Evaluation of Extreme Temperatures over Australia in the Historical Simulations of CMIP5 and CMIP6 Models

Xu Deng[1], Sarah Perkins-Kirkpatrick[1], Sophie Caroline Lewis[2], and Elizabeth A. Ritchie[3]

[1]University of New South Wales
[2]University of New South Wales at ADFA
[3]PEMS, UNSW Canberra

November 23, 2022

Abstract

Historical simulations of models participating in the 6th phase of the Coupled Model Intercomparison Project (CMIP6) are evaluated over ten Australian regions for their performance in simulating extreme temperatures. Based on two observational datasets, the Australian Water Availability Project (AWAP) and the Berkeley Earth Surface Temperatures (BEST), we first analyze the models' abilities in simulating the probability distributions of daily maximum and minimum temperature (TX and TN), followed by the spatial patterns and temporal variations of temperature-related extreme indices, as defined by the Expert Team on Climate Change Detection and Indices (ETCCDI). Overall, the CMIP6 models are comparable to CMIP5, with modest improvements shown in CMIP6. Compared to CMIP5, the CMIP6 ensemble tends to have narrower interquartile model ranges for some cold extremes, as well as narrower ensemble ranges in temporal trends for most indices. Over southeast, tropical and southern south regions, both CMIP ensembles generally exhibit relatively large deficiencies in simulating temperature extremes. It is also noted that models with relatively coarse resolution sometimes show better performance, suggesting that some localized processes may need further improvement in finer-scale models. With the assessment on the probability distributions of TX and TN, the results of this study provide more robustness on the evaluation of extreme temperatures and more confidence on future projections. The findings of this study demonstrate only incremental improvement on the simulation of extremes over Australia from CMIP5 to CMIP6. However, they are useful in informing and interpreting future projections of temperature-related extremes over the region.

Evaluation of Extreme Temperatures over Australia in the Historical Simulations of CMIP5 and CMIP6 Models

Xu Deng[1,2*], Sarah E. Perkins-Kirkpatrick[2,3], Sophie C. Lewis[1], and Elizabeth A. Ritchie[1]

[1]School of Science, University of New South Wales, Canberra, ACT, Australia.

[2]ARC Centre of Excellence for Climate Extremes, University of New South Wales, Sydney, NSW, Australia.

[3]Climate Change Research Centre, University of New South Wales, Sydney, NSW, Australia.

Corresponding author: Xu Deng (xu.deng@student.adfa.edu.au)

Key Points:

- The assessment on the probability distributions of daily maximum/minimum temperature makes the evlatuion on temperature extremes more robust
- Temperature extremes over Australia are broadly similar in CMIP5 and CMIP6
- The CMIP6 ensemble exhibits narrower ensemble ranges in temporal trends for most extreme temperature indices compared to CMIP5

Abstract

Historical simulations of models participating in the 6th phase of the Coupled Model Intercomparison Project (CMIP6) are evaluated over ten Australian regions for their performance in simulating extreme temperatures. Based on two observational datasets, the Australian Water Availability Project (AWAP) and the Berkeley Earth Surface Temperatures (BEST), we first analyze the models' abilities in simulating the probability distributions of daily maximum and minimum temperature (TX and TN), followed by the spatial patterns and temporal variations of temperature-related extreme indices, as defined by the Expert Team on Climate Change Detection and Indices (ETCCDI). Overall, the CMIP6 models are comparable to CMIP5, with modest improvements shown in CMIP6. Compared to CMIP5, the CMIP6 ensemble tends to have narrower interquartile model ranges for some cold extremes, as well as narrower ensemble ranges in temporal trends for most indices. Over southeast, tropical and southern south regions, both CMIP ensembles generally exhibit relatively large deficiencies in simulating temperature extremes. It is also noted that models with relatively coarse resolution sometimes show better performance, suggesting that some localized processes may need further improvement in finer-scale models. With the assessment on the probability distributions of TX and TN, the results of this study provide more robustness on the evaluation of extreme temperatures and more confidence on future projections. The findings of this study demonstrate only incremental improvement on the simulation of extremes over Australia from CMIP5 to CMIP6. However, they are useful in informing and interpreting future projections of temperature-related extremes over the region.

1 Introduction

Extreme temperatures pose severe threats to human society and the natural environment, such as human health, energy consumption, agriculture and ecosystems (Intergovernmental Panel on Climate Change (IPCC), 2012). During recent decades, distinct warming trends have been documented (e.g., Donat et al., 2013; Perkins-Kirkpatrick & Lewis, 2020) and attributed to anthropogenic influence (e.g., Diffenbaugh et al., 2017; Fischer & Knutti, 2015; Min et al., 2011), which may further change the severity of these impacts (IPCC, 2013). In Australia, observations also show clear warming trends in extreme temperatures, which is represented by most global climate models (GCMs) relatively well (e.g., Alexander & Arblaster, 2009, 2017). However, to provide more confidence in future climate projections, it is critical to investigate whether new state-of-art climate models exhibit improved performance in simulating temperature extremes. Furthermore, the Australian climate is highly variable (e.g., Herold et al., 2018; Westra et al., 2016), which is related to a variety of physical mechanisms and teleconnections to modes of climate variability. For example, the frequency of heatwaves over southern and northern parts of Australia can be influenced by the El Niño-Southern Oscillation (ENSO); and for southeastern Australia, there is a positive correlation between the South Annular Mode (SAM) and heatwave frequency (Perkins et al., 2015). To better understand model deficiencies, extreme temperatures over different sub-continental regions should also be documented.

To investigate how extreme temperatures evolve in the past, present and future climate, global climate models are the main tools available. The global climate models in the 6th phase of the Coupled Model Intercomparison Project (CMIP6; Eyring et al., 2016), organized by the Working Group on Coupled Modelling (WGCM) of the World Climate Research Programme

(WCRP), recently became available and will contribute to the Intergovernmental Panel on Climate Change (IPCC) 6th Assessment Report (AR6). Compared to the previous phase, CMIP5 (Taylor et al., 2012), the models in CMIP6 generally have finer model resolution and improved physical processes (Eyring et al., 2016; Stouffer et al., 2017). However, the improvements in model configuration may not always lead to better simulations. Recent studies (e.g., Meehl et al., 2020; Tokarska et al., 2020; Zelinka et al., 2020) have shown that equilibrium climate sensitivity (ECS), a quantity of how global surface temperature changes once equilibrium is reached in response to an instantaneous doubling of CO_2, has a greater range in CMIP6 (1.8 to 5.6°C). As documented in Meehl et al. (2020), 12 of the 39 CMIP6 models exceed the upper end of the assessed ECS range in CMIP5 (1.5 to 4.5 °C). Though this new attribute of CMIP6, compared to CMIP5, suggests that there will be more severe impacts of future warming in some models, the higher values may not be realistic (Tokarska et al., 2020), which is likely due to how such models resolve cloud feedbacks and cloud-aerosol interactions (e.g., Meehl et al., 2020; Zelinka et al., 2020).

Extreme temperature can be measured in many ways. The Expert Team on Climate Change Detection and Indices (ETCCDI), organized by the joint World Meteorological Organization (WMO) Commission on Climatology (CCl)/World Climate Research Programme (WCRP) project on Climate Variability and Predictability (CLIVAR)/Joint Technical Commission for Oceanography and Marine Meteorology (JCOMM), defines 16 core indices (Zhang et al., 2011), which are based on daily-scale data and describe extremes that typically occur once a year or shorter. Compared to other indices or methods that describe temperature extremes, such as extreme value theory (e.g., Coles, 2001; Kharin et al., 2007; Kharin et al., 2013; Perkins et al., 2014; Zwiers et al., 2011) and the frequency of record-breaking high or low

87 monthly temperatures (Meehl et al., 2009), the ETCCDI indices are consistent, widely used and
88 easy to interpret (e.g., Alexander & Arblaster, 2017; Kim et al., 2020; Klein Tank et al., 2009;
89 Sillmann et al., 2013; Zhang et al., 2011). In a global study, using ETCCDI indices, Sillmann et
90 al. (2013) found that the inter-model spread in CMIP5 decreases for extreme temperatures,
91 compared to CMIP3. As an updated analysis of Sillmann et al. (2013), Kim et al. (2020)
92 concluded that there is limited improvement for CMIP6 models in simulating temperature
93 extremes, both globally and regionally; however, some systematic biases (e.g., the cold bias in
94 cold extremes over high-latitude regions) still exist. In Australia, there are distinct warming
95 trends in CMIP5 models for most locations, but cold extremes are generally overestimated, and
96 warm extremes underestimated (Alexander & Arblaster, 2017). CMIP6 has not been analyzed in
97 terms of ETCCDI indices over sub-regions for Australia, nor have the CMIP5 or CMIP6 indices
98 been compared as yet.

99 Since CMIP6 has not been analyzed in terms of ETCCDI indices over Australian regions,
100 the aim of this study is to investigate the performance of CMIP6 models in simulating
101 temperature extremes over Australian regions, compared to the models in CMIP5. The paper is
102 organized as follows: Section 2 introduces the observed and model data. The methods are
103 summarized in Section 3. Section 4 describes the results and the discussion and conclusions are
104 presented in Section 5.

