Systematic differences in future 20 year temperature extremes in AR4 model projections over Australia as a function of model skill

Author:
Perkins, Sarah; Pitman, Andrew; Sisson, Scott

Publication details:
International Journal of Climatology
v. 33
Chapter No. 5
pp. 1153-1167
0899-8418 (ISSN)

Publication Date:
2012

Publisher DOI:
http://dx.doi.org/10.1002/joc.3500

INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **33**: 1153–1167 (2013)
Published online 27 April 2012 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/joc.3500

RMetS
Royal Meteorological Society

Systematic differences in future 20 year temperature extremes in AR4 model projections over Australia as a function of model skill

S. E. Perkins,[a]* A. J. Pitman[a] and S. A. Sisson[b]
[a] *Climate Change Research Centre, University of New South Wales, NSW, Australia*
[b] *School of Mathematics and Statistics, University of New South Wales, NSW, Australia*

ABSTRACT: The projection of temperature extremes by climate models participating in the Intergovernmental Panel on Climate Change Fourth Assessment Report (AR4) are examined regionally over Australia. Minimum and maximum temperature extremes are defined as the 20 year return value calculated using extreme value theory. Three measures of model evaluation, a means-based, a distribution-based [via probability density functions (PDFs)] and an extreme-based (via the tails of PDFs) method, are used to compare daily model data to observed daily data over various climatic regions for a 20 year period. Model ensembles consisting of the 'better' and 'poorer' models determined by each measure of skill are created for each region. These are compared with an all-model ensemble to examine the difference in more skilled ensemble projections of temperature extremes in the A2 (high emissions) scenario for 2046-2065 and 2081-2100. If either of the distribution-based evaluation methods were used to distinguish models, the higher skilled models projected smaller increases in the 20 year return values than the all-model ensemble for both maximum temperature and minimum temperature. For some regions, the 90% confidence intervals of the better and poorer ensemble ranges did not overlap, indicating that projections are statistically significantly different. We show that the means-based evaluation produces less consistent results to the two distribution-based evaluation methods. We conclude that specific AR4 models, shown to be relatively poor over most regions of Australia by different skill metrics, bias the projected increase in the 20 year temperature extremes towards higher values. We also suggest that performance in simulating the mean climate is an unreliable measure of climate model capacity used to select models for projecting changes in extremes over Australia. Copyright © 2012 Royal Meteorological Society

KEY WORDS temperature extremes; model evaluation; AR4 models; Australia

Received 17 November 2010; Revised 25 March 2012; Accepted 26 March 2012

1. Introduction

Most future climate studies have focused on changes in the mean (e.g. Allen *et al*., 2003; Moise and Hudson, 2008; Smith *et al*., 2009). However, changes in climate extremes due to global warming may affect human and biological systems more than changes in the mean (Mearns *et al*., 1984; Katz and Brown, 1992; Easterling *et al*., 2000; Kharin and Zwiers, 2000; Kharin *et al*., 2007; Fischer and Schar, 2010). Although changes in temperature extremes are related to shifts in the mean (Kharin *et al*., 2007), the magnitude of changes in extremes cannot be inferred solely from this relation (Schaeffer *et al*., 2005). Earlier studies, including those by Mearns *et al*. (1984), Katz and Brown (1992), Colombo *et al*. (1999) and Meehl *et al*. (2000), suggest that extremes may change more than indicated by a change in the mean, particularly if both the location and shape parameters of the probability density function (PDF) change.

There have been several recent studies using climate models assessed by the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4) that have investigated changes in temperature extremes at global or regional scales. Hegerl *et al*. (2004) used two AR4 climate models and showed that changes in temperature extremes were significantly different from changes in seasonal means in up to 66% of model grid points. Tebaldi *et al*. (2006) used nine AR4 models to demonstrate that the twentieth century trend in temperature extremes would likely be amplified under higher greenhouse forcing. Kharin *et al*. (2007) used 16 AR4 models and showed that globally averaged cold extremes warmed faster than warm extremes under all available emission scenarios. There have been few studies focusing on extremes over the Australian region, although changes in the mean are well documented (Moise and Hudson, 2008). Pitman and Perkins (2008) explored projected changes in the *annual* return values for maximum temperature (T_{max}) and minimum temperature (T_{min}) but these

* Correspondence to: S. E. Perkins, Climate Change Research Centre, Mathews Building, The University of New South Wales, Sydney, NSW 2052, Australia. E-mail: sarah.perkins@unsw.edu.au

are not 'extreme' in the sense that they may dramatically affect human health, ecosystems or other biophysical systems. Alexander and Arblaster (2009) focussed on the potential change in extreme indices over Australia for the twenty-first century using the AR4 models and demonstrated that the models were generally skilful in replicating trends in twentieth-century indices. Gallant and Karoly (2010) employed a combined climate extremes index to study observational trends in temperature and precipitation extremes. In terms of temperature, they found an increase (decrease) in the spatial extent of hot (cold) extremes of 1% (2%) decade^{-1}.

In this article, we use the 20 year return value as an indicator of an extreme at a magnitude that might severely affect humans and natural ecosystems. The nature of extreme values means that they occur at the tails of a distribution, which are sparsely populated (Wehner, 2004; Wehner *et al*., 2010). Extreme value theory (EVT) is used to study and extrapolate these rare events, as small discrepancies in the estimation of the empirical distribution may lead to considerable errors in the distribution of extreme values (Coles, 2001; Rusticucci and Tencer, 2008). EVT, in terms of the generalized extreme value (GEV) distribution, was first employed to study climate extremes by Zwiers and Kharin (1998) and is now widely used by the climate science community (Kharin and Zwiers, 2000; Kharin *et al*., 2005, 2007; Schaeffer *et al*., 2005; Brown *et al*., 2008; Rusticucci and Tencer, 2008; Sterl *et al*., 2008; Fowler *et al*., 2010; Wehner *et al*., 2010; Perkins, 2011).

This article uses the GEV distribution, following the approach of Kharin *et al*. (2007), but focused on Australia. We build on Kharin *et al*. (2007) by implementing an additional step before the calculation of the return values. Before using climate models to project future conditions, the models are usually compared with the current climate. Obviously, a model that can simulate current conditions well is not necessarily able to simulate future conditions (Jun *et al*., 2008; Weigel *et al*., 2010), although Macadam *et al*. (2010) suggested there is a decade-to-decade consistency in climate model skill in the simulation of mean temperature. There is no agreed 'best way' to evaluate a climate model (Knutti *et al*., 2010; Weigel *et al*., 2010; Irving *et al*., 2011; Klocke *et al*., 2011). While evaluation based on comparing means is common (Macadam *et al*., 2010), this tells us little about a model's capacity to simulate extremes. Perkins *et al*. (2007) suggested a metric that measures the amount of overlap between an observed and modelled PDF, which was employed to produce skill scores over Australia for daily T_{min}, T_{max} and precipitation. They found that overall most models captured the variability in the observed PDF reasonably well, but some AR4 models exhibited poorer skill relative to others and the same models tended to be the poorest in many regions. Perkins and Pitman (2009) and Pitman and Perkins (2008) use this skill score to select the more capable models (omit less capable models) for future projections under various scenarios over Australia and explore changes up to the scale of the annual event. Since the all-model ensembles were biased by poorer models on timescales up to the annual event, Stainforth *et al*.'s (2007) suggestion of omitting 'models whose simulations are "substantially" worse than state-of-the-art models' was implemented. However, a PDF-based evaluation does not focus solely on the tails (i.e. extremes) of a distribution. Thus, this article extends the approach of Perkins *et al*. (2007) by adding a metric focusing on the tails of a daily-based probability distribution.

This article therefore explores the AR4 model projections of the 20 year return levels for T_{max} and T_{min} over Australia. EVT is used to calculate, region-by-region, the 20 year return levels over Australia. Three measures of climate model skill are used to evaluate the AR4 models. We then explore the sensitivity of the projected 20 year returns to each skill measure to determine whether a systematic bias exists in some models to a degree that affects the ensemble projection of 20 year return values.

2. Methods

2.1. Modelled and observed data

All model data were downloaded from the Program for Climate Model Diagnosis and Intercomparison (PCMDI) at the Lawrence Livermore National Laboratory in the USA (http://www-pcmdi.llnl.gov/about_ipcc.php). Daily data for T_{min} and T_{max} for the Climate of the Twentieth Century and A2 emission scenarios were utilized for all models that had data common for all experiments. This included 11 models each for T_{max} and T_{min}, common across both variables. Table I lists all models used, their respective resolutions and the number of independent realizations for each variable. Models that had multiple runs were aggregated to form a single ensemble, as it was found by Kharin *et al*. (2007) and Perkins *et al*. (2007) there was negligible difference between individual realizations from a given model. Australia was extracted from the global data sets and native land–ocean masks were fitted to delete ocean values. For the twentieth century, the period 1981–2000 was used, as it was common among all models and was an appropriate time period to compare in terms of the mean and return values to the A2 scenario. There are two time periods for the A2 scenario, both of length 20 years, representative of 2050 and 2100.

Daily observed T_{min} and T_{max} were obtained from 1178 Australian Bureau of Meteorology stations for 1981–2000. The use of this data, their spatial distribution and homogeneity issues is fully discussed in the study by Perkins *et al*. (2007). Figure 1 shows the spatial distribution of temperature observation stations over Australia, and Table II lists the latitudinal and longitudinal boundaries (consistent with those in Figure 1) of the regions used in this study, as well as their climatic types.

2.2. GEV distribution

The GEV distribution is used to estimate a return value over a given period. A return value is an event of a

Table I. List of the AR4 models used with daily data common to all scenarios. The acronym used for each model in the text is italicized in column 1. Numbers in the fourth rightmost columns indicate the number of runs used for each model.

