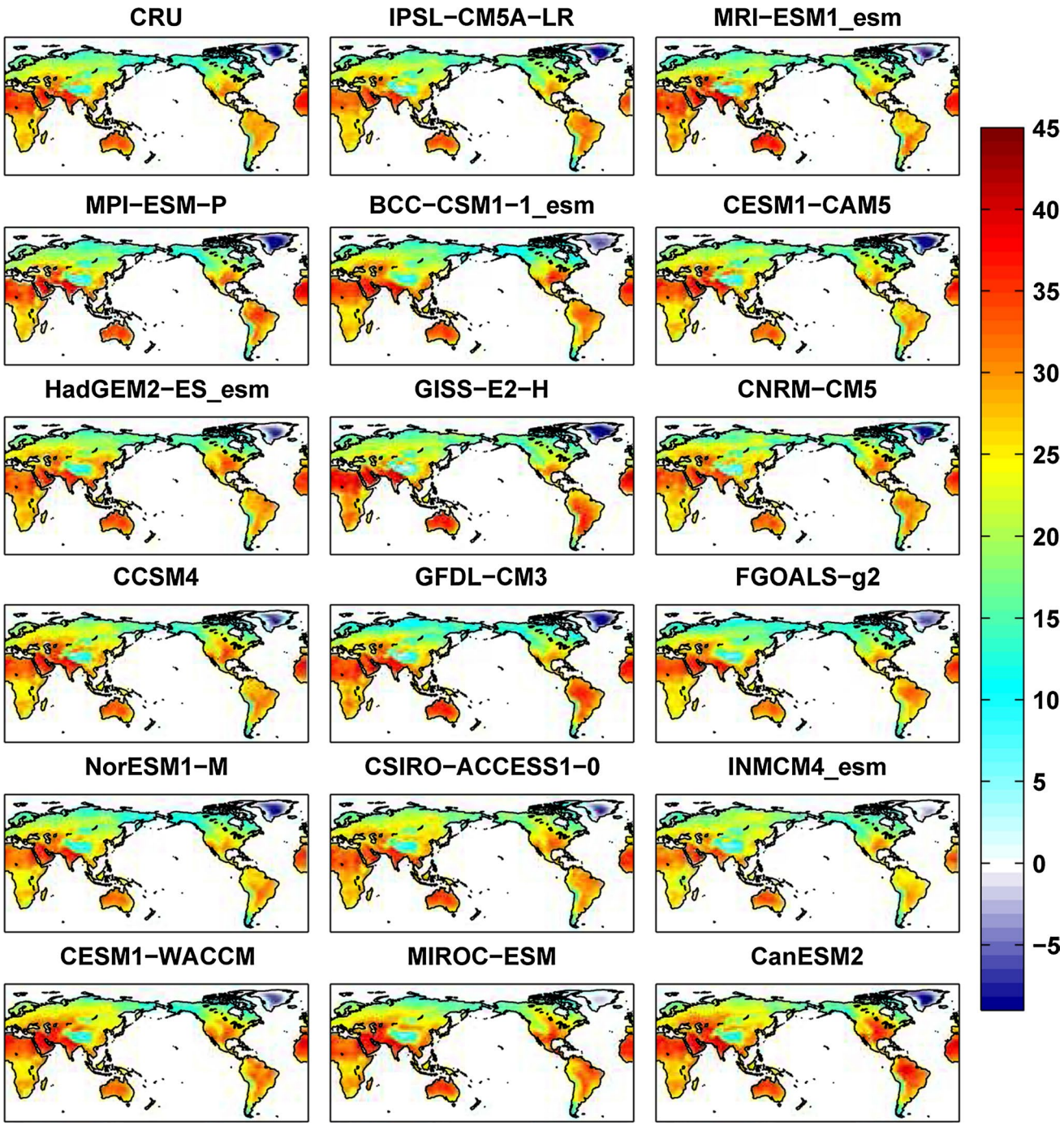

Fig. 6 100-Year return level (in °C) of the annual temperature maxima based on the CRU observations (*upper left panel*), and on selected CMIP5 climate model simulations (*remaining panels*)

4.2 Return levels of temperature extremes

Using the annual temperature maxima from CMIP5 multi-model simulations and CRU observations, temperature return levels are derived for different return periods by fitting the appropriate type of GEV (stationary/non-stationary) to the block maxima of temperature extremes. Return levels of annual temperature maxima are derived and reported for the return periods T of 2, 10, 25, 50 and 100 years.