105 Table 1. CMIP6 models used in this study

Model	Institution	Horizontal Resolution (lon × lat)
1. ACCESS-CM2	Commonwealth Scientific and Industrial Research Organisation (CSIRO) and Australian Research Council Centre of Excellence for Climate System Science (ARCCSS), Australia	192 × 145
2. ACCESS-ESM1-5	Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia	192 × 144
3. AWI-CM-1-1-MR	Alfred Wegener Institute, Germany	384 × 192
4. AWI-ESM-1-1-LR		192 × 96
5. BCC-CSM2-MR	Beijing Climate Center, China Meteorological Administration, China	320 × 160
6. BCC-ESM1		128 × 64
7. CanESM5	Canadian Centre for Climate Modelling and Analysis, Canada	128 × 64
8. CNRM-CM6-1	Centre National de Recherches Meteorologiques and Centre	256 × 128
9. CNRM-CM6-1-HR	Europeen de Recherche et Formation Avancees en Calcul Scientifique, France	720 × 360
10. CNRM-ESM2-1		256 × 128
11. FGOALS-f3-L	Chinese Academy of Sciences, China	288 × 180
12. FGOALS-g3		180 × 80
13. GFDL-CM4	National Oceanic and Atmospheric Administration (NOAA)	288 × 180
14. GFDL-ESM4	Geophysical Fluid Dynamics Laboratory, United States	288 × 180
15. GISS-E2-1-G	National Aeronautics and Space Administration (NASA) Goddard Institute for Space Studies, United States	144 × 90
16. HadGEM3-GC31-LL	Met Office Hadley Centre, United Kingdom	192 × 144
17. HadGEM3-GC31-MM		432 × 324
18. INM-CM4-8	Institute for Numerical Mathematics, Russia	180 × 120
19. INM-CM5-0		180 × 120
20. IPSL-CM6A-LR	Institut Pierre-Simon Laplace, France	144 × 143
21. MIROC-ES2L	Japan Agency for Marine-Earth Science and Technology,	128 × 64
22. MIROC6	Atmosphere and Ocean Research Institute at the University of Tokyo, National Institute for Environmental Studies and RIKEN Center for Computational Science, Japan	256 × 128
23. MPI-ESM-1-2-HAM	HAMMOZ-Consortium	192 × 96
24. MPI-ESM1-2-HR	Max Planck Institute for Meteorology, Germany	384 × 192
25. MPI-ESM1-2-LR		192 × 96
26. MRI-ESM2-0	Meteorological Research Institute, Japan	320 × 160
27. NorCPM1	Norwegian Climate Center, Norway	144 × 96
28. NorESM2-LM		144 × 96
29. NorESM2-MM		288 × 192
30. SAM0-UNICON	Seoul National University, South Korea	288 × 192
31. UKESM1-0-LL	Met Office Hadley Centre, United Kingdom	192 × 144

2 Data

In this study, we used daily maximum and minimum temperatures (TX and TN) in the historical simulations from 31 CMIP6 (Table 1) and 26 CMIP5 models (Table 2). Only one ensemble member (typically the first member) from each model is considered, as using all members would overemphasize some models with a large number of simulations (Seneviratne & Hauser, 2020).

As suggested by previous studies (Alexander & Arblaster, 2017; Sillmann et al., 2013; Srivastava et al., 2020), there are large differences between observational datasets. To robustly validate the simulated results produced by the models from CMIP5 and CMIP6, the Australian Water Availability Project (AWAP; Jones et al., 2009) and the Berkeley Earth Surface Temperatures (BEST; Rohde, Muller, Jacobsen, Muller, et al., 2013; Rohde, Muller, Jacobsen, Perlmutter, et al., 2013) are employed here.

AWAP is generated by the Commonwealth Scientific and Industrial Research Organization (CSIRO), the Australian Bureau of Agricultural and Resource Economics and Sciences (ABARES) and the Australian Bureau of Meteorology (BoM), which aims to understand the terrestrial water balance of Australia and the responses of land surface changes to climate variability and change (Jones et al., 2009). The gridded dataset includes rainfall, temperature, vapor pressure, solar exposure and the normalized difference vegetation index (NDVI) at the horizontal resolution of 0.05° × 0.05° (approximately 5 km × 5 km) over the period 1911–present. Although the analyses over data-sparse regions (e.g., central Western Australia) should be taken caution as the station network is changed over time and method of gridding can make timeseries inhomogeneous (Alexander & Arblaster, 2017; King et al., 2013),

AWAP is a high-quality observed dataset over Australia (King et al., 2017), which in this study is the primary reference dataset.

Table2. CMIP5 models used in this study

Model	Institution	Horizontal Resolution (lon × lat)
1. ACCESS−1.0	Commonwealth Scientific and Industrial Research Organization	192 × 145
2. ACCESS1-3	(CSIRO) and Bureau of Meteorology (BOM), Australia	192 × 145
3. bcc-csm1–1	Beijing Climate Center, China Meteorological Administration, China	128 × 64
4. BNU-ESM	Beijing Norml University, China	128 × 64
5. CanESM2	Canadian Centre for Climate Modelling and Analysis, Canada	128 × 64
6. CCSM4	National Center for Atmospheric Research (NCAR), United States	288 × 192
7. CESM1-BGC	National Science Foundation, Department of Energy and NCAR, United States	288 × 192
8. CMCC-CM	Centro Euro-Mediterraneo sui Cambiamenti Climatici, Italy	480 × 240
9. CNRM-CM5	Centre National de Recherches Meteorologiques and Centre Europeen de Recherche et Formation Avancees en Calcul Scientifique, France	256 × 128
10. CSIRO-Mk3–6–0	Commonwealth Scientific and Industrial Research Organization (CSIRO) and Queensland Climate Change Centre of Excellence, Australia	192 × 96
11. FGOALS-s2	State Key Laboratory of Numerical Modelling for Atmospheric Sciences and Geophysical Fluid Dynamics, Institute of Atmospheric Physics, Chinese Academy of Sciences, China	128 × 108
12. GFDL-ESM2G	National Oceanic and Atmospheric Administration (NOAA)	144 × 90
13. GFDL-ESM2M	Geophysical Fluid Dynamics Laboratory, United States	144 × 90
14. GISS-E2-R	National Aeronautics and Space Administration (NASA) Goddard Institute for Space Studies, United States	144 × 90
15. HadGEM2-CC	Met Office Hadley Centre, United Kingdom	192 × 144
16. HadGEM2-ES		192 × 144
17. IPSL-CM5A-LR	Institut Pierre-Simon Laplace, France	96 × 96
18. IPSL-CM5A-MR		144 × 143
19. IPSL-CM5B-LR		96 × 96
20. MIROC5	Japan Agency for Marine-Earth Science and Technology,	256 × 128
21. MIROC-ESM	Atmosphere and Ocean Research Institute at the University of	128 × 64
22. MIROC-ESM-CHEM	Tokyo, and National Institute for Environmental Studies, Japan	128 × 64
23. MPI-ESM-LR	Max Planck Institute for Meteorology, Germany	192 × 96
24. MPI-ESM-MR		192 × 96
25. MRI-CGCM3	Meteorological Research Institute, Japan	320 × 160
26. NorESM1-M	Norwegian Climate Center, Norway	144 × 96

As a globally observed dataset, BEST is also analyzed in this study, which provides daily high and low temperatures from 1880–present (Rohde, Muller, Jacobsen, Muller, et al., 2013; Rohde, Muller, Jacobsen, Perlmutter, et al., 2013). Compared to other global observational datasets (e.g., Global Precipitation Climatology Project (GPCP), Global Historical Climatology Network (GHCN)), the resolution of the Berkeley data is 1° × 1°, which is relatively higher and covers a longer period. Moreover, more records (around 37,000) are incorporated into the dataset, compared to 5,000–7,000 records incorporated into other global datasets. Since the Berkeley Earth claims to address some major concerns (e.g., data selection, data adjustment, poor station quality and the urban heat island effect) systematically and objectively, it is also an opportunity to check its validity in measuring temperature extremes over Australia.

3 Methods and Data Processing

3.1 Perkins' Skill Score

As a measure of how well each model can capture the probability distributions of weather variables in the observations, Perkins' skill score (PSS; Perkins et al., 2007) is defined as follows:

$$\text{PSS} = \sum_1^n \min(Z_o,\ Z_m)$$

where n is the number of bins used to calculate the probability distribution, Z_o is the frequency of the observed values, and Z_m is the frequency of simulated values in a given bin. A score of 0

indicates no overlapping area between the simulated and observed data, and a score of 100% means the two distributions are identical.

In this study, since the definitions of ETCCDI indices are based on TX and TN, it is necessary to examine the models' ability in simulating the distributions of TX and TN before applying the metrics to conduct research. It is noted that the definitions of some ETCCDI indices (e.g., cold nights (TN10p)) are not always based on TX or TN which are located in the tails of their probability distributions. Consequently, we utilized PSS to assess the overall similarity between the observed and simulated data (e.g., Kumar et al., 2014; Lewis, 2018; Perkins et al., 2007).

Table 3. Extreme temperature indices used in this study, defined by ETCCDI

Label	Index Name	Description	Unit
TXx	Hottest day	Annual maximum value of daily maximum temperature	°C
TXn	Coldest day	Annual minimum value of daily maximum temperature	°C
TNx	Warmest night	Annual maximum value of daily minimum temperature	°C
TNn	Coldest night	Annual minimum value of daily minimum temperature	°C
DTR	Diurnal temperature range	Annual mean difference between daily maximum and minimum temperature	°C
TX90p	Warm days	Percentage of time when daily maximum temperature is greater than 90th percentile (using running 5-day window)	%
TX10p	Cold days	Percentage of time when daily maximum temperature is less than 10th percentile (using running 5-day window)	%
TN90p	Warm nights	Percentage of time when daily minimum temperature is greater than 90th percentile (using running 5-day window)	%
TN10p	Cold nights	Percentage of time when daily minimum temperature is less than 10th percentile (using running 5-day window)	%
WSDI	Warm spell duration index	Annual count when at least six consecutive days of maximum temperature is greater than 90th percentile (using running 5-day window)	days
CSDI	Cold spell duration index	Annual count when at least six consecutive days of minimum temperature is less than 10th percentile (using running 5-day window)	days
SU	Summer days	Annual count when daily maximum temperature is greater than 25°C	days
TR	Tropical nights	Annual count when daily minimum temperature is greater than 20°C	days

FD	Frost days	Annual count when daily minimum temperature is less than 0°C	days

163

164 ### 3.2 ETCCDI Indices

165 ETCCDI indices used in this study are outlined in Table 3. The indices defined in Zhang
166 et al. (2011) can be classified into four categories: absolute indices (e.g., hottest day (TXx)),
167 threshold-based indices (e.g., frost days (FD)), percentile indices (e.g., cold days (TX10p)), and
168 duration indices (e.g., cold spell duration index (CSDI)). Since the definitions of growing season
169 length (GSL) and ice days (ID) are not suitable over most of Australia (Alexander & Arblaster,
170 2017), they are excluded here. Furthermore, compared to previous studies (e.g., Alexander &
171 Arblaster, 2017; Sillmann et al., 2013), the bootstrap resampling procedure proposed by Zhang et
172 al. (2005) is also applied to the calculations of warm spell duration index (WSDI) and CSDI, and
173 the spells crossing year boundaries are taken into consideration.