Model name and *acronym*	Affiliation/country	Resolution		Number of realizations			
		Horizontal	Vertical	T_{max}		T_{min}	
				20thC	A2	20thC	A2
BCCR-BCM2.0	Bjerknes Centre for Climate Research, Norway	1.9° × 1.9°	L31	1	1	1	1
CGCM 3.1	Canadian Centre for Climate Modeling & Analysis, Canada	~2.8° × 2.8°	L31	4	3	5	3
CSIRO3.0	CSIRO Atmospheric Research, Australia	~1.9° × 1.9°	L18	2	1	3	1
CSIRO3.5	CSIRO Atmospheric Research, Australia	~1.9° × 1.9°	L18	1	1	1	1
ECHAM 5/MPI-OM	Max Planck Institute for Meteorology, Germany	~1.9° × 1.9°	L31	1	1	1	1
ECHO-G	Meteorological Institute of the University of Bonn, Meteorological Research Institute of KMA, and Model & Data Group, Germany/Korea	~3.9° × 3.9°	L19	2	3	3	3
GFDL2.0	Geophysical Fluid Dynamics Laboratory, USA	~2.5° × 2.5°	L24	1	1	1	1
GFDL2.1	Geophysical Fluid Dynamics Laboratory, USA	~2.5° × 2.5°	L24	1	1	1	1
IPSL-CM4	Institut Pierre Simon Laplace, France	~2.5° × 3.75°	L19	2	1	2	1
MIROC 3.2 (medres)	Center for Climate System Research (University of Tokyo), National Institute for Environmental Studies, and Frontier Research Center for Global Change (JAMSTEC), Japan	~2.8° × 2.8°	L20	2	3	1	3
MRI-CGCM2.3.2	Meteorological Research Institute, Japan	~2.8° × 2.8°	L30	3	4	2	5

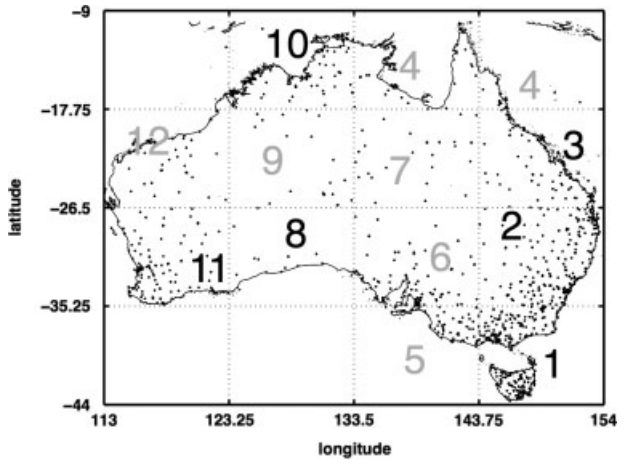

Figure 1. Spatial distribution of temperature observations and the locations of regions 1, 2, 3, 8, 10 and 11, as discussed in the text. Regions marked in grey are the further six regions used by Perkins *et al*. (2007).

certain magnitude that occurs once, on average within the return period, t. For example, an event with a return period of $t = 20$ years has a probability of $1/20 = 0.05$ (5%) of occurring within a given year. When considering the cumulative density function [CDF, $F(X)$] based on annual data, the 20 year return value for T_{max} is $F(X) = 1 - 1/t = 0.95$ and for T_{min} is $F(X) = 1/t = 0.05$.

Extreme value samples are extracted from the original daily data set before fitting the GEV distribution, taken as the annual maxima for T_{max} and the annual minima for T_{min}. At the regional scale, samples were formed by taking the annual maxima (minima) for each model grid box in a given region and concatenating to create a region-specific sample per model. The sample size in this case is dependant on the model's horizontal resolution (Table I). At the continental scale, samples are formed separately for each land grid element at the model's native resolution.

The theory behind the derivation of the GEV distribution as the limiting distribution of the largest observation in a sample is given in detail by Zwiers and Kharin (1998), Kharin and Zwiers (2000) and Kharin *et al*. (2005). The GEV distribution has three parameters: location, scale and shape (k). The three distributional subfamilies are distinguished by k. In the limit as $k \to 0$, the GEV distribution reduces to the Gumbel distribution, which exhibits exponential (light) tail decay; $k < 0$ leads to the Frechét distribution with polynomial (heavy) tail decay and $k > 0$ leads to the Weibull distribution, which has a finite upper endpoint. There are two methods that are commonly used to estimate the GEV distribution parameters: L-moments (probability-weighted moments) and maximum likelihood. Although the method of L-moments assumes stationarity of annual extremes, the method of maximum likelihood is less efficient for

Table II. The latitudinal and longitudinal boundaries of the regions used in this study, and their climatic types. Figure 1 shows the location of these regions, as well as the other six regions used by Perkins *et al.* (2007).

Region	Latitude	Longitude	Climate
1	35.25 °S–44 °S	143.75 °E–154 °E	Temperate
2	26.5°–35.25 °S	143.75 °E–154 °E	Desert/grassland/temperate
3	17.75 °S–26.5 °S	143.75 °E–154 °E	Desert/grassland/subtropical
8	26.5 °S–35.25 °S	133.5 °E–123.25 °E	Grassland/desert
10	9 °S–17.75 °S	133.5 °E–123.25 °E	Grassland/tropical
11	26.5 °S–35.25 °S	113 °E–123.25° E	Temperate/grassland/desert

short samples (Coles, 2001; Kharin *et al.*, 2005), making L-moments the preferable method of parameter estimation for this study.

Kharin *et al.* (2007) note that the GEV distribution is valid only when extremes are drawn from increasingly larger samples. It is therefore important to determine whether the GEV distribution explains the nature of observed annual extremes reasonably, as the seasonal cycle within daily data substantially reduces the parent sample size (annual maxima does not occur during winter and annual minima does not occur during summer). Following Kharin *et al.* (2007), we used the Kolmogorov–Smirnov test (Stephens, 1970) to determine whether there was any substantial difference between the empirical and fitted CDF. We conducted the test for each grid box in the model's native resolution and also for each of the regions defined by Perkins *et al.* (2007). Overall, our results are similar to that of Kharin *et al.* (2007) in that the GEV distribution is a reasonable approximation for $T_{\min}$ and $T_{\max}$ at the 1% significance level.

Once the location, scale and shape parameters have been estimated, a CDF is produced and inverted to estimate the return value for the given return period. Our study focuses on 20 year return values for $T_{\min}$ ($T_{\min}{}^{20}$) and $T_{\max}$ ($T_{\max}{}^{20}$); we did not estimate changes in longer return values from the available 20 year data sets, given concerns over the small sample size (e.g. Kharin *et al.*, 2007).

To quantify in-sample uncertainty, the nonparametric bootstrap (Efron and Tibshirani, 1993) was employed. As there are no known analytical expressions for such information when calculating parameters by L-moments, 1000 bootstrap samples were generated for each model grid box at its native resolution. Return values were calculated for each sample to provide 90% bootstrap confidence intervals and estimates of standard errors.

2.3. Model evaluation

Three methods of model evaluation were used to assess each AR4 model's ability to simulate $T_{\max}$ and $T_{\min}$ in the current climate. All 12 regions defined by Perkins *et al.* (2007) were used in this study for each evaluation method (Figure 1 and Table II). The three evaluation methods were performed separately for $T_{\max}$ and $T_{\min}$ using observed data for 1981–2000. All models resolved multiple grid squares for each region. For each of the three evaluation methods outlined below, *two* ensembles were created, one consisting of the best (most skilled) models and one consisting of the poorest (least skilled) models. Division of the models into small samples allowed for the quantification of biases (if any) related to evaluation performance, which may be hampered when considering larger subsets. On division into the better and poorer ensembles, all models were assigned equal weightings since recent literature suggests the implausibility of finding the optimum weights, due to uncertainty in observations, models and the statistics used to evaluate them (Weigel *et al.*, 2010; Klocke *et al.*, 2011).

Our first validation method is the absolute difference between the annual 1981–2000 mean of a given model and the observed for each variable and region. While the mean may not be a good indicator of extreme values, it is still widely used for model validation.

Our second validation method is the skill score developed by Perkins *et al.* (2007). This calculates the cumulative minimum value of two distributions of each binned value, thereby measuring the common area between two PDFs (Figure 2). If a model simulates the observed conditions perfectly, the skill score will equal one, which is the total sum of the binned values in a given PDF:

$$\mathrm{PDF}_{\mathrm{score}} = \sum_{1}^{n} \min(Z_{\mathrm{m}}, Z_{\mathrm{o}}) \qquad (1)$$

where n is the number of bins used to calculate the PDF for a given region, Z_{m} is the proportion of values in a given bin from the model and Z_{o} is the proportion of values in a given bin from the observed data. Perkins

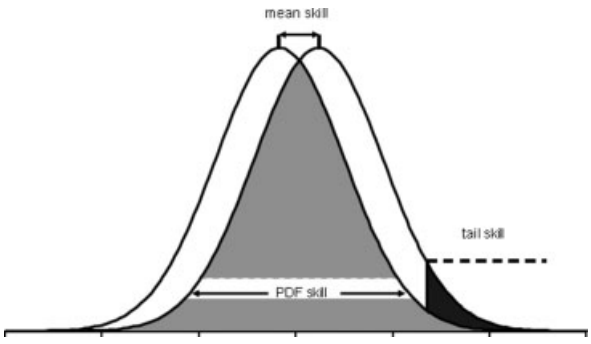

Figure 2. Schematic showing the regions of the PDF for each evaluation measure.

et al. (2007) found the skill score to be robust against data limitations (e.g. outliers and gaps in data) and to be a clear and straightforward way of comparing the entire modelled and observed data set. The skill score is also directly comparable across all variables and is easily interpreted.