As an example, Fig. 3 shows the 2-year temperature return levels based on CRU observations (top left) and on the selected subset of 17 CMIP5 climate model simulations. In Fig. 3, the 2-year return levels of global temperature maxima range from −11 to 35 °C. Overall, Fig. 3

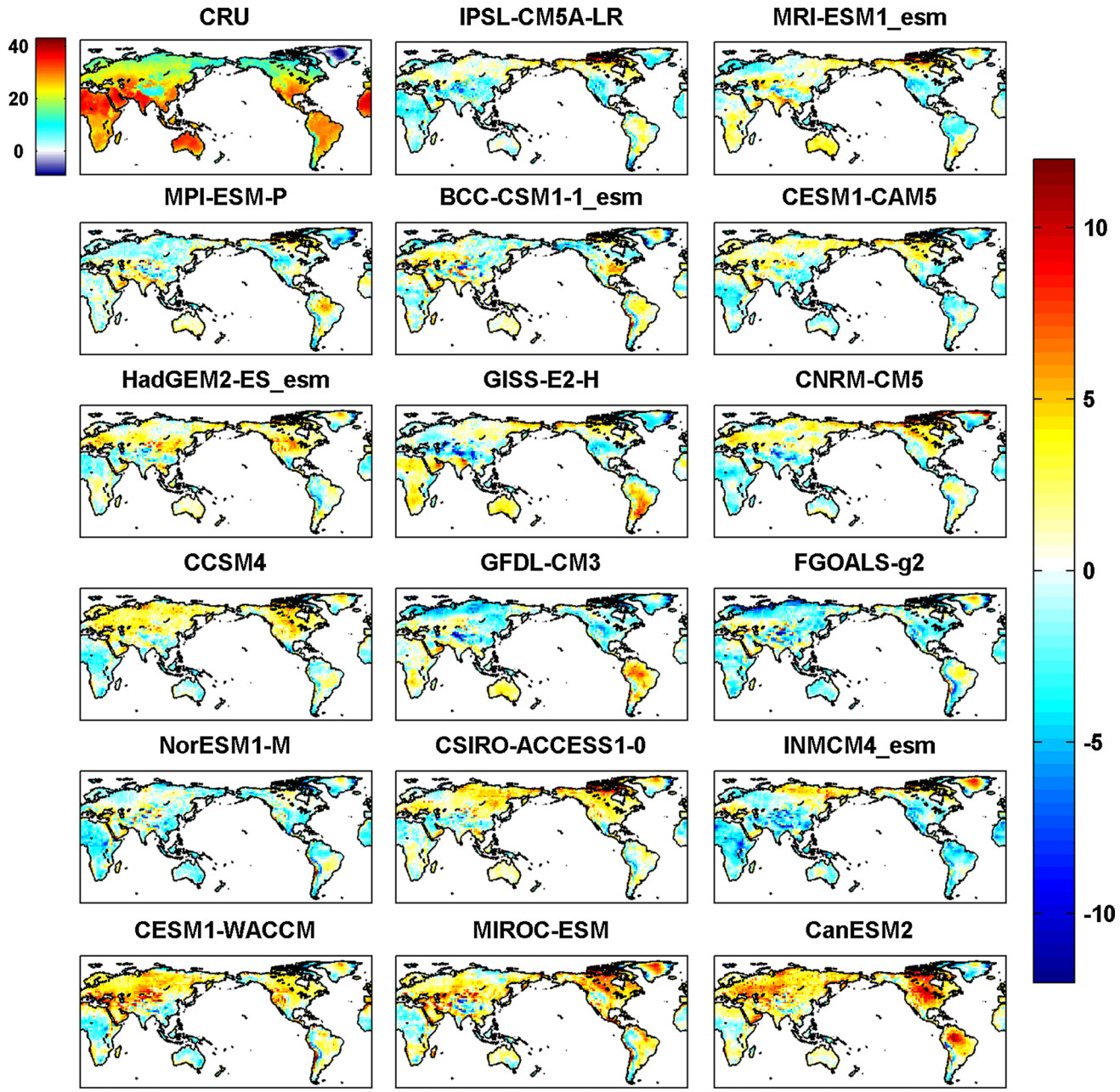

Fig. 7 100-Year return level (in °C) of the annual temperature maxima from CRU observations (*upper left panel*), and the 100-year return level differences between selected CMIP5 climate simulations and CRU reference data (CMIP5–CRU; *remaining panels*)

indicates that there are good agreements between the observed and CMIP5 simulated *spatial patterns* of 2-year annual temperature maxima, but that the magnitudes of 2-year annual temperature maxima represented by the selected CMIP5 models differ substantially.

Figure 4 presents the differences in CMIP5 simulated 2-year annual temperature maxima with respect to CRU observations. One can see that there are variations in both the magnitude and sign of the error of 2-year return levels across CMIP5 climate simulations. This implies that CMIP5 climate models capture the spatial patterns of temperature extremes well; however, individual models may be biased with respect to observations. As shown, over most parts of the world, the biases are within ±4 °C. For a higher return level, one expects the differences in temperature simulations to increase relative to observations. For example, Fig. 5 presents the differences in 25-year-return annual temperature maxima, as simulated by CMIP5 models with respect to CRU temperature observations. As shown, the patterns of differences remain similar, but