174 The linear trends of the time series of ETCCDI indices are estimated by Theil-Sen
175 estimator and Mann-Kendall non-parametric test is used as the significance test (e.g., Alexander
176 & Arblaster, 2009; Dey et al., 2019).

177 ### 3.3 Model Performance Metric

178 Following Sillmann et al. (2013), the evaluation of model performance is based on root-
179 mean-square error (RMSE), which is calculated as:

$$RMSE = \sqrt{\langle (X - Y)^2 \rangle}$$

180 where X is the model climatology of an ETCCDI indicator, Y represents the corresponding
181 climatology in the observed data, and the angular brackets denote spatial averaging over a

182 particular domain. It is a quantity that measures the degree of agreement between the observed
183 and simulated data.

184 3.4 Data Processing

185 The observed and simulated datasets of TX and TN are first regridded to 1° × 1°
186 resolution using bilinear interpolation; the calculations of ETCCDI indices are then performed. It
187 is noted that reversing the order of operation may have significant effects on the resulting
188 gridded values (e.g., Avila et al., 2015; Chen & Knutson, 2008; Herold et al., 2017; Zhang et al.,
189 2011). For example, indices sensitive to resolution choice (e.g., Maximum 1-day precipitation
190 amount) are substantially altered when the order of operation is changed (Herold et al., 2017). In
191 addition, following the practice in King et al. (2015), gridboxes containing less than 75% land
192 are masked out.

193 To investigate the Australian extreme temperatures in more detail, Australia is also
194 divided into nine sub-regions shown in Table 4 and Fig. 1, which is based on a study by Perkins
195 et al. (2014) and the BoM
196 (http://www.bom.gov.au/climate/change/about/temp_timeseries.shtml). Ten regions were
197 determined according to climatological and geographical conditions, abbreviated AUS
198 (Australia), NA (Northern Australia), SA (Southern Australia), SEA (South East Australia),
199 MEA (Middle Eastern Australia), TA (Tropical Australia), SWA (South West Australia), SSA
200 (Southern South Australia), CAU (Central Australia) and MWA (Mid-Western Australia). Since
201 there has been an increase in in-situ observations since 1950, the analysis is carried out for the
202 period of 1950 – near-present, and the base period is from 1961 to 1990, which is commonly
203 used and allows for a standardized quantification. In the next section, each model is first
204 evaluated for TX and TN using the PSS for each region. Then, all extreme temperature indices

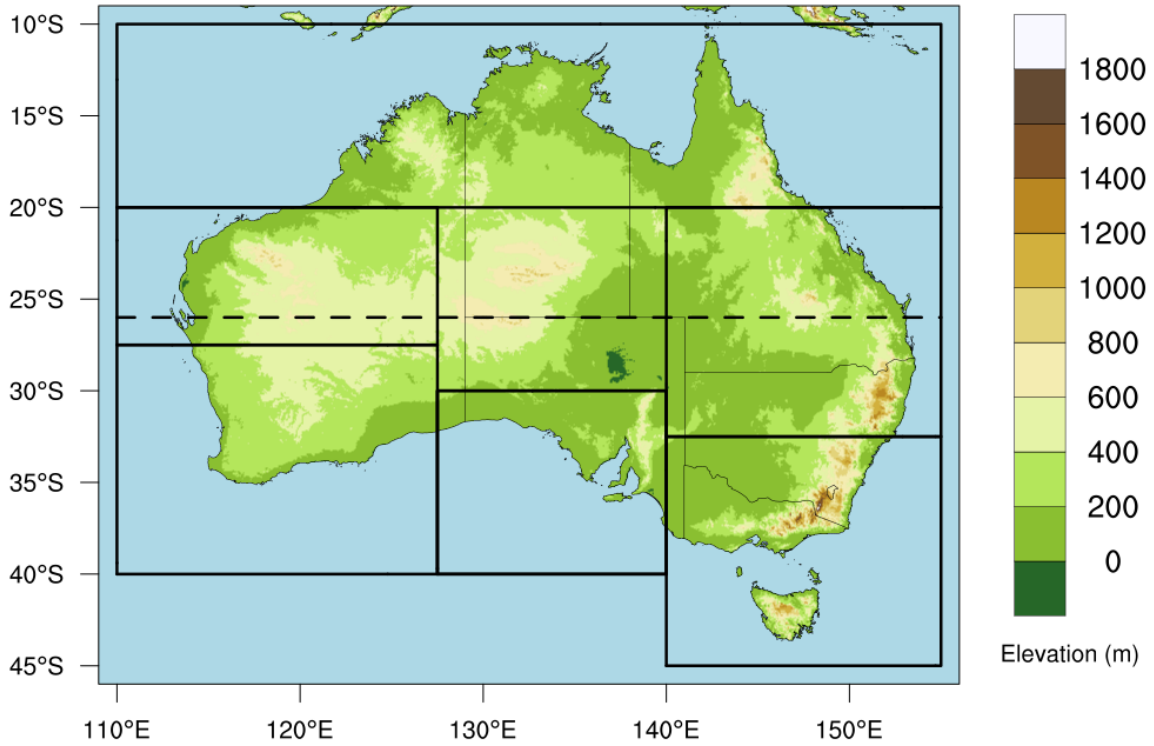
205 are analyzed in terms of the spatial patterns and temporal evolution, as well as the RMSE being
206 performed.

207

208 Table 4. Latitude and longitude boundaries of Australian regions

Label	Region	Lat (°S)	Lon (°E)
1. AUS	Australia	10-45	110-155
2. NA	Northern Australia	10–26	110–155
3. SA	Southern Australia	26–45	110–155
4. SEA	South East Australia	32.5–45	140–155
5. MEA	Middle Eastern Australia	20–32.5	140–155
6. TA	Tropical Australia	10–20	110–155
7. SWA	South West Australia	27.5–40	110–127.5
8. SSA	Southern South Australia	30–40	127.5–140
9. CAU	Central Australia	20–30	127.5–140
10. MWA	Mid-Western Australia	20–27.5	110–127.5

209



210
211 Figure 1. Regions used in the study. Northern Australia (NA) and Southern Australia (SA) are divided by the dashed
212 line at 26°S, and solid lines denote the boundaries of other Australian subregions.
213

4 Results

4.1 Probability Distributions and PSS

Figs. 2-3 and Figs. 4-5 show the probability distributions of TX and TN and their PSSs over the Australian regions during the period 1950-2005 for AWAP, BEST, CMIP6 and CMIP5 models. Bin sizes of 0.5°C were used. For the probability distributions of TX (Fig. 2), the two observations are generally comparable over the regions, though there are slight differences between AWAP and BEST over the regions SWA and SSA. In contrast, the probability distributions of TN (Fig. 3) in the two observed datasets show larger differences over most regions (except NA). Overall, for TN, BEST tends to have right shifted distributions (warmer-side tails), with higher peaks over the northern regions and lower peaks over the southern regions, compared to AWAP.

For both TX and TN, the multi-model medians in CMIP6 and CMIP5 are generally similar over all regions (Figs. 2-3). Compared to AWAP, the medians of the two CMIP ensembles in the probability distributions of TX tend to overestimate the lower tails and underestimate the upper tails in Fig. 2. For TN (Fig. 3), the lower tails are underestimated and the upper tails overestimated. Furthermore, the medians in CMIP6 and CMIP5 are more analogous to AWAP than BEST. The model spread, as measured by the full range of the multi-model ensemble in each CMIP, tends to be larger in the upper tails and narrower in the lower tails for CMIP6 when compared to CMIP5 (Figs. 2-3). This suggests that more models in CMIP6 tend to show warmer patterns. In particular, for the probability distributions of TX in CMIP6 models, the larger spread in the upper tail is mainly caused by the three models CanESM5, MIROC6 and MRI-ESM2 (not shown).