The third validation method concentrates on the tail of the PDF (Figure 2, right tail for T_{max}, left tail for T_{min}). The 'tail-skill' focuses on the top (bottom) 5% for T_{max} (T_{min}), based on the observed PDF and is the weighted sum of absolute differences between the model and observed PDF proportions:

$$\mathrm{Tail}_{skill} = \sum_{i=1}^{n} W_i |Z_o^i - Z_m^i| \quad (2)$$

where n is the number of equally spaced bins, W_i is the specific weighting for bin i, Z_o is the frequency of values in a given bin for the observed data and Z_m is the frequency of values in a given bin for each model. All bins below (above) the observed 5% limit are weighted zero. The weighting above (below) the observed threshold is based on the number of bins in the observed tail, n_t. The weighting for bin i_t, where t is the number of the bin above (below) the threshold is $i_t \times 10/n_t$ for $i_t = 1 \ldots n_t$. The weighting is normalized is capped at bin n_t. If there are differences between the observed and modelled tails beyond where the observed tail ceases, the weighting is the same as bin n_t. The direction of the weighting increases to the right for T_{max} and to the left for T_{min}. In this method, a perfect skill equals zero (no difference between the observed and modelled tail) and poor skill scores exceed 1.0 (when the model tail is much larger than the observed and the model is over estimating the magnitude of the extreme values).

Models were ranked from highest to lowest for each region using each skill measure. For each variable, the best and poorest four models were selected to form the two ensembles based on the measure of skill. The change in 20 year return values estimated by the GEV distribution over Australia was examined using continental maps and regional analysis. Regional analysis is presented in 'stock plots' – these show the regionally calculated minimum and maximum for each ensemble based on skill and return value. Both the best (higher skilled models) and poorest (lower skilled models) ensembles are shown to demonstrate the influence the poorest models have over the all-model ensemble. Regions 1, 2, 3, 8, 10 and 11 defined by Perkins *et al.* (2007) are analysed for the twentieth century and A2 scenario, for 2050 and 2100. Figure 1 shows the location of these regions, and Table II lists their boundaries and climatic types. These regions were selected to cover a range of climate types over Australia and because of the larger observed and model sample sizes (stations and grid boxes, respectively).

3. Continental results

3.1. Maximum temperature

Figure 3 shows the A2 2050 20 year return values (T_{max}^{20}) for the all-model ensemble, the ensemble of the four best models for each skill-based ensemble, the change from the twentieth century and the bias compared with the all-model ensemble for each skill-based ensemble. Figure 3(a) shows T_{max}^{20} in the all-model ensemble exceeding 48 °C over large areas of the centre and northwest of Australia. The rest of the continent has return values of 46–48 °C, except for the majority of the coastline with return values of 40–44 °C. These return levels are 2–4 °C higher than the twentieth-century values (Figure 3(b)) over most of the continent.

Figure 3(c) shows the 20 year return values for the PDF-based ensemble (an ensemble over the best four models in each region). While the PDF-based ensemble projects increases in the 20 year return levels over twentieth-century levels (Figure 3(d)), the amount of increase is 1–4 °C less than the all-model ensemble over eastern Australia (Figure 3(e)), a substantial reduction in the magnitude of the return value. This result is largely mirrored for the tail-skill-based ensemble (Figure 3(f)). The mean-based ensemble (Figure 3(i)) is commonly 2–4 °C warmer than the all-model ensemble and 4–6 °C warmer over southeast Australia. In each case, the skill-selected models project increases in T_{max}^{20} due to increased CO_2 (Figure 3(d), (g) and (j)), but the increase is generally smaller in the PDF and tail-skill-selected models (Figure 3(e) and (h)) and larger in the mean skill-selected models (Figure 3(k)).

Figure 4 shows that by 2100, T_{max}^{20} may experience similar patterns of change as 2050, but at magnitudes 2 °C higher. The all-model ensemble projects increases in T_{max}^{20} of 4–6 °C over all regions. The PDF-based skill ensemble (Figure 4(c)) shows areas of 50–52 °C in central and west Australia, but overall, the amount of warming is projected to be similar or less compared with the all-model ensemble (Figure 4(e)), particularly in eastern regions. The tail-based skill ensemble also simulates less increase in T_{max}^{20} compared with the all-model ensemble over the southeast. The mean-based ensemble shows larger increases (Figure 4(k)) than the all-model ensemble over eastern and northern regions, but smaller increases in some parts of southern Australia. Thus, as with 2050, the PDF and tail-skill-selected models simulate lower increases in T_{max}^{20} compared with an all-model ensemble.

Figure 5 shows for selected regions (Figure 1 and Table II) the range in T_{max}^{20} and the 90% confidence intervals calculated from bootstrapped samples for *all* the ensembles over the twentieth century, 2050 and 2100. In virtually every case, for virtually every region, T_{max}^{20} estimated for the twentieth century by the models with better skill is smaller in magnitude and range that the poorer models (Figure 5(a) and (b)). The expansion in the simulated range shown by the all-model ensemble is therefore substantially caused by the inclusion of poorer

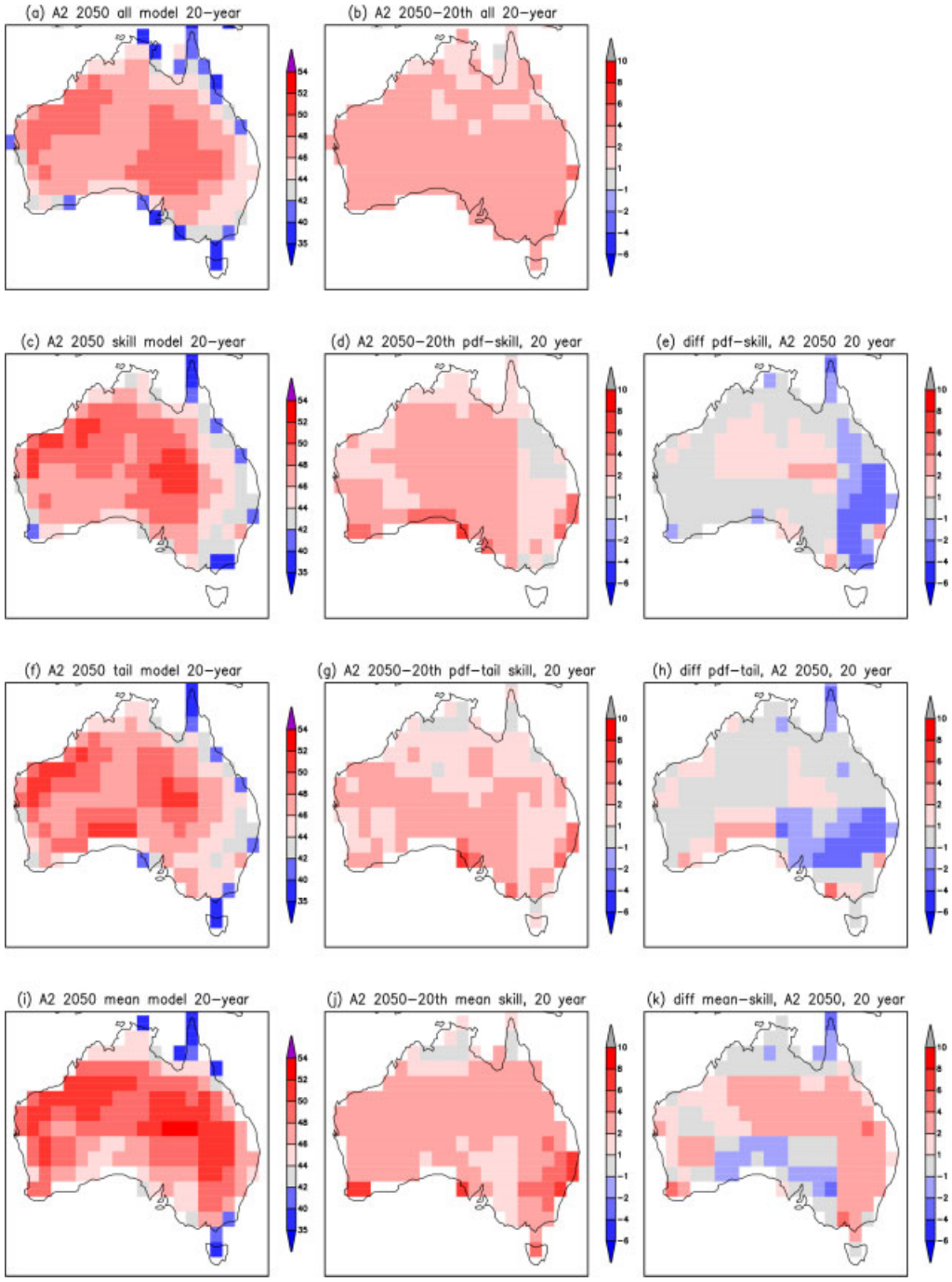

Figure 3. $T_{max}{}^{20}$ for the A2 emission scenario for 2050 for (a) actual values and (b) their change from the twentieth century for the all-model ensemble; (c) actual values from the PDF-based ensemble; (d) change from the twentieth century; (e) difference between the PDF and all-model ensembles; (f) actual values from the tail skill-based ensemble; (g) change from the twentieth century; (h) difference from the all-model ensemble; (i) actual values from the mean-based ensemble; (j) difference from the twentieth century and (k) difference from the all-model ensemble.

models and this is true irrespective of how model skill is determined. In the case of regions 1, 8, 10 and 11, the skill-based ensembles tend to populate the upper end of the all-model range. In regions 2 and 3, the skill-based ensembles tend to be more central or towards the bottom of the all model range. In Figure 5(a) and (b), the two samples for each skill score are not statistically significantly different since in all cases the uncertainty derived from bootstrapping (shown in Figure 5 as thin bars) overlap.