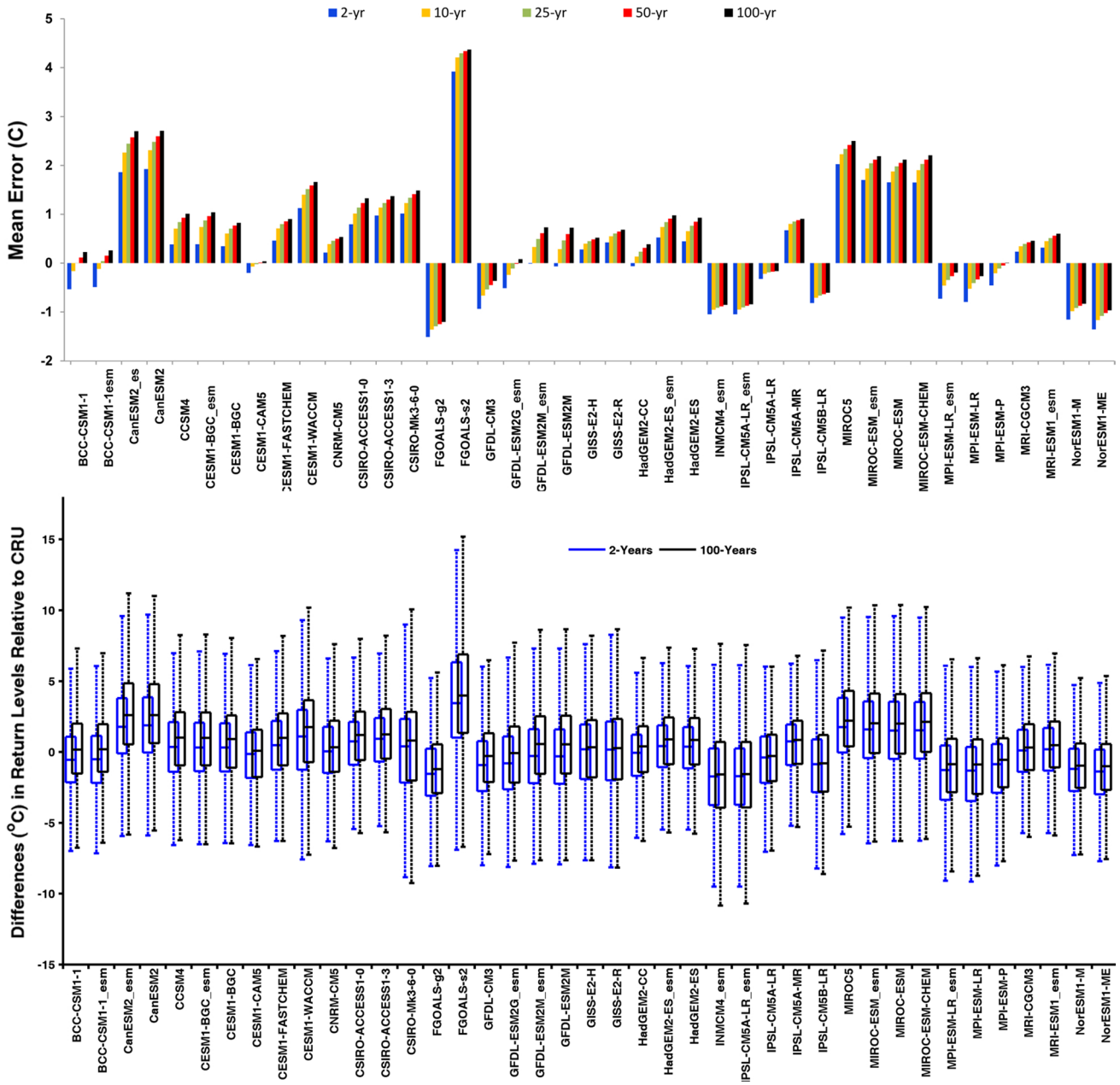

Fig. 8 (*top*) Mean error (ME) of the 2-, 10-, 25-, 50- and 100-year temperature maxima (°C) simulations based on 41 CMIP5 simulations relative to Climatic Research Unit (CRU) observations; (*bottom*): *boxplots* of differences (°C) between CMIP5 2- and 100-year return levels relative to CRU observations

the range of differences between simulated and observed annual temperature maxima increases at the 25-year return level relative to the 2-year return level.

As another example, Fig. 6 displays the 100-year return levels for the CRU observations and the selected CMIP5 simulations. One can see that the patterns of annual temperature maxima are similar to those of Fig. 3, but with higher magnitudes of annual temperature maxima (as expected). The figure shows that the warmest months across the globe typically occur over the Sahara, the Middle East, and Australia. The differences in CMIP5 100-year simulated and observed annual temperature maxima are presented in Fig. 7. As shown, the biases of the 25-, and 100-year return temperature simulations are larger than those of 2-year-return simulations in Fig. 4. However, the spatial patterns of temperature extremes are in a good agreement with CRU observations and consistent across different return periods (compare the model simulations with the upper left panels in Figs. 3, 4, 5, 6). Overall, the regional biases of simulated annual temperature extremes at high return levels