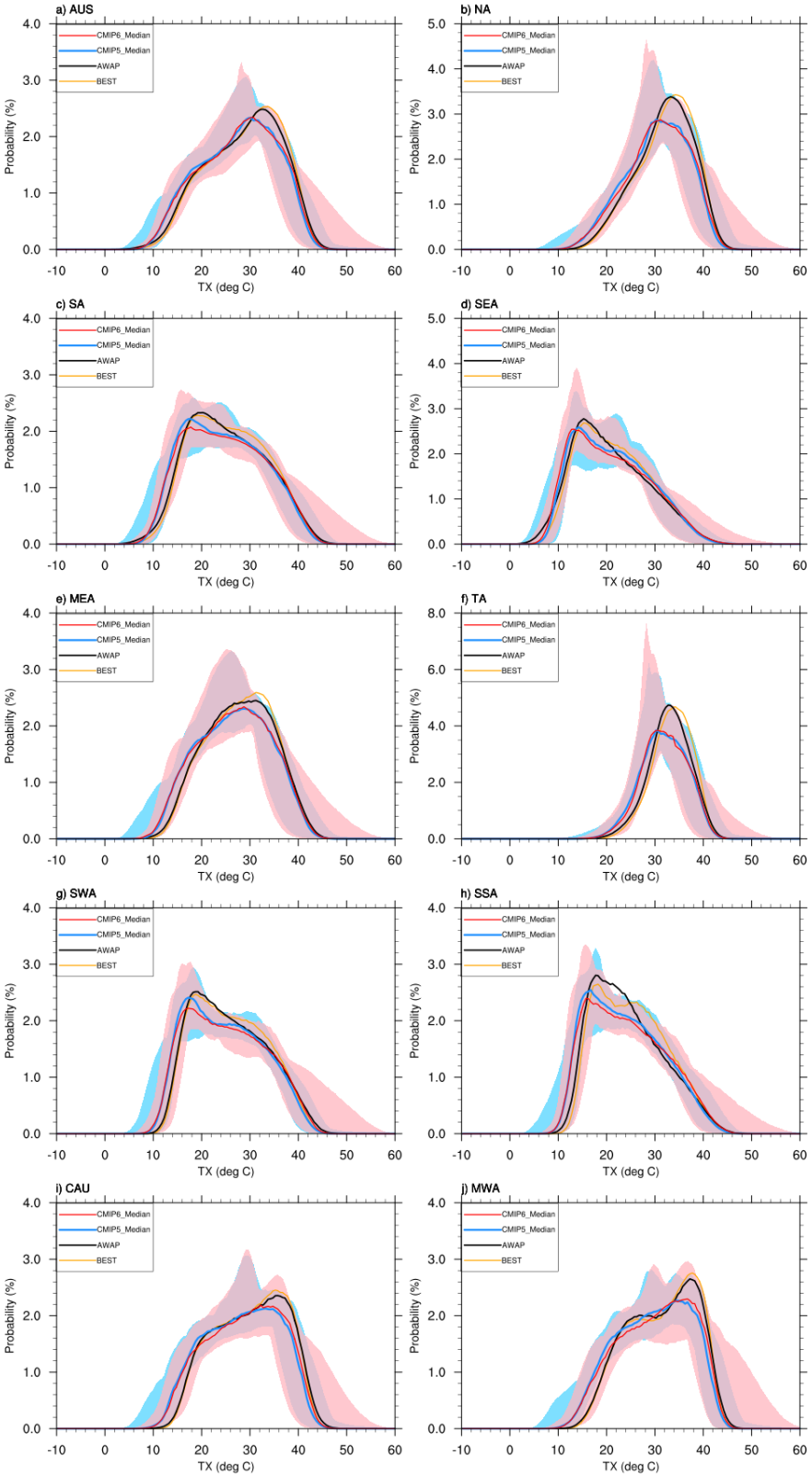


Figure 2. Probability distributions of daily maximum temperature (TX) during the period 1950-2005 over Australian regions for AWAP (black), BEST (yellow), CMIP6_Median (red) and CMIP5_Median (blue); shading denotes the full range across the models in CMIP6 (red) and CMIP5 (blue).

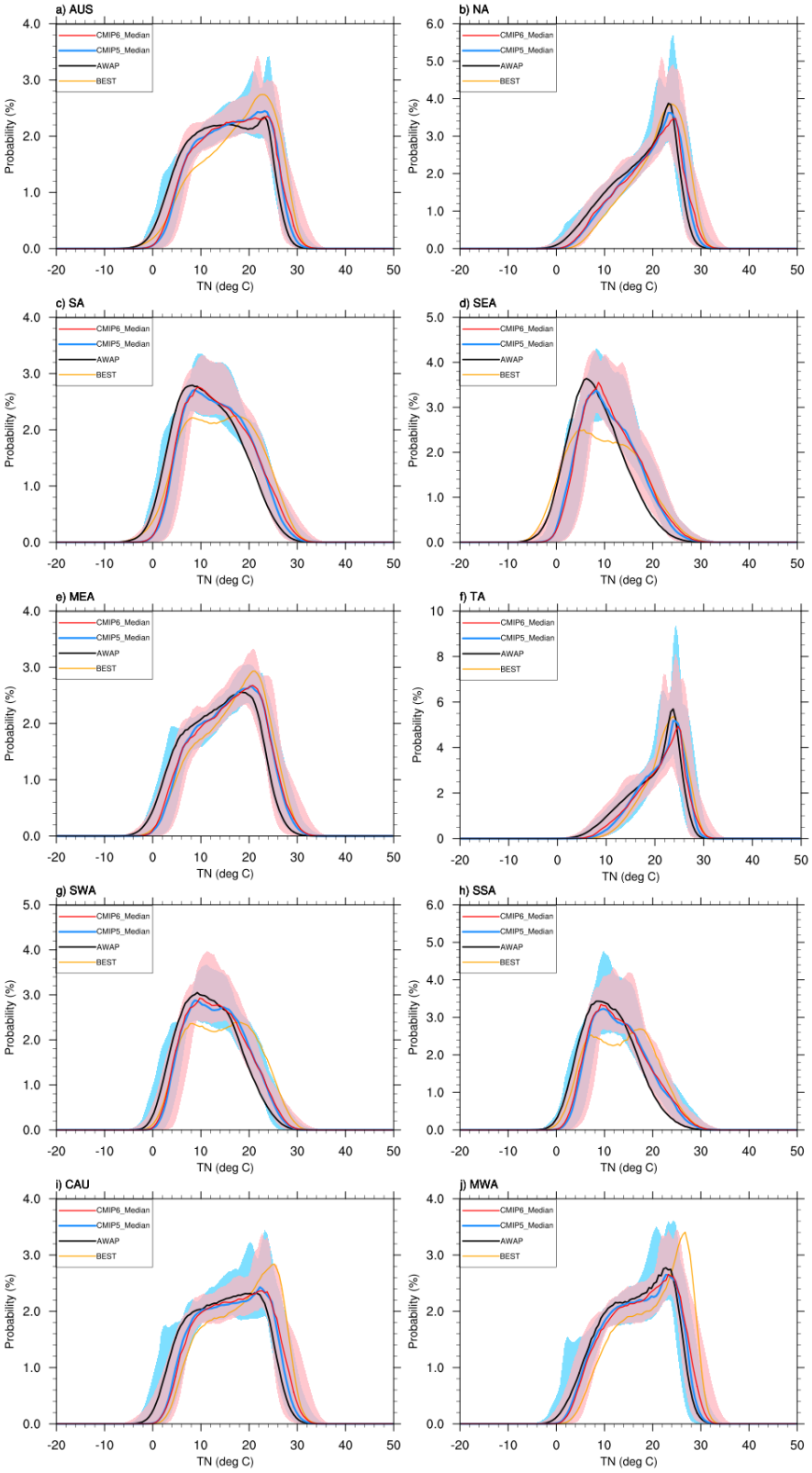
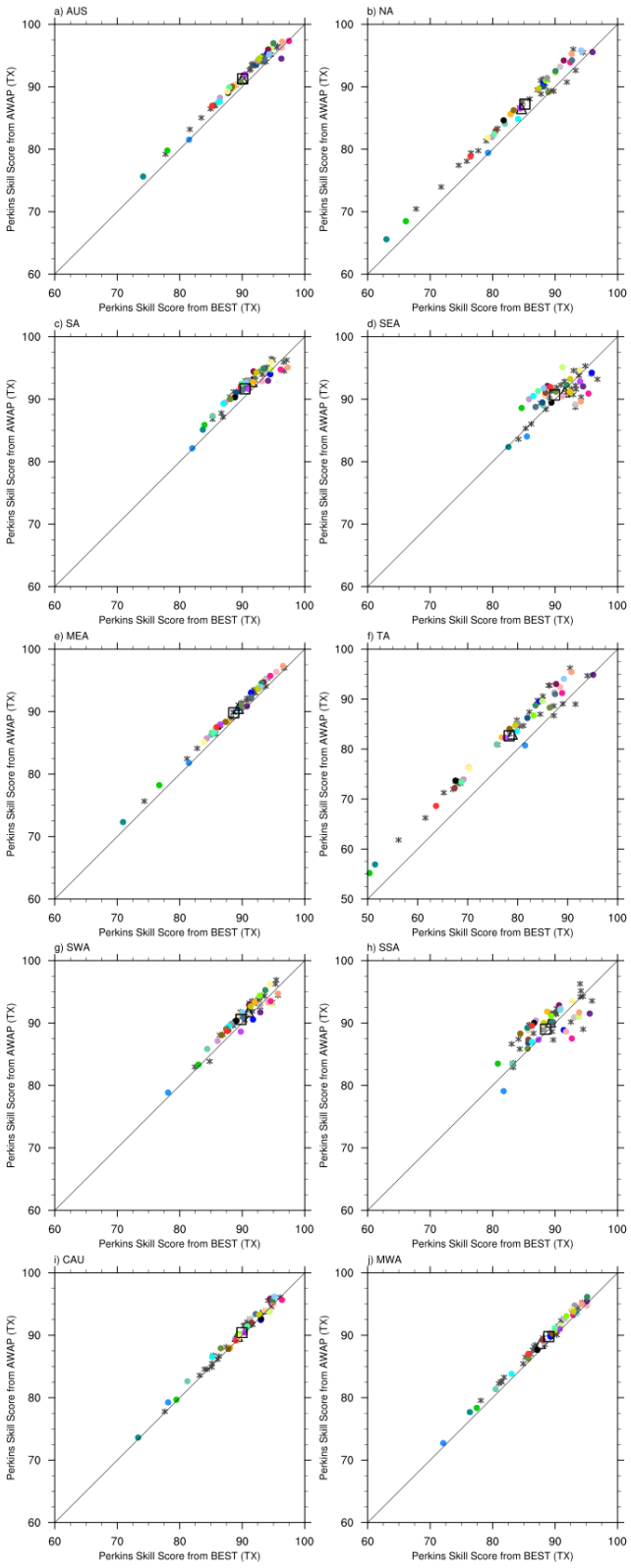


Figure 3. Same as Fig. 2, but for daily minimum temperature (TN).


a) AUS
b) NA
c) SA
d) SEA
e) MEA
f) TA
g) SWA
h) SSA
i) CAU
j) MWA
Perkins Skill Score from AWAP (TX)
Perkins Skill Score from BEST (TX)

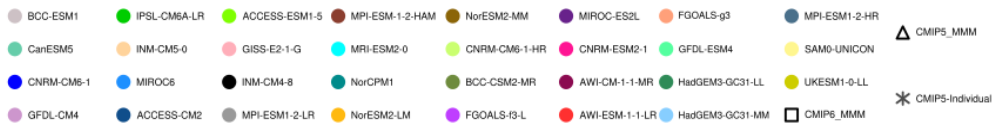


Figure 4. Perkins' skill scores for probability distributions of TX over the Australian regions for the period 1950–2005; the colored circles represent CMIP6 models and the models in CMIP5 are denoted by the black asterisks; the triangles and squares are the multi-model means from CMIP5 and CMIP6.