While the best and worst skill-based samples are not statistically significantly different during the twentieth century, Figure 5(c) and (d) demonstrates that there are examples where there is a clear distinction between the best and poorest skill-based ensembles for 2050 (e.g. regions 2 and 3). For other regions, the differentiation between the poorest and best models is not significantly different. However, the range of values projected by the poorest model ensembles is similar in scale to the all-model ensemble, while the range from the

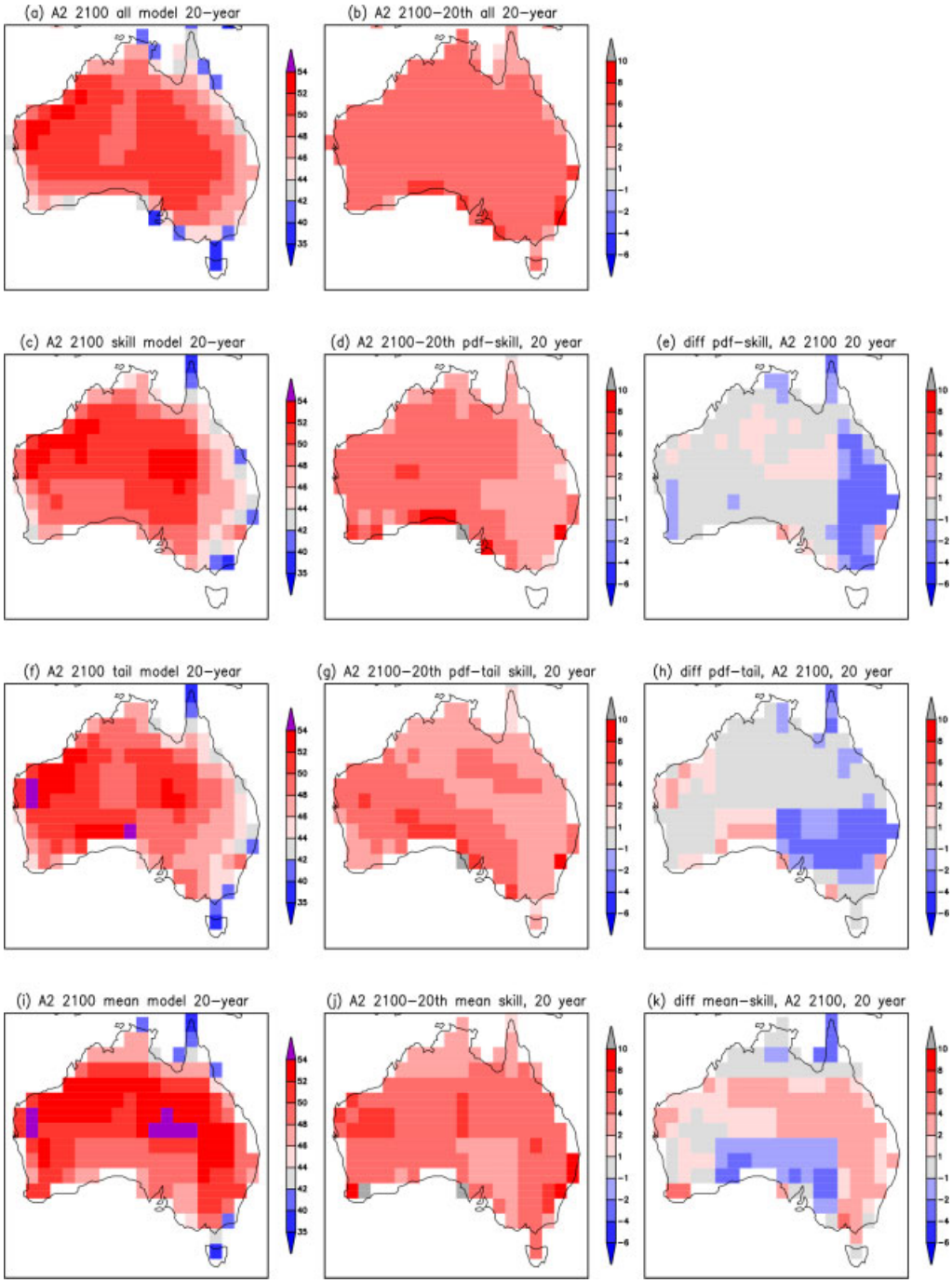

Figure 4. As Figure 3 but for 2100.

skill-selected ensembles tends to be much smaller and more consistent. The expansion in the poorest model ranges shown in Figure 5(c) and (d) compared with the best model ensemble is not related to sample size (which is common, four models per sample) and mainly occurs due to both the top and the bottom of the range being expanded. A noteworthy difference is visible when the mean skill is used. In all cases shown in Figure 5(c) and (d), the upper values of the projected all-model ensemble range are only simulated by the better models and while the poorer models show a larger range, this is due to the lowest values in the all-model ensemble being included at the same time as higher values being excluded. That is, only the better models determined by the mean capture the higher end of the projected range in $T_{\max}{}^{20}$. A similar result is seen for $T_{\max}{}^{20}$ by 2100 (Figure 5(e) and (f)).

3.2. Minimum temperature

The 20 year return level for $T_{\min}{}^{20}$ is shown in Figure 6 for 2050 and in Figure 7 for 2100. In Figure 6(a), the all-model ensemble shows $T_{\min}{}^{20}$ in the tropics reaching 10–20 °C with cooler values to the south. There is an area in the southeast with return values of between −1 and −3 °C associated with higher orography. These

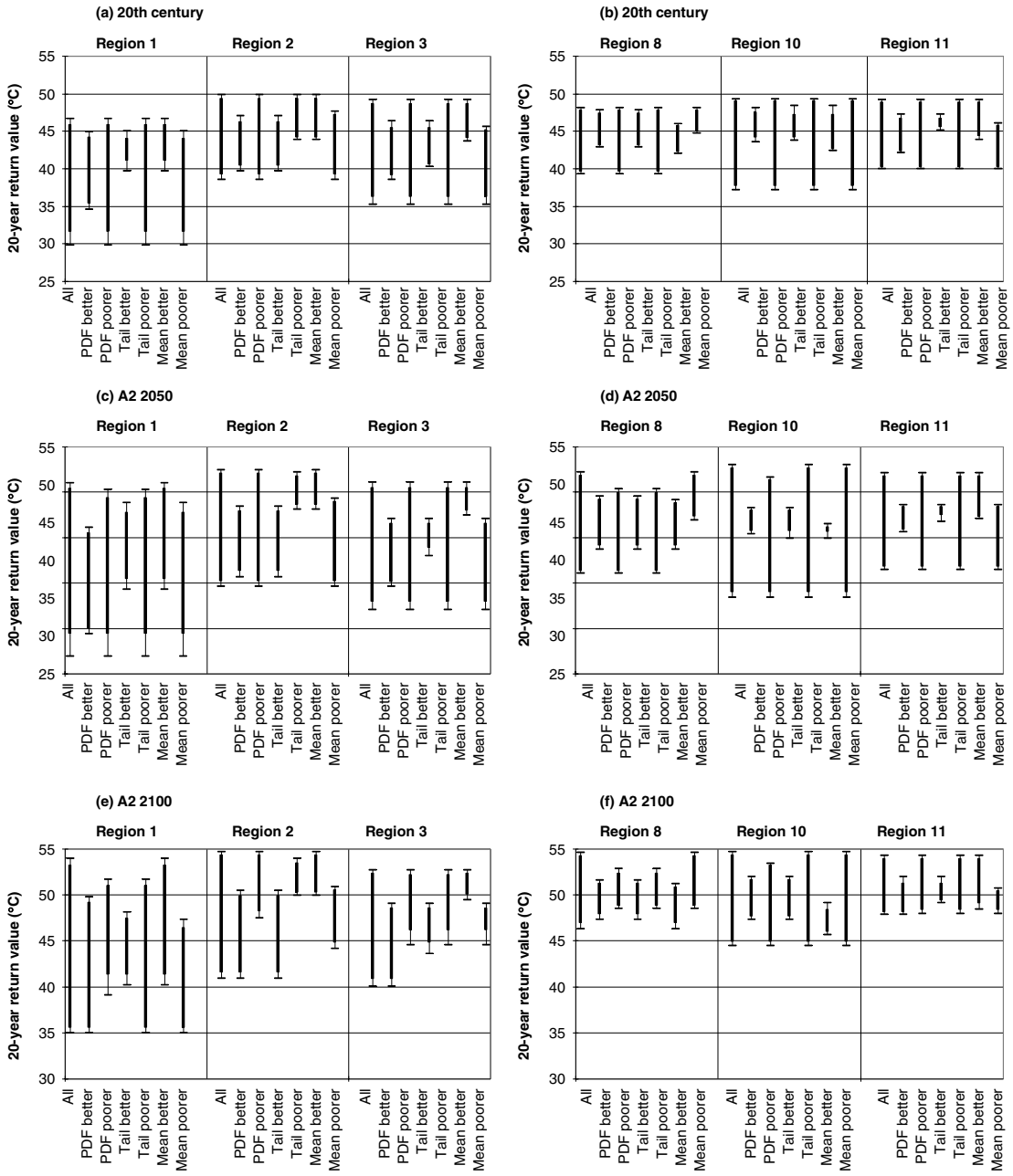

Figure 5. The simulated range and 90% bootstrapped confidence interval for all ensembles for $T_{max}{}^{20}$ for (a) regions 1, 2 and 3, (b) regions 8, 10 and 11 for the twentieth century; (c) and (d) shows the equivalent regions for 2050 and (e) and (f) 2100. For each panel/region, the first bar is the all-model ensemble with the bootstrapped uncertainty shown as thin 'error bars'. The second and third bars are the better and poorer models using the PDF skill score, the fourth and fifth bars are the better and poorer models from the tail skill score and the sixth and seventh bars are the better and poorer models using the mean skill. The models that make up each ensemble are listing in Table III.

return values are 1–2 °C warmer than the twentieth century over virtually the whole continent (Figure 6(b)), with the exception of the far southwest of Western Australia and Victoria. The PDF-based skill ensemble (Figure 6(c)) shows broadly similar patterns to the all-model ensemble but different absolute values. Return values are cooler over almost all the continent (Figure 6(e)) by 2–4 °C compared with the all-model ensemble but are still warmer than the twentieth century (Figure 6(d)). Similar results are found for the tail-based (Figure 6(f)) ensembles. The mean-based ensemble (Figure 6(i)) is warmer than the twentieth century (Figure 6(j)) by 1–4 °C and is warmer than the all-model ensemble over the central east and west Australia, while being 2–5 °C cooler over the Great Australian Bight (Figure 6(k)).