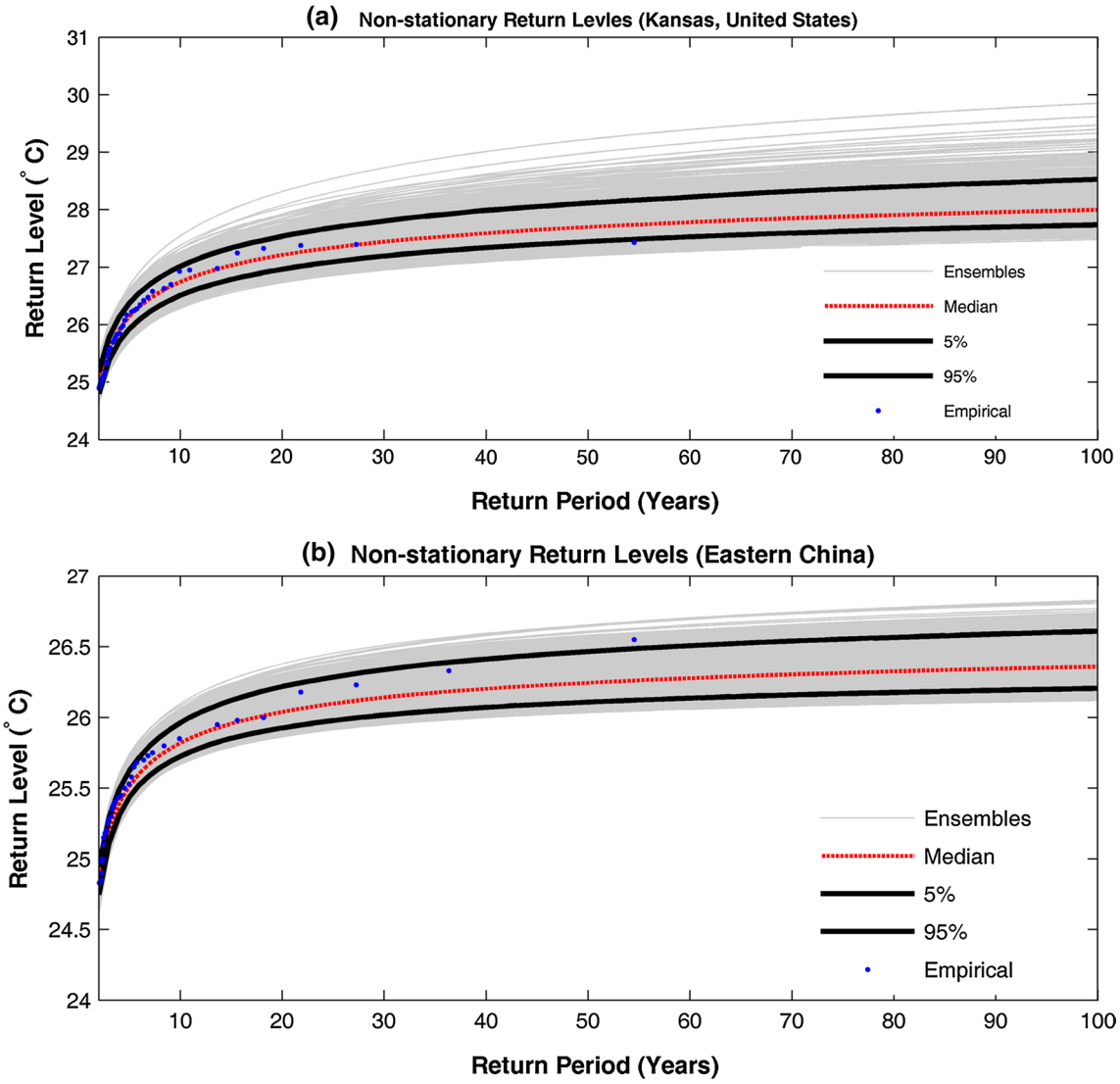

Fig. 9 Sample uncertainty bounds, median, and the 5 and 95 % confidence bounds of the annual temperature maxima based on the DE-MC model for CRU reference and over two pixels in **a** Kansas, United States and **b** Eastern China

(e.g., 100-year) are consistent with those of the lower return levels (e.g., 2-year).

Not shown here for brevity are the spatial biases of 10-, 25- and 50-year return levels of extreme temperature simulations by CMIP5 models, which are consistent with the results presented in Figs. 3, 4, 5, 6 and 7. For a quantitative evaluation of the extremal simulation by CMIP5 models, Fig. 8 (top) summarizes the mean error (ME) for all the 41 CMIP5 climate model simulations of 2-, 10-, 25-, 50 and 100-year annual temperature maxima return levels with respect to CRU observations. As anticipated, ME values are larger at higher return levels. One can see that considering the global averages, most models overestimate the simulated return levels of the annual temperature maxima, while fewer models (e.g., FGOALS-g2, INMCM4_esm, NorESM1-ME) underestimate the temperature extremes. Among the models, FGOALS-s2, CanESM2 and MIROC5 exhibit the highest global averages of the ME of the annual temperature maxima. Most models either systematically overestimate or underestimate the extreme return levels, except the BCC model experiments in which the shorter return levels (2- and 10-year) are underestimated and the longer ones are overestimated. Figure 8 (bottom) displays boxplots of the differences between CMIP5 simulations and CRU observations. The figure shows medians, 25th and 75th percentiles, and whiskers (variability outside respective percentiles) of differences in °C. Figure 8 (bottom) indicates that while local differences can be large, most differences (between 25th and 75th percentiles) fall within ±2 °C.