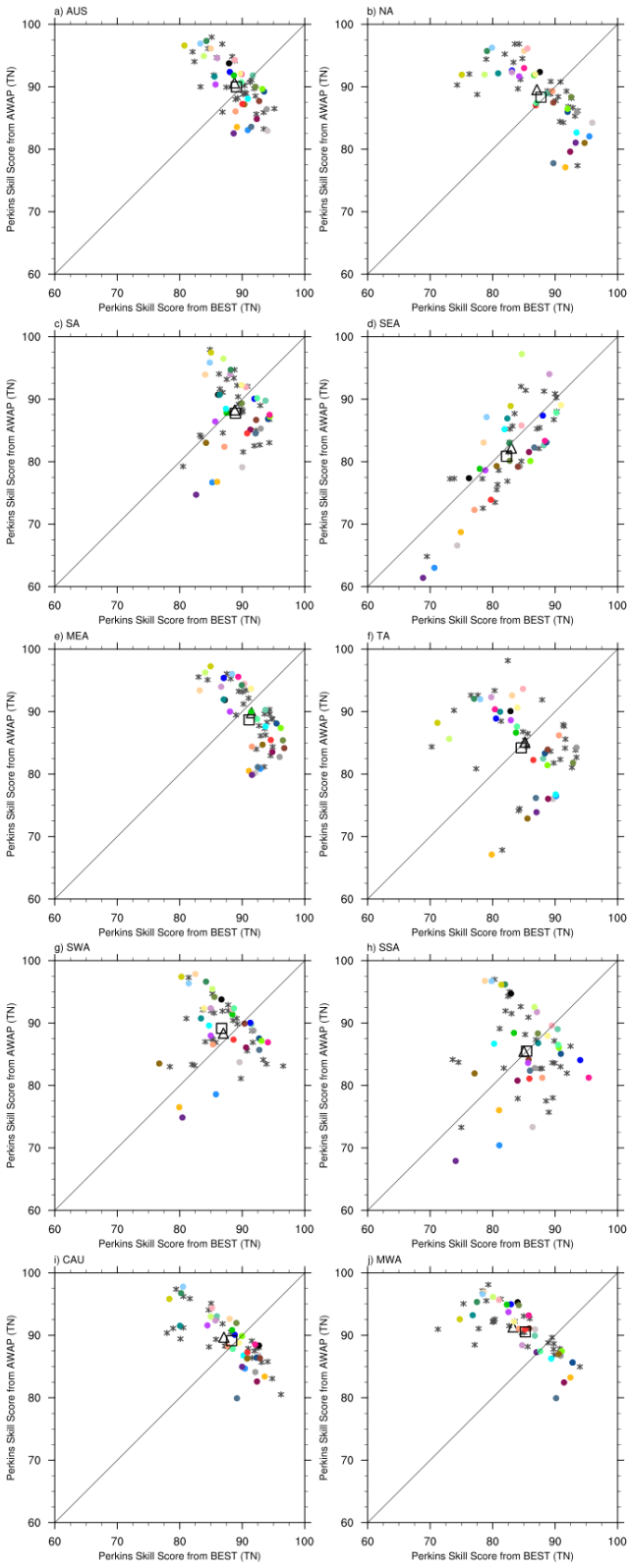


Figure 5. Same as Fig. 4, but for daily minimum temperature (TN).

In Figs. 4 and 5, compared to both observations, the multi-model means of PSSs in CMIP6 and CMIP5 models are generally around 90%, which implies that both CMIP ensembles simulate the daily-scale extreme temperatures similarly and relatively well. The lower multi-model mean PSSs are found for TX over TA (~83%) and TN over SEA (~82%), TA (~84%) and SSA (~84%). Also, over most regions shown in Fig. 5 (e.g., AUS, NA, MEA), higher scores for AWAP does not mean higher scores in BEST, suggesting that the two observed datasets are significantly different. For the model spreads of PSSs, the full ranges of the probability distributions for TX and TN in CMIP6 are commonly wider than CMIP5 over the regions. This could be due to the fact that several models in CMIP6, such as MIROC6 and NorCPM1, show relatively lower scores. It is also noted that the models with higher resolution (e.g., MRI-ESM2-0) do not generally show higher scores than those with relatively coarse resolution (e.g., FGOALS-g3; Figs. 4 and 5). As the change of temperature may be more related to large-scale meteorological patterns (Grotjahn et al., 2016), the relatively lower PSSs in some models with higher resolution may result from the generation of unrealistic local details (e.g., soil moisture) in simulations (Lau & Nath, 2012).

In general, models in CMIP6 and CMIP5 can be evaluated quite differently based on AWAP or BEST; and the multi-model means and spreads of PSSs over most regions in CMIP6 are comparable to that in CMIP5, though the multi-model means are typically slightly lower in CMIP6 for both TX and TN over most regions (Figs. 4 and 5). This is because some models in CMIP6, which usually produce lower scores, collectively reduce the ensemble mean. Compared to AWAP (Fig. 4), MIROC6, NorCPM1, IPSL-CM6A-LR and CanESM5 usually have lower scores. Of those, NorCPM1 and IPSL-CM6A-LR have cold shifts while warm shifts occur for MIROC6 and CanESM5 (not shown). In contrast, the PSSs of MIROC-ES2L, MIROC6, MPI-

ESM1-2-HR and NorESM2-LM for the probability distributions of TN are usually lower over the Australian regions (Fig. 5), which all have warm shifts (not shown). It is interesting to note that the model MIROC-ES2L typically has lower PSSs in Fig. 5 but relatively higher scores in Fig. 4, implying that MIROC-ES2L tends to simulate higher temperatures over Australia. Furthermore, as the model CanESM5 shows a warmer upper side in the probability distributions of TX and relatively lower scores in Fig. 4, the extreme heat calculated from CanESM5 may be unrealistic. In addition, although the ECS is a measure of global climate sensitivity, the higher values in CanESM5 documented in recent studies (e.g., Meehl et al., 2020; Zelinka et al., 2020) may be doubtful as well. The results based on PSSs suggest that when using historical simulations from the above models to calculate extremes, the results should be interpreted with caution.

4.2 Spatial Patterns of Climatologies

Examining the extreme temperature indices averaged over the period 1961-1990 helps us to determine the magnitude and spatial distributions of model bias. The 30-year climatologies of TXx, (coldest night) TNn and diurnal temperature range (DTR) for the observations and the historical simulations from CMIP6 and CMIP5 models are shown in Figs 6-8, as well as the biases between the simulated and observed datasets. The climatological patterns of other indices, including coldest day (TXn), warmest night (TNx), WSDI, CSDI, summer days (SU), tropical nights (TR) and FD, are shown in Supplementary Material Figs. S1-S7. Except for DTR (Fig. 8), AWAP and BEST exhibit similar patterns for other temperature indices (Figs. 6-7 and Figs. S1-S7). Overall, compared to AWAP, the magnitude in BEST for most indices is higher over most parts of Australia, although the absolute values of TXx (Fig. 6) and FD (Fig. S7) in BEST are generally lower. The negligible variation of DTR in BEST (Fig. 8b) is likely caused by the

minimization process in the Berkeley's homogenization algorithm, which minimizes the mean square of the local weather term and suppresses regional differences to some extent (Rohde, Muller, Jacobsen, Muller, et al., 2013; Rohde, Muller, Jacobsen, Perlmutter, et al., 2013).

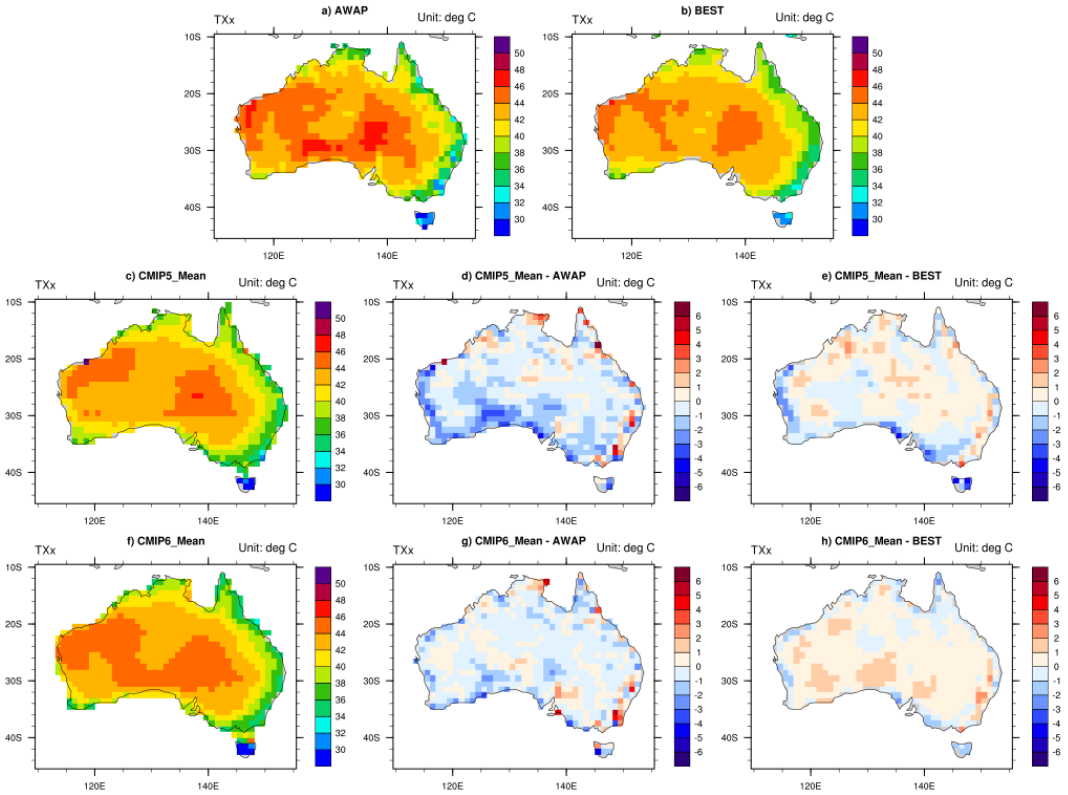


Figure 6. Spatial patterns of 30-year climatological TXx (1961–1990) over Australia for a) AWAP, b) BEST, c) the multi-model mean of CMIP5 (termed "CMIP5_Mean") and f) the multi-model mean of CMIP6 (termed "CMIP5_Mean"), and the biases for d) CMIP5_Mean - AWAP, e) CMIP5_Mean - BEST, g) CMIP6_Mean - AWAP and h) CMIP6_Mean – BEST.