Figure 7 (for A2 2100) shows an analogous result to 2050. Every ensemble projects warming in $T_{min}{}^{20}$ by at least 1 °C over most of Australia and warming by 2–4 °C over many areas (Figure 7(b), (d), (g) and (j)). However, the PDF- and tail-skill-based ensembles projects less warming in $T_{min}{}^{20}$ than the all-model ensemble by 1–2 °C over most of Australia and by 2–4 °C over many areas (Figure 7(e) and (h)). Some areas in the west suggest higher increases in $T_{min}{}^{20}$ in the tail-based

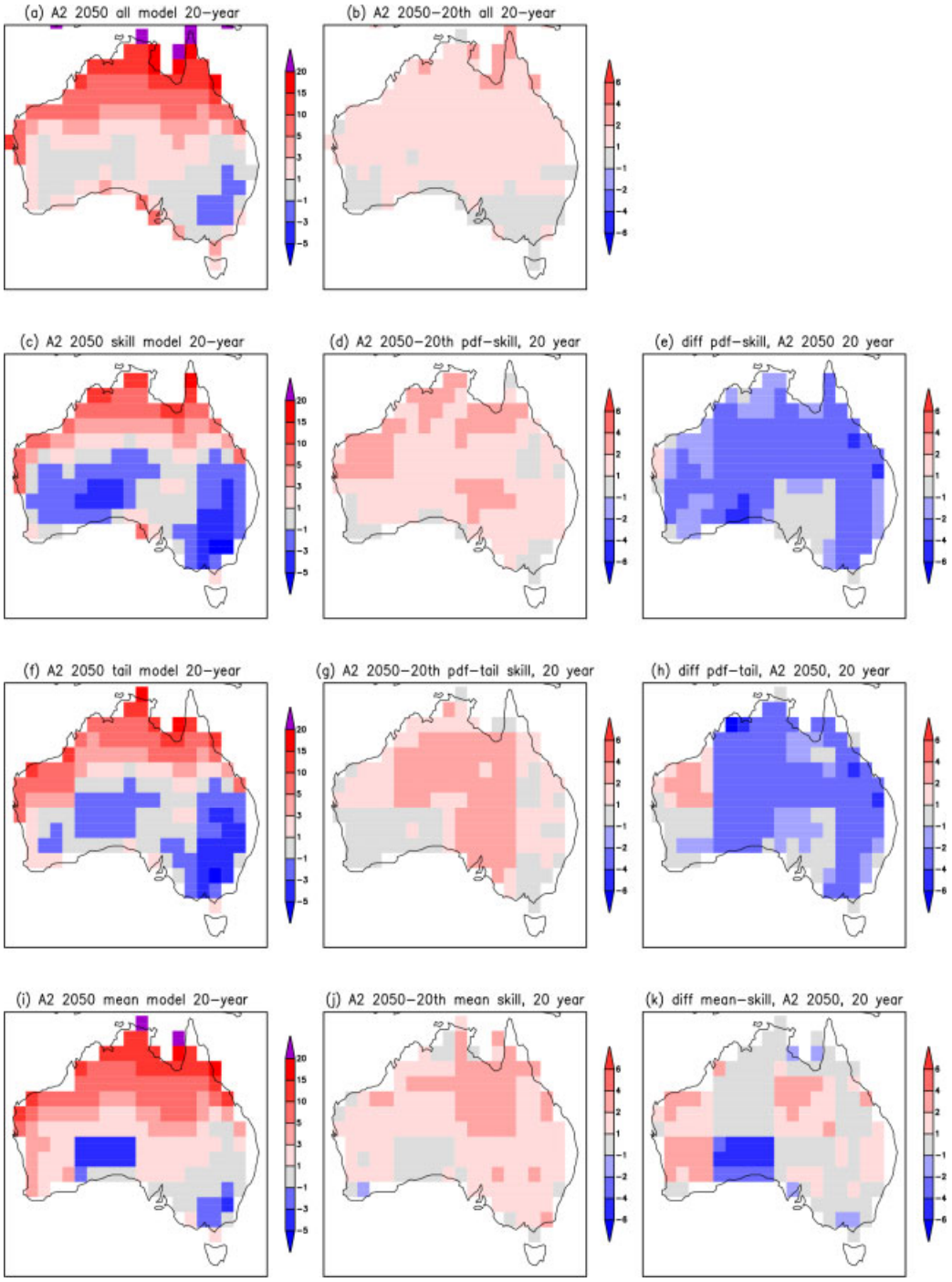

Figure 6. As Figure 3 but for T_{min}^{20}.

ensembles (Figure 7(h)). The mean-based ensemble simulates a similar change to the all-model ensemble except over Western Australia (Figure 7(k)) where higher values are projected.

Figure 8 shows for selected regions the range in the ensembles for the twentieth-century T_{min}^{20} and the 90% confidence intervals calculated from bootstrapped samples. The results are very different in comparison to T_{max}^{20} (Figure 5). For magnitudes T_{min}^{20}, the twentieth-century simulations by the models with better skill is commonly statistically significantly smaller than the poorer models in regions 1, 3 and 8. This is particularly clear for the PDF- and tail-based skill ensembles. The selection of models into the best and poorest ensembles tends to have an impact on the range in the resulting projections (the length of each bar is not noticeably smaller in the best or poorest ensemble).

Figure 8(c) and (d) (2050) and Figure 8(e) and (f) (2100) show clear distinctions between the 'better' and 'poorer' skill ensembles. In contrast to T_{max}^{20}, the differences are sometimes statistically significant. For example,

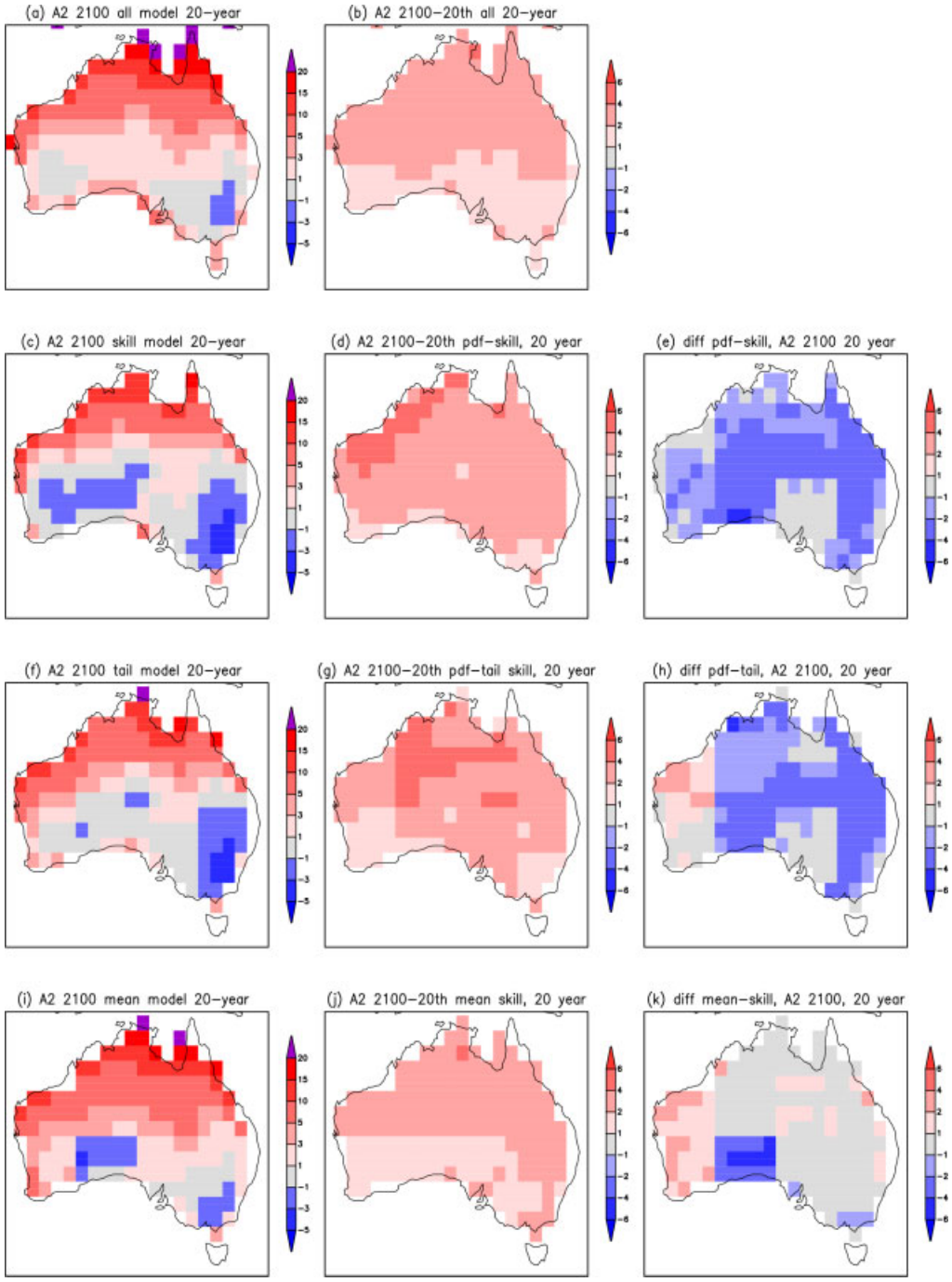

Figure 7. As Figure 4 but for $T_{\text{min}}{}^{20}$.

there are clear differentiations between the two ensembles in regions 3 and 8 in both 2050 and 2100.