Table 2 Results of the Bayes factor at the selected pixels in the Central US and Eastern China

Location	K	Test interpretation
Bayes factor		
Kansas, USA	0.49	Reject stationary model
Eastern China	0.003	Reject stationary model

Bayes factor (K) larger than one indicates the stationary assumption (i.e., null hypothesis) is representative, whereas $K < 1$ rejects the null hypothesis

The MCMC component of the DE-MC model used in this study allows the upper and lower bounds and credible intervals of the temperature return levels to be derived based on all model parameters. The uncertainty bounds would be different across either models or space (simulation grid boxes). As an example, Fig. 9 shows sample uncertainty bounds, median, and the 5 and 95 % confidence bounds of the annual temperature maxima based on the DE-MC for CRU reference data and over grid boxes in two different locations in Kansas, United States and eastern China under the non-stationary assumption. The figure confirms that the inference uncertainty is larger at higher return levels (e.g., because of larger sampling errors). One can see that the uncertainties of the estimated return levels also vary over different regions. It should be noted that this approach provides uncertainties associated with the statistical analysis of extremes, but does not include uncertainties associated with model physics.

As mentioned in the Methodology Section, the initial assumption of stationarity and non-stationarity by the Mann–Kendal trend test is tested using the Bayes factor. As an example, for the selected locations in Fig. 9, the Bayes factor results are provided for testing non-stationarity in extremes to make sure the initial assumption from the Mann–Kendal test is reasonable (see Table 2). As shown in Table 2, the method confirms the initial assumption of non-stationarity by the Mann–Kendal trend test.

5 Discussion and concluding remarks

The objective of this study is to evaluate to what extent the CMIP5 climate model simulations can represent observed warm monthly temperature extremes under a changing climate. The biases of simulated annual temperature maxima are quantified for the selected CMIP5 models. Furthermore, the 2-, 10-, 25-, 50- and 100-year return levels of the annual temperature maxima from CMIP5 simulations are compared with those derived from CRU observations.

The results show that most, but not all, CMIP5 climate models tend to underestimate the mean annual temperature maxima over the United States and Amazonia. The CMIP5 models particularly disagree with each other over cold regions (e.g., Russia, northern Canada), with the exception of Greenland where most climate models underestimate the mean annual temperature maxima. This underestimation of the annual temperature maxima is likely to affect model-based representations of changes in ice-sheets and snow/glacier melt. In contrast, over Alaska, Northern Canada and Siberia, most CMIP5 simulations overestimate the annual temperature maxima compared to those derived from the CRU reference data.

Over arid and semi-arid regions (e.g., the Sahara, southwestern US, and Middle East), most climate models also underestimate the mean annual temperature maxima. Considering the magnitudes of deviations from the CRU, however, there is a better agreement between CMIP5 model simulations and observations in subtropical regions than in high-latitude cold regions.

The return level analyses show that there are good agreements between the observed and CMIP5 simulated spatial patterns of 2-, 10-, 25-, 50- and 100-year annual temperature maxima. While the simulated spatial patterns of the temperature extremes are similar, the magnitudes of the return levels of the annual temperature maxima represented by CMIP5 climate models are biased with respect to CRU observations. In addition, there are variations in both the magnitude and sign of the biases of the annual temperature maxima return levels across the CMIP5 simulations. The results reveal that most CMIP5 simulations overestimate the global averages of the annual temperature maxima at different return periods (see Fig. 8).

Given the state of the science in climate modeling, one would not expect the coupled atmosphere/ocean general circulation model (AOGCMs) and ESMs to reproduce the magnitudes of the observed historical extremes very accurately. Rather, one expects the models to reasonably simulate large-scale patterns of change in occurrences of climate extremes (Tebaldi et al. 2006). Overall, the results of this study indicate that the models capture the spatial patterns of temperature extremes well, but that individual models are biased relative to the CRU observations.

Acknowledgments The authors thank the three anonymous reviewers for their constructive suggestions which significantly improved the paper. The financial support for authors LC and AA was made available by the Environmental Sciences Division of the Army Research Office Award No. W911NF-14-1-0684. The contributions of author TJP were performed under the auspices of the Lawrence Livermore National Laboratory under Contract DE-AC52-O7NA27344. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate-modeling groups for producing and making available their model output. For CMIP, the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and leads the development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

References

AghaKouchak A, Easterling D, Hsu K, Schubert S, Sorooshian S (2013) Extremes in a changing climate. Springer, Dordrecht, p 426

AghaKouchak A, Mehran A (2013) Extended contingency table: performance metrics for satellite observations and climate model simulations. Water Resour Res 49:7144–7149. doi:10.1002/wrcr.20498