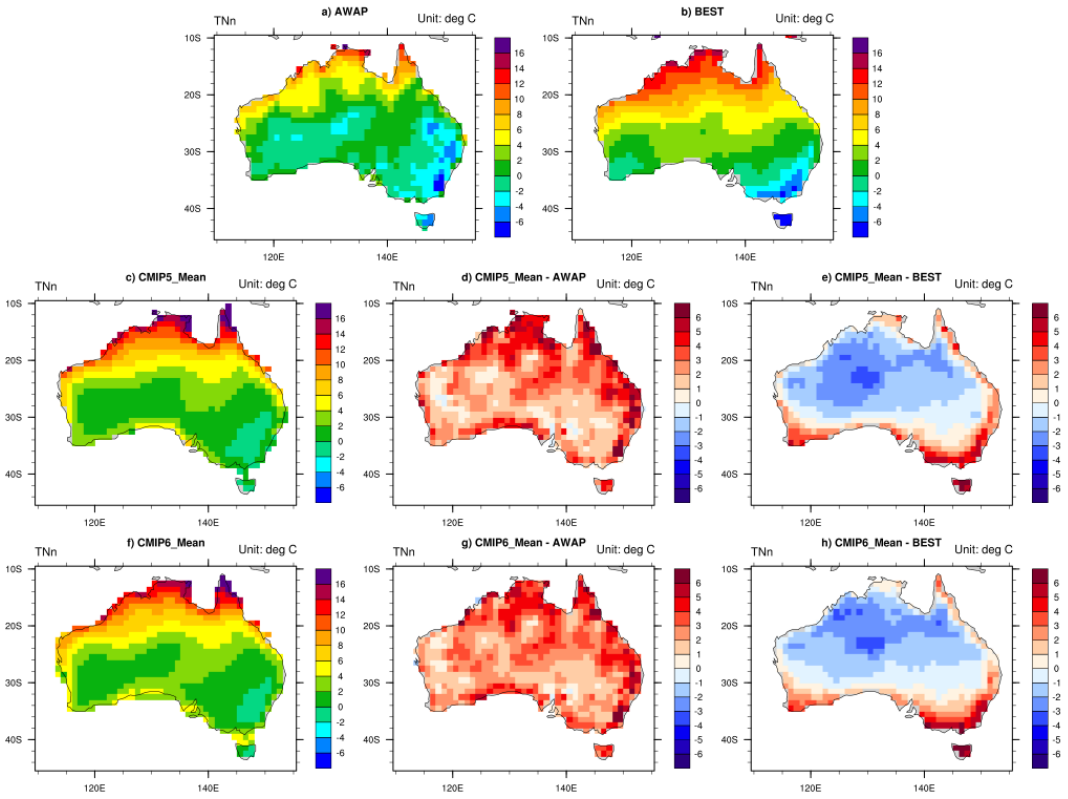


Figure 7. Same as Fig. 6, but for TNn.

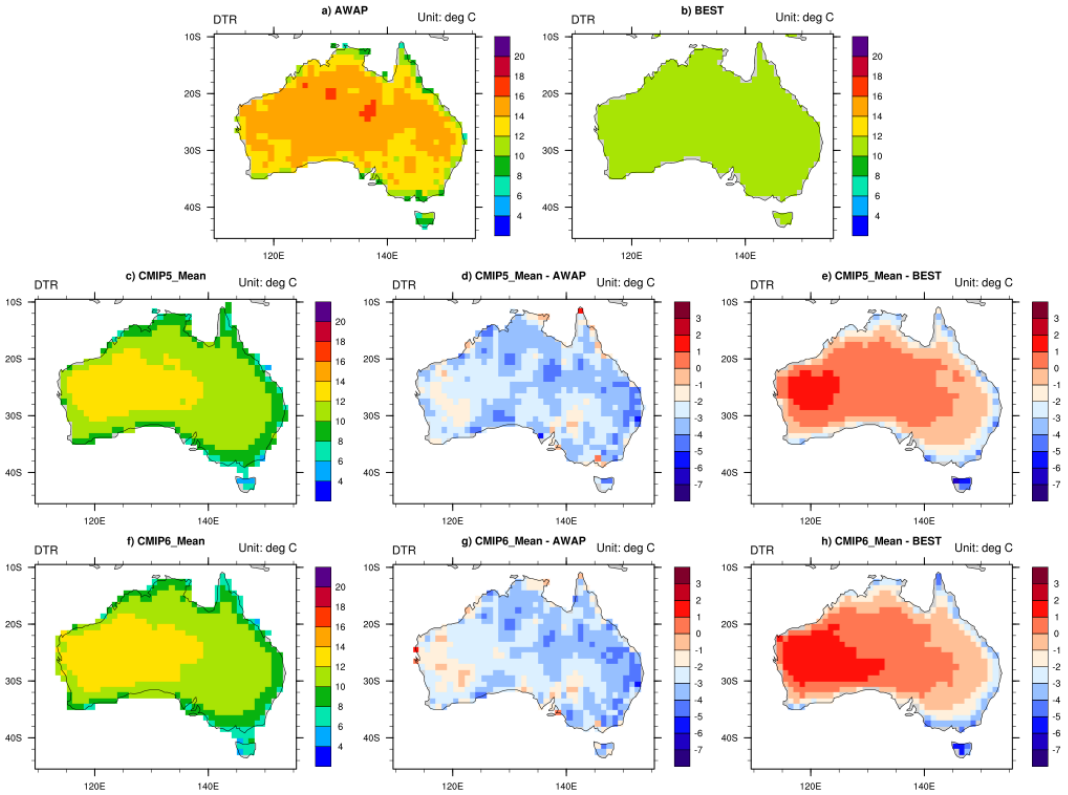


Figure 8. Same as Fig. 6, but for DTR.

316 The observed climatological indices are reasonably well represented by the models from
317 CMIP6 and CMIP5. However, similar to CMIP5, systematic errors still exist in the CMIP6
318 multi-model mean. As shown in Figs 6-8 and Figs. S1-S7, the distinct differences are usually
319 located over the eastern part of tropical Australia, southeast and western Australia. For example,
320 for TXx, there are cold biases over southwest Australia and warm biases over southeast Australia
321 (Fig. 6g). In general, compared to AWAP, the multi-model means of CMIP6 appear to show
322 improvements for some indices (e.g., TXx, TXn, CSDI, SU).

323 To investigate regional performance of CMIP6 models, box-and-whisker plots are
324 employed to show ETCCDI indices over the Australian regions (Fig. 9). The boxes indicate the
325 interquartile model spreads (range between the 25th and 75th quantiles), the black lines within
326 the boxes are the multi-model medians, the whiskers extend to the edges of 1.5×interquartile
327 ranges, and "outlier points" that fall outside of the whiskers are denoted by diamonds. Except for
328 DTR, BEST exhibit broadly higher values than AWAP over most regions (Figs. 9b-i). However,
329 for TXx (Fig. 9a) and FD (Fig. 9j), the magnitudes of indices in BEST are generally lower than
330 AWAP. Moreover, the differences between the observational datasets may be comparable to the
331 interquartile range of the models from CMIP6 and CMIP5 over most regions for many indices
332 (except TXx, TXn, SU and FD), which may be due to the homogenization algorithm and
333 relatively poor observational network coverage. This further implies that based on different
334 observational data, the model evaluation results may differ, which is consistent with previous
335 studies (e.g., Kim et al., 2020; Sillmann et al., 2013; Srivastava et al., 2020).

336 Compared to AWAP, the multi-model medians of CMIP6 tend to overestimate the
337 duration indices (i.e., WSDI and CSDI) over all Australian regions. For absolute and threshold
338 indices, TXx, TXn, DTR, SU and FD are commonly underestimated by the CMIP6 over most

regions (except TXx, TXn and SU over SEA); while the medians in CMIP6 models overestimate TNx, TNn and TR. Over some regions such as SEA, MEA, TA and CAU, there are relatively higher biases between AWAP and the medians in the CMIP6 models.