4. Discussion

Global warming is already causing increases in warm nights and hot days in many regions (Alexander *et al.*, 2007; Brown *et al.*, 2008; Coelho *et al.*, 2008; Rusticucci and Tencer, 2008; Caesar *et al.*, 2010; Perkins, 2011). Observed trends in the twentieth century over Australia highlight warm temperature extremes increasing and cool extremes decreasing over most of the country (Plummer *et al.*, 1999; Collins *et al.*, 2000). Our results suggest that these trends will continue in the future under increasing atmospheric concentrations of greenhouse gases in agreement with previous global analyses (e.g. Kharin *et al.*, 2007). Our results are also complementary with that of Alexander and Arblaster (2009), who reported warming trends in temperature extremes over Australia under future emission scenarios and analyses at the annual timescale by Perkins and Pitman (2009) and Pitman and Perkins (2008).

Our analysis of the AR4 models therefore provides additional evidence that the 20 year minimum and

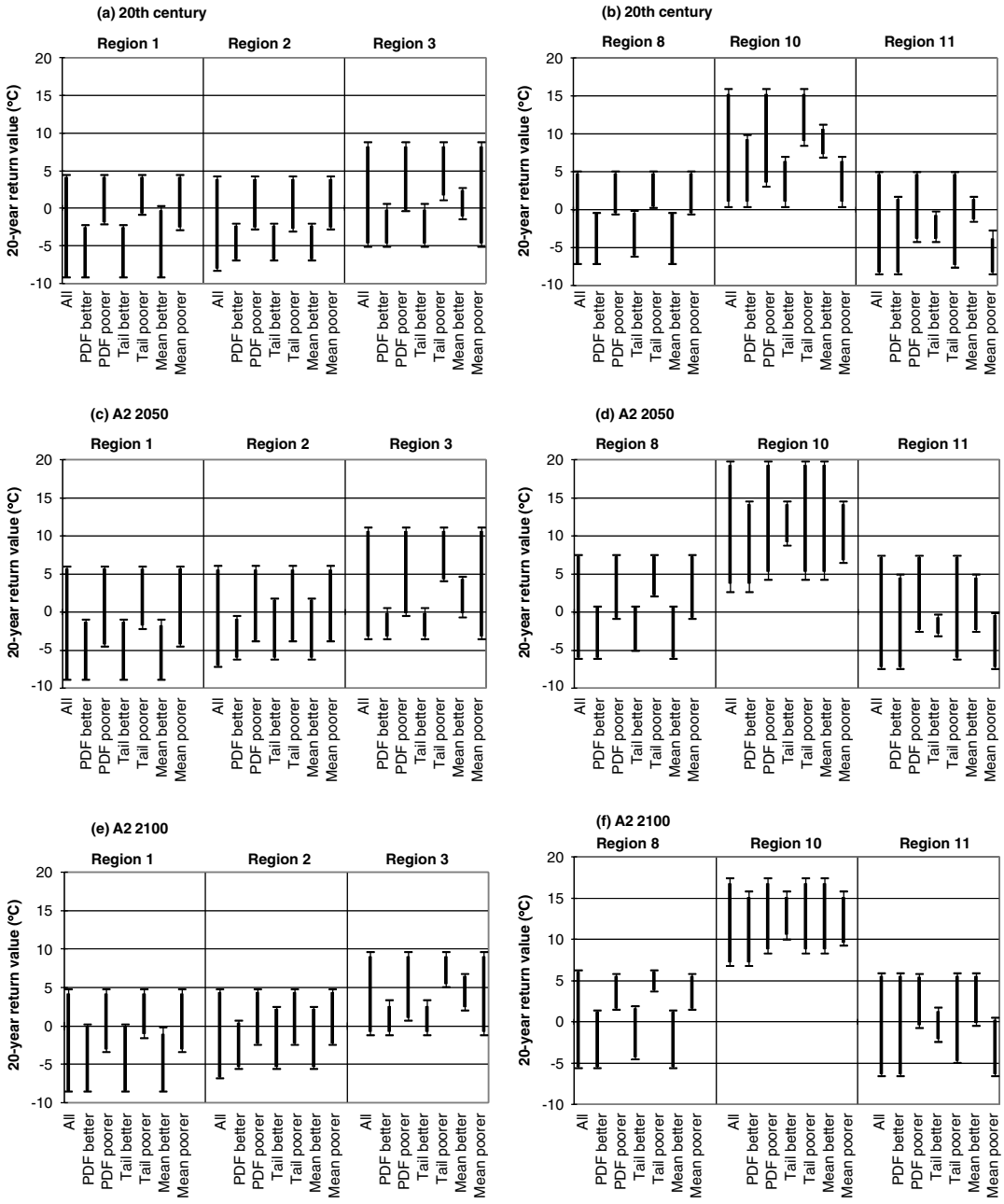

Figure 8. As Figure 5 but for $T_{min}{}^{20}$.

maximum temperatures will increase over Australia by 2050 and increase further by 2100. Although increases in $T_{max}{}^{20}$ and $T_{min}{}^{20}$ are projected by *all* ensembles presented in this article, our results show a systematic reduction in the amount of warming in $T_{max}{}^{20}$ and $T_{min}{}^{20}$, and a smaller range, in skill-selected ensembles if a PDF or tail-based skill score compared with an all-model ensemble, particularly over the eastern part of the continent (Figures 3(e), 3h, 4(e), 4(h), 6(e), 6(h), 7(e) and 7(h)). If a means-based skill score is used, higher amounts of projected warming are commonly simulated (Figures 3(k), 4(k), 6(k) and 7(k)). The amount of warming in $T_{max}{}^{20}$ is commonly 2–4 °C less over eastern states than an all-model ensemble (Figures 3 and 4) compared with the all-model ensemble. Similar differences in the projected changes in $T_{min}{}^{20}$ occur in the skill-selected AR4 models, but the impact is more geographically distributed. It was shown (Figure 5) that the better models formed a statistically significantly different population in $T_{max}{}^{20}$ in some regions in both 2050 and 2100 (Figure 5). Figure 8 showed that the better models tended to provide the lower values in the total distribution for $T_{min}{}^{20}$, and the results were commonly statistically significant. Our results therefore suggest that over Australia, a PDF- or tail-based ensemble commonly projects smaller increases in both $T_{max}{}^{20}$ and $T_{min}{}^{20}$ than an all-model ensemble. This leads to three questions: Is the way model skill is measured important? Do common models populate the better and poorer ensembles determined by each skill metric? And is this consistent over Australia?

Table III. Frequency of models appearing in the better and poorer ensembles based on each skill score for T_{max}^{20} and (in parentheses) T_{min}^{20}.

	Better ensemble			Poorer ensemble		
	PDF-based skill	Tail skill	Mean	PDF-based skill	Tail skill	Mean
CGCM	0 (4)	0 (7)	6 (2)	12 (5)	12 (1)	5 (7)
CSIRO3.0	2 (5)	4 (5)	2 (2)	9 (1)	3 (3)	7 (6)
ECHO-G	4 (4)	9 (2)	6 (8)	0 (0)	1 (6)	2 (1)
IPSL	5 (3)	1 (5)	7 (6)	2 (4)	5 (3)	1 (3)
MIROC	8 (4)	9 (2)	6 (8)	1 (0)	1 (6)	2 (1)
MRI	1 (7)	1 (5)	5 (7)	9 (1)	10 (3)	5 (4)
BCCR	0 (0)	0 (1)	0 (4)	11 (10)	11 (10)	11 (7)
GFDL2.0	10 (12)	10 (8)	4 (2)	0 (0)	0 (3)	5 (7)
GFDL2.1	10 (10)	10 (10)	2 (6)	0 (1)	0 (0)	5 (4)
ECHAM	6 (3)	5 (4)	5 (1)	0 (4)	2 (4)	1 (6)
CSIRO3.5	2 (0)	5 (1)	6 (6)	4 (10)	2 (3)	4 (1)

Our results suggest that the measure of skill used in model selections is important. Over Australia, the differences between the three skill-selected ensembles are large (recall, the differences are in extreme values, not the mean). Omitting the poorest models using a distribution-based measure has a systematic impact on the ensemble projections by removing the poorest models that are consistently relatively more sensitive to increasing CO_2 over Australia than the better models. Using a means-based measure systematically removes models with a relatively low sensitivity. This is clearly seen in Figures 5 and 8, where the poorer models project larger increases in return values and the better models project smaller increases. Thus, the measure of climate model performance is important, and our results suggest that a PDF- or tail-based measure is preferable to the mean.

The second question – which models make up the more skilful ensembles – informs us on the consistency of the three measures. A 'poor' model should ideally appear poor in a range of measures; a 'best' model should ideally not be omitted by one measure and included in another. Table III shows the frequency (out of 12) of a given model forming part of the best and poorest ensembles for T_{max}^{20}. Given the three measures used assess very different statistics of temperature (Figure 2), some variability in models that score well against each measure is to be expected.