Aghakouchak A, Cheng L, Mazdiyasni O, Farahmand A (2014) Global warming and changes in risk of concurrent climate extremes: insights from the 2014 California drought. Geophys Res Lett 41:8847–8852. doi:10.1002/2014GL062308

AghaKouchak A, Nasrollahi N (2010) Semi-parametric and parametric inference of extreme value models for rainfall data. Water Resour Manag 24(6):1229–1249

Alexander L, Zhang X, Peterson T, Caesar J, Gleason B, Klein Tank A, Haylock M, Collins D, Trewin B, Rahimzadeh F, Tagipour A, Ambenje P, Rupa Kumar K, Revadekar J, Griffiths G (2006) Global observed changes in daily climate extremes of temperature. J Geophys Res 111:D05109

Alley RB, Clark PU, Huybrechts P, Joughin I (2005) Ice-sheet and sea-level changes. Science 310(5747):456–460

Bonnin G, Martin D, Lin B, Parzybok T, Yekta M, Riley D (2004) Precipitation-frequency Atlas of the United States. NOAA Atlas 14, vol 2. NOAA Natl

Cheng L, AghaKouchak A (2014) Nonstationary precipitation intensity-duration-frequency curves for infrastructure design in a changing climate. Sci Rep 4:7093. doi:10.1038/srep07093

Cheng L, AghaKouchak A, Gilleland E, Katz R (2014a) Non-stationary extreme value analysis in a changing climate. Clim Change 127(2):353–369. doi:10.1007/s10584-014-1254-5

Cheng L, Gilleland E, Heaton MJ, AghaKouchak A (2014b) Empirical Bayes estimation for the conditional extreme value model. Stat 3:391–406. doi:10.1002/sta4.71

Clarke RT (2002) Estimating trends in data from the Weibull and a generalized extreme value distribution. Water Resour Res 38(6):1089. doi:10.1029/2001WR000575

Coles S (2001) An introduction to statistical modeling of extreme values. Springer, London, p 209

Cooley D (2009) Extreme value analysis and the study of climate change. Clim Change 97(1–2):77–83

Cooley D (2013) Return periods and return levels under climate change. In: Extremes in a changing climate. Springer, Berlin, pp 97–114. doi:10.1007/978-94-007-4479-0_4

Cooley D, Nychka D, Naveau P (2007) Bayesian spatial modeling of extreme precipitation return levels. J Am Stat Assoc 102(479):824–840

Easterling D, Meehl G, Parmesan C, Changnon S, Karl T, Mearns L (2000) Climate extremes: observations, modeling, and impacts. Science 289(5487):2068–2074

Field CB, Barros V, Stocker TF, Qin D, Dokken D, Ebi K, Mastrandrea M, Mach K, Plattner G, Allen S et al (2012) Managing the risks of extreme events and disasters to advance climate change adaptation. Cambridge University Press, Cambridge, p 594

Fisher R, Tippett L (1928) Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proc Camb Philos Soc 24:180–190

Gelman A, Shirley K (2011) Inference from simulations and monitoring convergence. In: Brooks S, Gelman A, Jones GL, Meng X-L (eds) Handbook of Markov Chain Monte Carlo, 163–174. Chapman and Hall/CRC, London, p 619

Gilleland E, Katz RW (2006) Analyzing seasonal to interannual extreme weather and climate variability with the extremes toolkit. In: 18th Conference on climate variability and change, 86th American Meteorological Society (AMS) Annual Meeting, vol 29

Gleckler P, Taylor K, Doutriaux C (2008) Performance metrics for climate models. J Geophys Res Atmos 113(D6):D06104

Gnedenko B (1943) Sur la distribution limite du terme maximum d'une série aléatoire. Ann Math 44:423–453

Granger CW, Inoue T, Morin N (1997) Nonlinear stochastic trends. J Econ 81(1):65–92

Gumbel E (1942) On the frequency distribution of extreme values in meteorological data. Bull Am Meteorol Soc 23(3):95

Hansen J, Ruedy R, Sato M, Lo K (2010) Global surface temperature change. Rev Geophys 48(4):RG4004

Hao Z, AghaKouchak A, Phillips T (2013) Changes in concurrent monthly precipitation and temperature extremes. Environ Res Let 8:034014. doi:10.1088/1748-9326/8/3/034014

Hoerling M et al (2013) An interpretation of the origins of the 2012 Central Great Plains drought. Technical report, Assessment Report, National Oceanic and Atmospheric Administration, Drought Task Force

IPCC (2007) Climate Change 2007: impacts, adaptation, and vulnerability. In: Parry Martin L, Canziani Osvaldo F, Palutikof Jean P, van der Linden Paul J, Hanson Clair E (eds) Contribution of working group II to the third assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge, p 987

Karl T, Knight R (1998) Secular trends of precipitation amount, frequency, and intensity in the USA. Bull Am Meteorol Soc 79:231–241

Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90(430):773–795

Katz R (2010) Statistics of extremes in climate change. Clim Change 100(1):71–76

Katz R (2013) Statistical methods for nonstationary extremes. In: Extremes in a changing climate. Springer, Berlin, pp 15–37. doi:10.1007/978-94-007-4479-0