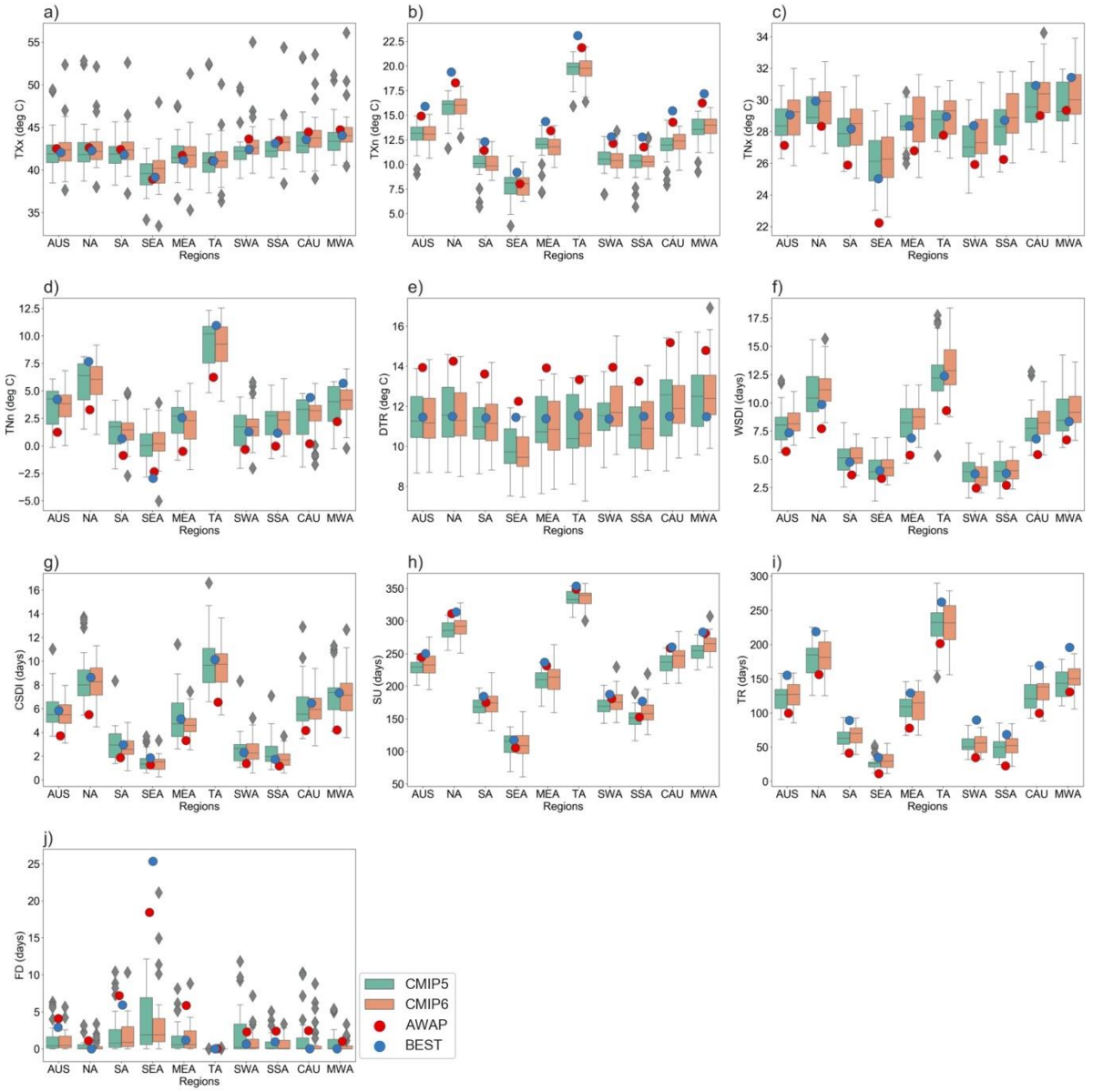


Figure 9. Box-and-whisker plots for the 10 ETCCDI indices calculated from 31 CMIP6 models (orange) and 26 CMIP5 models (green) over Australian regions. The boxes indicate the interquartile model spreads (range between the 25th and 75th quantiles), the black lines within the boxes are the multi-model medians, the whiskers extend to the edges of 1.5×interquartile ranges and "outliers" outside of the whiskers are denoted by diamonds. The round circles represent the indices in AWAP (red) and BEST (blue) datasets.

For the comparison between CMIP6 and CMIP5 models, the multi-model medians and interquartile model ranges are analyzed and shown to be broadly comparable. The distinct differences for the medians are among the absolute indices. For TXx and TNx, the medians in CMIP6 models are higher than CMIP5. In contrast, for TXn and TNn, CMIP6 shows lower values over most regions (expect for TXn over the regions CAU and MWA). The interquartile model ranges in CMIP6 tend to be lower than CMIP5 for TNn, WSDI and CSDI over most regions, which suggests that the model uncertainty in CMIP6 may be reduced. However, over some regions such as NA, TA and MEA, the interquartile range tends to be larger for some indices, compared to other regions, suggesting that models simulating the extremes over these regions may have more uncertainty.

4.3 Metric Evaluation

With respect to AWAP, the RMSEs for the CMIP6 and CMIP5 models are used to assess the models' overall performance in simulating extreme temperature indices averaged for the base period 1961-1990 over Australian regions (Fig. 10; RMSEs based on BEST is shown in Fig. S8). The medians in the two ensembles commonly have higher values over tropical and eastern Australia (Fig. 10). And the models do not perform consistently well over Australian regions (not shown), which suggests that there is large variability for the performance of the models in simulating different indices over different regions. For example, in CMIP6, the model MIROC-ES2L has higher RMSEs across all regions for TNn while its performance in simulating TXn is relatively better than other models (lower RMSEs). The values of RMSEs in CMIP6 also suggest that the models need further improvement over the regions MEA, TA, CAU and MWA. Overall, the models HadGEM3-GC31-MM, HadGEM3-GC31-LL and GFDL-CM4 are commonly among

the best performers, while NorCPM1, NorESM2-LM and MIROC6 tend to show higher RMSEs (not shown).

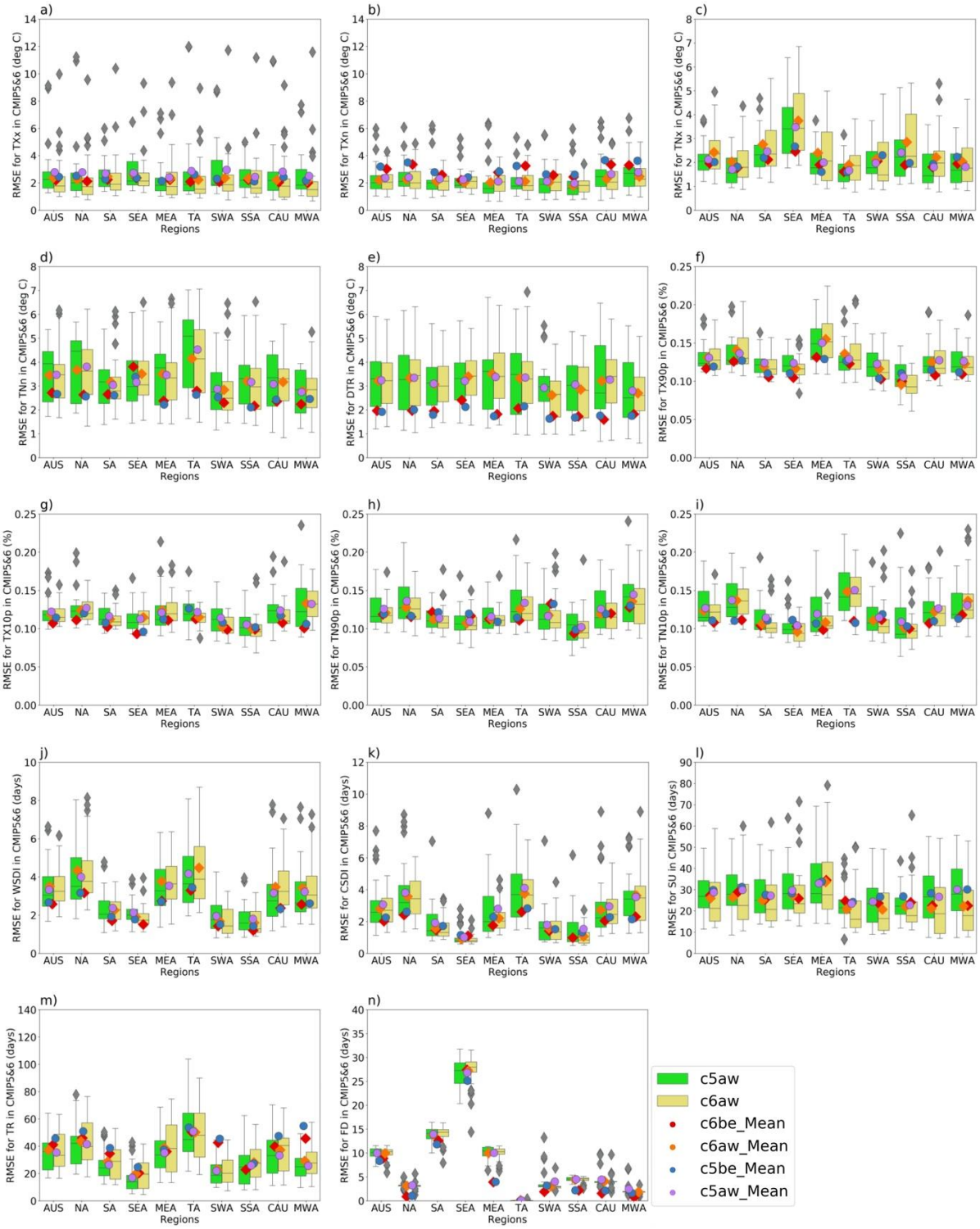


Figure 10. Box-and-whisker plots for the RMSEs of 14 ETCCDI indices calculated from 31 CMIP6 models (green) and 26 CMIP5 models (yellow) over Australian regions, with respect to AWAP. The boxes indicate the interquartile

spreads (range between the 25th and 75th quantiles), the black lines within the boxes are the multi-model medians, the whiskers extend to the edges of 1.5×interquartile ranges and "outliers" outside of the whiskers are denoted by diamonds. The round circles represent the multi-model means of RMSEs calculated from CMIP5 with respect to AWAP and BEST, termed "c5aw_Mean" (purple) and "c5be_Mean" (blue); the squares are the same but for CMIP6, termed "c6aw_Mean" (yellow) and "c6be_Mean" (red).

Compared to the RMSEs in CMIP5 models, there are some improvement shown in CMIP6. Usually, for some cold extremes (e.g., TNn, warm nights (TN90p), TN10p, CSDI and FD), the interquartile model ranges are commonly narrower in CMIP6. For TXx, TNn and SU, the means and medians of RMSEs in CMIP6 are generally lower than CMIP5.

4.4 Temporal Variations

Time series of the anomalies and the actual values for extreme temperature indices averaged over Australia (10-45°S, 110-155°E) are shown in Fig. 11 and Fig. S9, respectively. Furthermore, the boxplots representing trends over Australia regions are displayed in Fig. 12, and the number of models that show trends of ETCCDI indices significant at 95% level is summarized in Table. 5.

As shown in Fig. 11, the temporal variations of the two observations for the extremes are quite similar and they are reasonably well captured by both the CMIP ensembles. However, for some indices, differences between AWAP and BEST are substantial. For example, the differences between the two observations for TR (Fig. S9m) can be as large as the total inter-model range, further indicating the observational uncertainty can be quite large. Consistent with Alexander and Arblaster (2017), the temporal variations of TNx, TNn and TR in AWAP is close to the lower end of the model spread in CMIP6 and CMIP5, while the observed TXn, DTR and FD tend to be at the upper end (Fig. S9). In terms of the model spread, some outliers shown in

CMIP5 are relieved in CMIP6 (e.g., the outliers shown in TN10p and CSDI produced by the model GFDL-ESM2G in the year 1964).

In Fig. 12, the trends of temperature indices in the observed and simulated data are displayed for each region. For all the temperature indices, the warming trends of BEST are generally higher than AWAP over most regions, with the lower warming trends in BEST usually located over SSA, CAU and MWA, which are data-sparse regions. Again, the differences between the observations can be as large as the interquartile model range (e.g., TN10p). Compared to the medians of CMIP5 models, the medians in CMIP6 are commonly closer to AWAP (e.g., TXx, warm days (TX90p) and SU). Moreover, both the spreads and interquartile model ranges tend to be narrower in CMIP6, and there is a larger portion of models in CMIP6 that show the trends significant as compared to CMIP5 (Table 5). This may imply that the model uncertainty in CMIP6 is somewhat reduced. Over the regions, the interquartile model ranges in CMIP6 and CMIP5 are usually larger over NA, TA and MWA for some indices (e.g., TNn, TX90p, TN90p, WSDI and CSDI).