Some models are almost always in the best ensemble for T_{max}^{20} using the PDF-based skill score or the tail-based skill (GFDL2.0, GFDL2.1 and MIROC). Some models are almost always in the poorest ensemble, region-by-region over Australia (e.g. CGCM, BCCR and MRI). The results for T_{max}^{20} are therefore explained by omission of CGCM, BCCR and MRI and the inclusion of GFDL2.0, GFDL2.1 and MIROC in the skilled ensembles and the recognition that these better models simulate a lower sensitivity to increasing CO_2 in terms of increasing T_{max}^{20} over most regions of Australia. Table III also shows the equivalent result for T_{min}^{20}. In this case, GFDL2.0, GFDL2.1 and MRI typically form the skilful ensemble, while ECHO-G and CSIRO3.5 are commonly in the poorer ensembles. The models that appear best/poorest in the means-based measure are not as systematic. Only one model (BCCR) never appears in the best ensemble for any region (Table III). While GFDL2.0 and GFDL2.1 stand out as particularly good in the PDF- and tail-based measures, no model stands out as particularly good in the means-based measure. That is, as an evaluation measure, the mean does not appear to discriminate between models well.

There have been several papers that have noted that skill in simulating the present may be a weak guide to the reliability in the future (e.g. Räisänen, 2007; Weigel *et al*., 2010; Klocke *et al*., 2011). To the best of our knowledge, all analyses relating skill in the present to reliability in the future have used mean-based performance measures (Räisänen, 2007; Jun *et al*., 2008) sometimes on decadal or longer timescales (e.g. Reifen and Toumi, 2009) and there are contrasting results suggesting that specific models retain skill decade-by-decade through the twentieth century (Macadam *et al*., 2010). However, as implied by Räisänen (2007), Jun *et al*. (2008) and Reifen and Toumi (2009), there is no clear relationship between mean skill in simulating the present and the amount of warming in the 20 year return values simulated by the models in the results presented here. Perkins and Pitman (2009) discussed the merits of the PDF-based skill scores in terms of the high overlap of the PDF representing 2100 with the present-day PDF. Even under high global warming, it was noted that the present and future PDFs would overlap considerably, and it could be inferred that for when this overlap exists, broadly similar physical climates observed in the present will exist in the future. A climate model able to simulate the whole of the present-day PDF has therefore shown a capacity to simulate a useful amount of a future PDF, to the degree that the present and future PDFs overlap. This finding is consistent across the Australian continent, thereby answering the third question stated above.

Overall, the individual models that make up the better (i.e. more skilful) and poorer (i.e. less skilful) ensembles

are similar for all regions (Tables II and III), which means there are very limited inconsistencies in our methodology, region-by-region, across Australia. There is, however, some inconsistency between the models that perform well for $T_{max}{}^{20}$ compared with $T_{min}{}^{20}$ (e.g. MRI). This is not surprising since the simulation of $T_{max}{}^{20}$ requires clouds, incoming solar radiation, albedo, aerosols and the partitioning of net radiation between sensible and latent heat fluxes to be captured well. Jones and Trewin (2000) highlight the role of radiative and latent heat interactions in explaining maximum temperature variations over Australia. Variations in any processes that control the supply of water for evaporation (rainfall, soil moisture, root distribution and stomatal conductance; Pitman, 2003) can affect evaporative cooling and therefore maximum temperatures (Collatz *et al*., 2000). Furthermore, extreme maxima over (particularly eastern) Australia are commonly related to drought, which are linked to large-scale teleconnection patterns such as the Indian Ocean Dipole (IOD; Ummenhofer *et al*., 2009) and the El Niño/Southern Oscillation (ENSO; Jones and Trewin, 2000; Nicholls, 2004). Ummenhofer *et al*. (2009) found that extreme droughts over southern Australia have only occurred during the positive and neutral phases of IOD, while the El Niño phase of ENSO is associated with below-average rainfall for the eastern two thirds of the country. Such conditions are linked to high land surface temperatures. Thus, to simulate the change in $T_{max}{}^{20}$ requires many processes to be captured well ranging from local surface energy balance through regional scale advection of heat and moisture through to large-scale coupling of the ocean-atmosphere system at low-frequency timescales.

Drivers of changes in $T_{min}{}^{20}$ can also be associated with large-scale ocean-atmosphere coupling as changes in rainfall also affect minimum temperatures. However, while annual minima can commonly be explained through changes in cloud cover and associated increases in infrared loss, extreme minima tend to be associated with outbreaks of Antarctic air masses that can affect minimum temperatures as far north as the subtropics in eastern Australia.

There are some unavoidable caveats given the nature of climate models, and the data available. First, present-day skill in calculating the mean, PDF or the tail of the PDF may not reliably guide how well the models can simulate future changes in the 20 year return levels. A second caveat relates to sample size and model independence. Twenty years of daily data from 11 AR4 climate models is a small sample size. Omitting poorest models reduces the sample size; where two or three models are omitted, this may not be a problem, but where only two or three models are included, conclusions should be treated with caution. This is particularly important, given the issues around model independence (Abramowitz and Gupta, 2008; Jun *et al*., 2008).

5. Conclusions

Temperature extremes have a large impact on many human, industrial and biophysical systems. Earlier studies using EVT consistently show increases in extreme temperatures in the future under higher atmospheric concentrations of CO_2. Most of these studies explore these changes via multimodel ensembles (e.g. Kharin *et al*., 2007) using all available climate model results.

In this article, we utilize the approach of Kharin *et al*. (2007) but with a prior step; we evaluate each AR4 climate model over Australia using three methods that test performance against mean, the PDF and the tail of the PDF – all derived using daily climate model data and compared with daily observational data. This is important as most temperature-related extremes occur over several days and it is difficult to test the skill of a model on this timescale using monthly averages.

Our results show for T_{max} and T_{min} over Australia that regardless of the evaluation procedure considered, the poorest performing models project larger increases in the 20 year return values than the best performing models. Thus, an all-model ensemble, used most commonly in the literature, is biased towards overestimating the amount of increase in the 20 year return values at regional and continental scales for both 2050 and 2100. Models that performed relatively poorly in representing the observed climate were generally consistently poor for most regions (CGCM, BCCR and MRI for T_{max} and CSIRO3.5 and ECHO-G and BCCR for T_{min}). The best-performing models were generally commonly best performing for most regions for T_{max} (GFDL2.0, GFDL2.1 and MIROC) and for T_{min} (GFDL2.0, GFDL2.1 and MRI).

The better skill-based ensembles project 20 year return values for T_{min} and T_{max} that are 2–4 °C cooler than the all-model ensemble at the continental scale and were commonly at the lower end of the all-model ensemble range at the regional scale. For T_{max}, the confidence intervals for some regions for the better and poorer ensembles do not overlap, such that the range of 20 year return values in the two ensembles for the respective measure of skill are statistically significantly different. This also occurs for T_{min}, the ranges of the poorer ensembles are commonly at the higher end of the all-model range.

We emphasize some limitations in our methodology – in particular, sample size and concerns over how independent the AR4 models are from each other, whether there are systematic biases to all models, whether skill in modelling the present is a guide to predictive skill and how to manage the situation of the better models for T_{min} being the worst for T_{max} and vice versa. However, with these caveats in mind, we note that an all-model ensemble is biased over Australia by specific poor models and that excluding these reduces the amount of projected warming in $T_{max}{}^{20}$ and $T_{min}{}^{20}$. We suggest our results reinforce the case for excluding demonstrably poor models from ensembles – noting that this analysis is necessarily regionally specific. We also note that our

results support earlier analyses that suggest that a good performance by a model against observed averages may not provide confidence that the model will perform well in the future. However, we provide some evidence that a PDF- or tail-skill-based measure is a more rigorous and valuable measure to evaluate climate models since this appears to provide a means to more consistently discriminate between model performances.

Acknowledgements

We acknowledge the international modelling groups for providing their data for analysis, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) for collecting and archiving the model data, the JSC/CLIVAR Working Group on Coupled Modeling (WGCM) and their Coupled Model Intercomparison Project (CMIP) and Climate Simulation Panel for organizing the model data analysis activity, and the IPCC WG1 TSU for technical support. The IPCC Data Archive at Lawrence Livermore National Laboratory is supported by the Office of Science, U.S. Department of Energy. SAS is supported by the Australian Research Council (DP0877432). This research was partially supported by the Australian Research Council Centre of Excellence for Climate System Science (CE110001028).

References

Abramowitz G, Gupta H. 2008. Toward a model space and model independence metric. *Geophysical Research Letters* **35**: L05705, DOI: 10.1029/2007GL032834.

Alexander LV, Arblaster JM. 2009. Assessing trends in observed and modelled climate extremes over Australia in relation to future projections. *International Journal of Climatology* **29**: 417–435, DOI: 10.1002/joc.1730.

Alexander LV, Hope P, Collins D, Trewin B, Lynch A, Nicholls N. 2007. Trends in Australia's climate means and extremes: a global context. *Australian Meteorological Magazine* **56**: 1–18.

Allen MR, Stott PA, Mitchell JFB, Schnur R, Delworth TL. 2003. Quantifying the uncertainty in forecasts of anthropogenic change. *Nature* **407**: 617–620.

Brown SJ, Caesar J, Ferro CAT. 2008. Global changes in extreme daily temperature since 1950. *Journal of Geophysical Research* **113**: D05115, DOI: 10.1029/2006JD008091.

Caesar J, Alexander LV, Trewin B, Tse-ring K, Sorany L, Vuniyayawa V, Keosavang N, Shimana A, Htay MM, Karmacharya J, Jayasinghearachchi DA, Sakkamart J, Soares E, Hung LT, Thuong LT, Hue CT, Dung NTT, Hung PV, Cuong HV, Cuong NM, Sirabaha S. 2010. Changes in temperature and precipitation extremes over the Indo-Pacific region from 1971 to 2005. *International Journal of Climatology* **36**: 791–801, DOI: 10.1002/joc.2118.

Coelho CAS, Ferro CAT, Stephenson DB, Steinskog DJ. 2008. Methods for exploring spatial and temporal variability of extreme events in climate date. *Journal of Climate* **21**: 2072–2092, DOI: 10.1175/2007JCLI1781.1.