Katz R, Parlange M, Naveau P (2002) Statistics of extremes in hydrology. Adv Water Resour 25:1287–1304

Kendall M (1976) Rank correlation methods, 4th edn. Griffin, London, p 272

Kharin V, Zwiers F (2005) Estimating extremes in transient climate change simulations. J Clim 18:1156–1173

Kharin V, Zwiers F, Zhang X, Hegerl G (2007) Changes in temperature and precipitation extremes in the IPCC ensemble of global coupled model simulations. J Clim 20(8):1419–1444

Kharin V, Zwiers F, Zhang X, Wehner M (2013) Changes in temperature and precipitation extremes in the CMIP5 ensemble. Clim Change 119(2):345–357

Kilbourne E (1997) Heat waves and hot environments. In: Noji Eric K (ed) The public health consequences of disasters. Oxford University Press, New York, pp 245–269

Kjær K, Khan S, Korsgaard N, Wahr J, Bamber J, Hurkmans R, van den Broeke M, Timm L, Kjeldsen K, Bjørk A et al (2012) Aerial photographs reveal late-20th-century dynamic ice loss in northwestern Greenland. Science 337(6094):569–573

Krabill W, Hanna E, Huybrechts P, Abdalati W, Cappelen J, Csatho B, Frederick E, Manizade S, Martin C, Sonntag J et al (2004) Greenland ice sheet: increased coastal thinning. Geophys Res Lett 31(24):L24402

Kunkel K (2013) Uncertainties in observed changes in climate extremes. In: Extremes in a changing climate. Springer, Berlin, pp 287–307. doi:10.1007/978-94-007-4479-0_10

Leadbetter M, Lindgren G, Rootzén H (1983) Extremes and related properties of random sequences and processes. Springer, New York, p 336. doi:10.1007/978-1-4612-5449-2

Li Y, Cai W, Campbell E (2005) Statistical modeling of extreme rainfall in southwest western Australia. J Clim 18(6):852–863

Liu Z, Mehran A, Phillips T, AghaKouchak A (2014) Seasonal and regional biases in CMIP5 precipitation simulations. Clim Res 60:35–50. doi:10.3354/cr01221

Makkonen L (2006) Plotting positions in extreme value analysis. J Appl Meteorol Climatol 45(2):334–340

Mann H (1945) Nonparametric tests against trend. Econometrica 13:245–259

Mays L (2010) Water resources engineering. Wiley, Hoboken, NJ, p 928

McMichael AJ et al (2003) Climate change and human health: risks and responses. World Health Organization, Geneva. ISBN 92 4 156248 X

Meehl G, Bony S (2011) Introduction to CMIP5. CLIVAR Exch 16(2):4–5

Meehl GA, Karl T, Easterling DR, Changnon S, Pielke R Jr, Changnon D, Evans J, Groisman PY, Knutson TR, Kunkel KE et al (2000) An introduction to trends in extreme weather and climate events: observations, socioeconomic impacts, terrestrial ecological impacts, and model projections. Bull Am Meteorol Soc 81(3):413–416

Mehran A, AghaKouchak A, Phillips T (2014) Evaluation of CMIP5 continental precipitation simulations relative to GPCP satellite observations. J Geophys Res 119:1695–1707. doi:10.1002/2013JD021152

Milly P, Julio B, Malin F, Robert M, Zbigniew W, Dennis P, Ronald J (2008) Stationarity is dead: whither water management? Science 319(5863):573–574

Mitchell T, Jones P (2005) An improved method of constructing a database of monthly climate observations and associated high-resolution grids. Int J Climatol 25(6):693–712

Morak S, Hegerl GC, Christidis N (2013) Detectable changes in the frequency of temperature extremes. J Clim 26:1561–1574

Morton R, Henderson BL (2008) Estimation of nonlinear trends in water quality: an improved approach using generalized additive models. Water Resour Res 44(7):W07420. doi:10.1029/2007WR006191

New M, Hulme M, Jones P (2000) Representing twentieth-century space-time climate variability. Part II: development of 1901–96 monthly grids of terrestrial surface climate. J Clim 13(13):2217–2238. doi:10.1175/JCLI-D-11-00678.1

NOAA (1980) Impact assessment: U.S. social and economic effects of the great 1980 heat wave and drought. U.S. Department of Commerce, National Oceanic and Atmospheric Administration, Environmental Data and Information Service, Center for Environmental Assessment Services, Washington, DC

NOAA (1995) Impact assessment: the July 1995 heat wave natural disaster survey report. U.S. Department of Commerce, National Oceanic and Atmospheric Administration, National Weather Service, Silver Spring, MD

Papalexiou SM, Koutsoyiannis D (2013) Battle of extreme value distributions: a global survey on extreme daily rainfall. Water Resour Res 49(1):187–201. doi:10.1029/2012WR012557

Parey S, Hoang TTH, Dacunha-Castelle D (2010) Different ways to compute temperature return levels in the climate change context. Environmetrics 21(7–8):698–718

Phillips PC (2007) Regression with slowly varying regressors and nonlinear trends. Econ Theory 23(04):557–614

Phillips T, Gleckler P (2006) Evaluation of continental precipitation in 20th century climate simulations: the utility of multimodel statistics. Water Resour Res 42(3):W03202

Reeh N (1989) Parameterization of melt rate and surface temperature on the Greenland ice sheet. Polarforschung 59(3):113–128

Reiss R-D, Thomas M (2007) Statistical analysis of extreme values: with applications to insurance, finance, hydrology and other fields. Springer, Berlin, p 512

Ren D, Fu R, Leslie L, Chen J, Wilson C, Karoly D (2011) The Greenland ice sheet response to transient climate change. J Clim 24(13):3469–3483

Renard B, Garreta V, Lang M (2006) An application of Bayesian analysis and Markov Chain Monte Carlo methods to the estimation of a regional trend in annual maxima. Water Resour Res 42(12). doi:10.1029/2005WR004591

Renard B, Sun X, Lang M (2013) Bayesian Methods for non-stationary extreme value analysis. In: Extremes in a changing climate. Springer, Berlin, pp 39–95. doi:10.1007/978-94-007-4479-0_3

Rootzén H, Katz RW (2013) Design life level: quantifying risk in a changing climate. Water Resour Res 49(9):5964–5972

Schlather M (2002) Models for stationary max-stable random fields. Extremes 5(1):33–44

Serinaldi F (2014) Dismissing return periods! Stoch Environ Res Risk Assess. doi:10.1007/s00477-014-0916-1

Serinaldi F, Kilsby CG (2015) Stationarity is undead: uncertainty dominates the distribution of extremes. Adv Water Resour. doi:10.1016/j.advwatres.2014.12.013

Sillmann J, Kharin V, Zhang X, Zwiers F, Bronaugh D (2013) Climate extremes indices in the CMIP5 multimodel ensemble: part 1. Model evaluation in the present climate. J Geophys Res Atmos 118(4):1716–1733

Smith R (2001) Extreme value statistics in meteorology and environment. In: Environmental statistics, Chapter 8, pp 300–357. http://www.unc.edu/depts/statistics/postscript/rs/envnotes

Solomon S, Qin D, Manning M, Marquis M, Averyt K, Tignor M, LeRoy Miller H, Chen Z (2007) Projections of future climate change. Climate change 2007: the physical science basis. Contribution of working group I to the fourth assessment report of the Intergovermental Panel on Climate Change. Cambridge University Press, New York, NY, pp 1007

Sorooshian S, AghaKouchak A, Li J (2014) Influence of irrigation on land hydrological processes over California. J Geophys Res Atmos doi:10.1002/2014JD022232

Tarroja B et al (2014) Evaluating options for balancing the water–electricity nexus in California: part 1 Securing water availability. Sci Total Environ 497–498:697–710. doi:10.1016/j.scitotenv.2014.06.060

Tarroja B et al (2014) Evaluating options for balancing the water-electricity nexus in California: part 2 Greenhouse gas and renewable energy utilization impacts. Sci Total Environ 497–498:711–724. doi:10.1016/j.scitotenv.2014.06.071

Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. Bull Am Meteorol Soc 93(4):485–498

Tebaldi C, Hayhoe K, Arblaster J, Meehl G (2006) Going to the extremes. Clim Change 79(3):185–211

ter Braak CJ (2004) Genetic algorithms and Markov Chain Monte Carlo: differential evolution Markov Chain makes Bayesian computing easy. Biometris, Wageningen UR, The Netherlands

ter Braak C (2006) A Markov Chain Monte Carlo version of the genetic algorithm differential evolution: easy Bayesian computing for real parameter spaces. Stat Comput 16(3):239–249

Villarini G, Smith JA, Ntelekos AA, Schwarz U (2011) Annual maximum and peaks-over-threshold analyses of daily rainfall accumulations for Austria. J Geophys Res Atmos (1984–2012) 116:D05103. doi:10.1029/2010JD015038

Vose R, Wuertz D, Peterson T, Jones P (2005) An intercomparison of trends in surface air temperature analyses at the global, hemispheric, and grid-box scale. Geophys Res Lett 32(18):L18718

Vrugt JA, ter Braak C, Diks C, Robinson BA, Hyman JM, Higdon D (2009) Accelerating Markov Chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. Int J Nonlinear Sci Numer Simul 10(3):273–290

Wehner M (2013) Methods of projecting future changes in extremes. In: Extremes in a changing climate. Springer, Berlin. doi:10.1007/978-94-007-4479-0

Winkler RL (1973) A Bayesian approach to nonstationary processes. IIASA

Zhang X, Zwiers F (2013) Statistical indices for the diagnosing and detecting changes in extremes. In: Extremes in a changing climate. Springer, Berlin. doi:10.1007/978-94-007-4479-0_1

Zwiers FW, Kharin VV (1998) Changes in the extremes of the climate simulated by CCC GCM2 under CO_2 doubling. J Clim 11(9):2200–2222