Coles SG. 2001. *An Introduction to Statistical Modeling of Extreme Values*. Springer: London 225pp.

Collatz GJ, Bounoua L, Los SO, Randall DA, Fung IY, Sellers PJ. 2000. A mechanism for the influence of vegetation on the response of the diurnal temperature range to a changing climate. *Geophysical Research Letters* **27**: 3381–3384.

Collins DA, Della-Marta PM, Plummer N, Trewin BC. 2000. Trends in annual frequencies of extremes temperature events in Australia. *Australian Meteorological Magazine* **49**: 277–292.

Colombo A, Etkin D, Karney B. 1999. Climate variability and the frequency of extreme temperature events for nine sites across Canada: implications for power usage. *Journal of Climate* **12**: 2490–2502, DOI: 10.1175/15200442(1999)012<2490:CVATFO>2.0.CO;2.

Easterling DR, Meehl GA, Parmesan C, Changnon SA, Karl TR, Mearns LO. 2000. Climate extremes: observations, modeling, and impacts. *Science* **289**: 2068–2074.

Efron B, Tibshirani R. 1993. *An Introduction to the Bootstrap*. Chapman and Hall: New York 436pp.

Fischer EM, Schar S. 2010. Consitent geographical patterns of changes in high-impact European heatwaves. *Nature Geoscience* **3**: 398–403, DOI: 10.1038/NGEO866.

Fowler HJ, Cooley D, Sain SR, Thurston M. 2010. Detecting change in UK extreme precipitation using results from the climateprediction.net BBC climate change experiment. *Extremes* **13**: 241–267.

Gallant AE, Karoly DJ. 2010. A combined climate extremes index for the Australian region. *Journal of Climate* **23**: 6153–6165, DOI: 10.1175/2010JCLI3791.1.

Hegerl GC, Zwiers FW, Stott PA, Kharin VV. 2004. Detectability of anthropogenic changes in annual temperature and precipitation extremes. *Journal of Climate* **17**: 3683–3700, DOI: 10.1175/1520-0442(2004)017<3683:DOACIA>2.0.CO;2.

Irving DB, Perkins SE, Brown JR, Sen Gupta A, Moise AF, Murphy BF, Colman RA, Power SB, Delage FB, Brown JN. 2011. Evaluating global climate models for the Pacific region. *Climate Research* **49**: 169–187.

Jones DA, Trewin B. 2000. The spatial structure of monthly temperature anomalies over Australia. *Australian Meteorological Magazine* **49**: 261–276.

Jun M, Knutti R, Nychka DW. 2008. Spatial analysis to quantify numerical model bias and dependence: how many climate models are there?. *Journal of the American Statistical Association* **103**(483):. DOI: 10.1198/016214507000001265.

Katz R, Brown B. 1992. Extreme events in a changing climate: variability is more important than averages. *Climate Change* **21**: 289–302.

Kharin V, Zwiers F. 2000. Changes in the extremes in an ensemble of transient climate simulations with a coupled atmosphere-ocean GCM. *Journal of Climate* **13**: 3760–3788, DOI: 10.1175/1520-0442(2000)013<3760:CITEIA>2.0.CO;2.

Kharin V, Zwiers F, Zhang X. 2005. Intercomparison of near surface temperature and precipitation extremes in AMIP-2 simulations. *Journal of Climate* **18**: 5201–5223, DOI: 10.1175/JCLI3597.1.

Kharin VV, Zwiers FW, Zhang X, Hegerl GC. 2007. Changes in temperature and precipitation extremes in the IPCC ensemble of global couple model simulations. *Journal of Climate* **20**: 1419–1444, DOI: 10.1175/JCLI4066.1.

Klocke D, Pincus R, Quaas J. 2011. On constraining estimates of climate sensitivity with present-day observations through model weighting. *Journal of Climate* **24**: 6092–6098, DOI: 10.1175/2011JCLI4193.1.

Knutti R, Abramowitz G, Collins M, Eyring V, Glecker PJ, Hewiston B, Mearns L. 2010. Good practice guidance paper on assessing and combining multi model climate projections. In *Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections*, Stocker TF, Qin D, Plattner G-K, Tignor M, Midgley PM, (eds). IPCC Working Group I Technical Support Unit, University of Bern: Bern, Switzerland.

Macadam I, Pitman AJ, Whetton PH, Abramowitz G. 2010. Ranking climate models by performance using actual values and anomalies: implications for climate change impact assessments. *Geophysical Research Letters* **37**: L16704, DOI: 10.1029/2010GL043877.

Mearns LO, Katz R, Schneider S. 1984. Extreme high-temperature events: changes in the probabilities with changes in mean temperature. *Journal of Applied Meteorology* **23**: 1601–1613.

Meehl G, Zwiers F, Evans J, Knutson T, Mearns L, Whetton P. 2000. Trends in extreme weather and climate events: issues related to modelling extremes in projections of future climate change. *Bulletin of the American Meteorological Society* **8**: 427–436.

Moise AF, Hudson DA. 2008. Probabilistic predictions of climate change for Australia and southern Africa using reliability ensemble average of IPCC CMIP3 model simulations. *Journal of Geophysical Research* **113**: D15113, DOI: 10.1029/2007JD009250.

Nicholls N. 2004. The changing nature of Australian droughts. *Climatic Change* **63**: 323–336.

Perkins SE. 2011. Biases and model agreement in the projections of climate extremes over the tropical Pacific. *Earth Interactions* **15**: 1–36, DOI: 10.1175/2011EI395.1.

Perkins SE, Pitman AJ. 2009. Do weak AR4 models bias projections of future climate changes over Australia? *Climatic Change* **93**: 527–558, DOI: 10.1007/s10584-008-9502-1.

Perkins SE, Pitman AJ, Holbrook NJ, McAneney J. 2007. Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature and precipitation over Australia using probability density functions. *Journal of Climate* **20**: 4356–4376, DOI: 10.1175/JCLI4253.1.

Pitman AJ. 2003. The evolution of, and revolution in, land surface schemes designed for climate models. *International Journal of Climatology* **23**: 479–510.

Pitman AJ, Perkins SE. 2008. Regional projections of future seasonal and annual changes in rainfall and temperature over Australia based on skill-selected AR4 models. *Earth Interactions* **12**: 1–50.

Plummer N, Slinger MJ, Nicholls N, Suppiah R, Hennessy KJ, Leighton RM, Trewin B, Page CM, Lough JM. 1999. Changes in climate extremes over the Australian region and New Zealand during the twentieth century. *Climatic Change* **42**: 183–202.

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

Reifen C, Toumi R. 2009. Climate projections: past performance no guarantee of future skill? *Geophysical Research Letters* **36**: L13704, DOI: 10.1029/2009GL038082.

Rusticucci M, Tencer B. 2008. Observed changes in return values of annual temperature extremes over Argentina. *Journal of Climate* **21**: 5455–5467, DOI: 10.1175/1520-0442(2004)017<4099:OTACIT>2.0.CO;2.

Schaeffer M, Selten FM, Opsteegh JD. 2005. Shifts in means are not a proxy for changes in extreme winter temperatures in climate projections. *Climate Dynamics* **25**: 51–63.

Smith JB, Schneider SH, Oppenheimer M, Yohe GW, Har W, Mastrandrea MD, Patwardhan A, Burton I, Corfee-Morlot J, Magazda CHD, Fûssel H-M, Pittock BA, Rahman A, Suarez A, van Ypersele J-P. 2009. Assessing dangerous climate change though an update of the Intergovernmental Panel on Climate Change (IPCC) "reasons for concern". *PNAS* **106**: 4133–4137.

Stainforth DA, Allen MR, Tredger ER, Smith LA. 2007. Confidence, uncertainty and decision-support relevance in climate predictions. *Philosophical Transitions of the Royal Society of London* **365**: 2145–2161.

Stephens MA. 1970. Use of the Kolmogorov–Smirnov, Cramer von-Mises and related statistics without extensive tables. *Royal Statistical Society (Series A)* **32B**: 115–122.

Sterl A, Severijns C, Dijkstra H, Hazeleger W, van Oldenborgh GJ, van den Brok M, Bugers G, van den Hurk D, van Leeuwen PJ, van Velthoven P. 2008. When can we expect extremely high surface temperatures? *Geophysical Research Letters* **35**: L14703, DOI: 10.1029/2008GL034071.

Tebaldi C, Hayhoe K, Arblaster JM, Meehl GA. 2006. Going to the extremes, an intercomparison of model-simulated historical and future changes in extremes events. *Climatic Change* **79**: 185–211.

Ummenhofer CC, England MH, McIntosh PC, Meyers GA, Pook MJ, Risbey JS, Sen Gupta A, Taschetto AS. 2009. What causes Southeast Australia's worst droughts? *Geophysical Research Letters* **36**: L04706, DOI: 10.1029/2008GL03680.

Wehner MF. 2004. Predicted twenty-first-century changes in seasonal extreme precipitation events in the parallel climate model. *Journal of Climate* **17**: 4281–4290, DOI: 10.1175/JCLI3197.1.

Wehner MF, Smith RL, Bala G, Duffy P. 2010. The effect of horizontal resolution on simulation of very extreme US precipitation events in a global atmosphere model. *Climate Dynamics* **34**: 241–247.

Weigel AP, Knutti R, Liniger MA, Appenzeller C. 2010. Risks of model weighting in multi model climate projections. *Journal of Climate* **27**: DOI: 10.1175/2010JCLI3594.1.

Zwiers F, Kharin V. 1998. Changes in the extremes of climate simulated by CCC GCM2 under CO_2 doubling. *Journal of Climate* **11**: 2200–2222, DOI: 10.1175/1520-0442(1998)011<2200:CITEOT>2.0.CO;2.