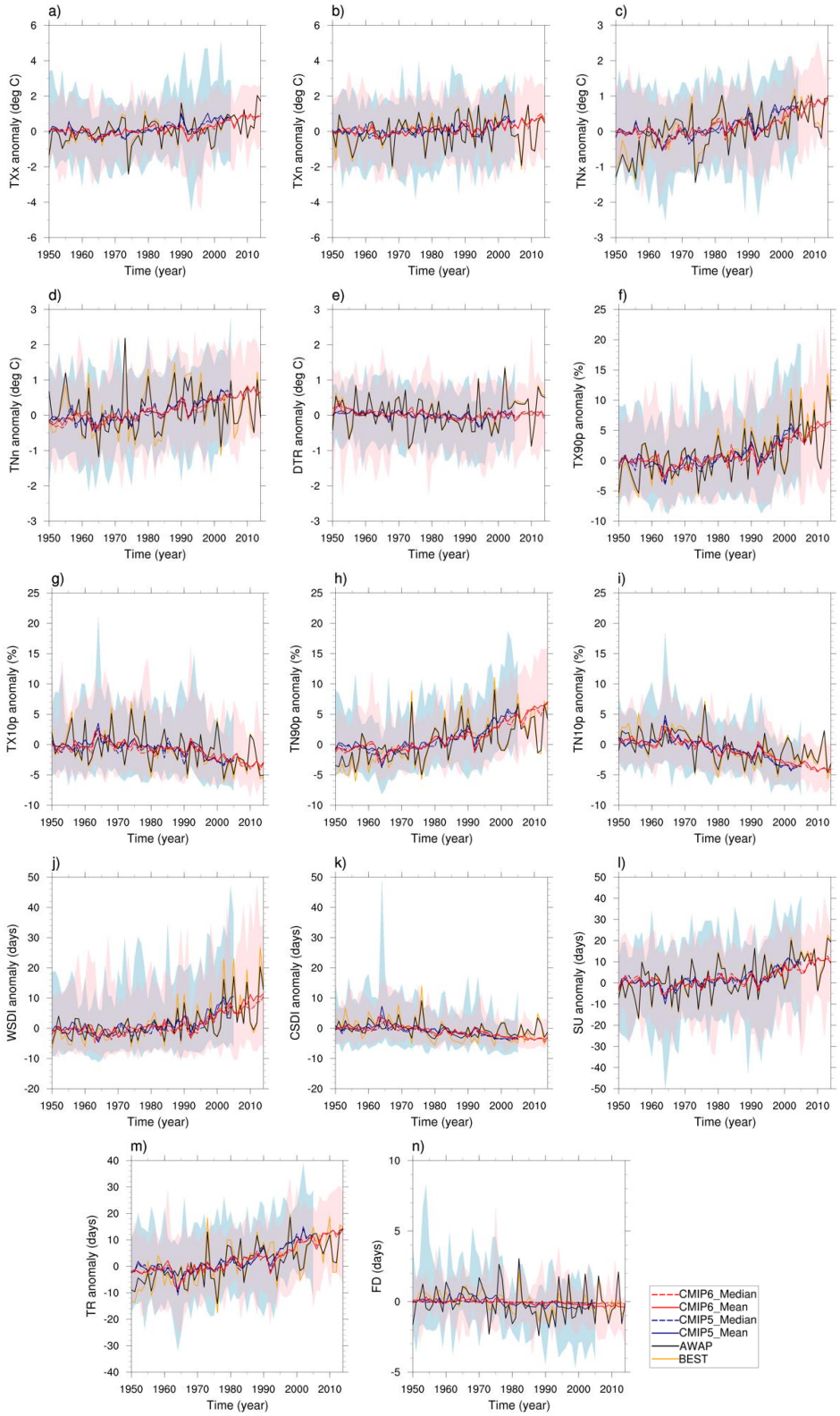


Figure 11. Time series for the anomalies of the 14 ETCCDI indices averaged over Australia (10-45°S, 110-155°E) from 1950 to 2014 for AWAP (red), BEST (yellow), CMIP5 (multi-model mean: red solid; multi-model median: red dashed) and CMIP6 (multi-model mean: blue solid; multi-model median: blue dashed); Shading indicates the full range of CMIP5 (blue) and CMIP6 (red) models.

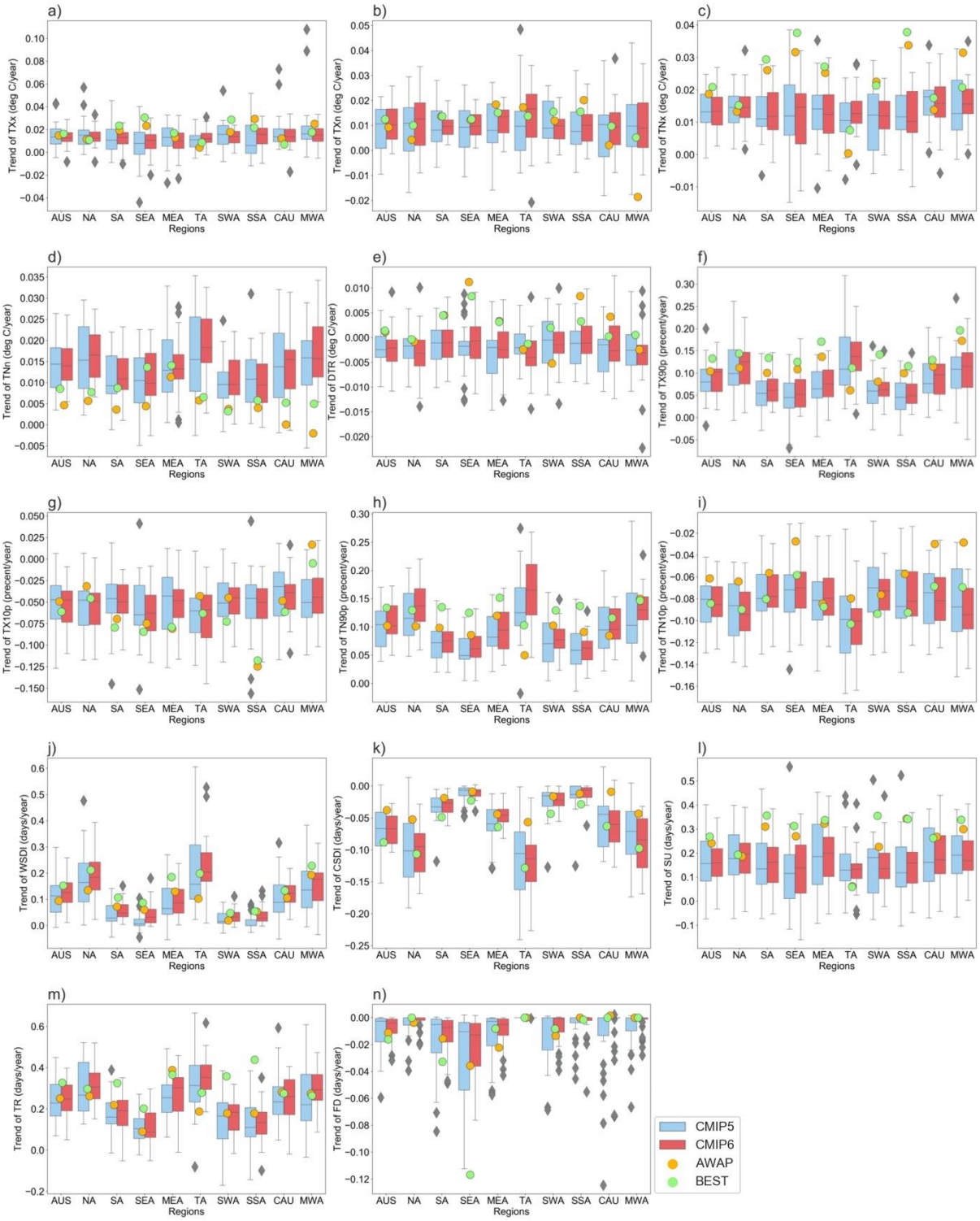


Figure 12. Box-and-whisker plots for the trends of 14 ETCCDI indices calculated from 31 CMIP6 models (red) and 26 CMIP5 models (blue) over Australian regions. The boxes indicate the interquartile spreads (range between the 25th and 75th quantiles), the black lines within the boxes are the multi-model medians, the whiskers extend to the edges of 1.5×interquartile ranges and "outliers" outside of the whiskers are denoted by diamonds. The round circles represent the indices in AWAP (yellow) and BEST (green) datasets.

443 Table 5. Number of models in CMIP6 and CMIP5 that show trends of ETCCDI indices significant at 95% level

Region	CMIP Phase	TXx	TXn	TNx	TNn	DTR	TX90p	TX10p	TN90p	TN10p	WSDI	CSDI	SU	TR	FD
AUS	CMIP6	18	12	26	25	5	24	25	31	31	24	30	23	30	22
	CMIP5	14	4	16	16	3	17	12	24	25	15	20	12	22	11
NA	CMIP6	19	8	27	21	7	25	22	30	31	25	30	20	31	13
	CMIP5	12	2	16	13	2	17	11	24	24	16	19	12	22	8
SA	CMIP6	13	13	18	25	4	20	23	30	31	16	25	17	25	21
	CMIP5	9	9	11	15	4	11	10	21	23	8	19	10	18	9
SEA	CMIP6	4	16	12	18	5	17	25	29	28	10	21	14	21	20
	CMIP5	5	10	5	10	5	10	15	19	22	4	9	7	11	9
MEA	CMIP6	8	11	13	22	3	19	15	28	28	16	25	16	25	19
	CMIP5	8	7	9	10	5	11	11	21	21	11	16	13	19	10
TA	CMIP6	19	7	27	21	6	24	21	30	30	25	29	18	29	1
	CMIP5	9	3	16	8	2	17	13	21	22	16	15	10	20	1
SWA	CMIP6	16	11	13	21	2	23	22	30	31	14	23	13	23	11
	CMIP5	13	10	10	13	5	11	11	16	22	5	12	6	14	4
SSA	CMIP6	14	13	9	17	5	17	23	27	28	11	18	13	20	9
	CMIP5	7	8	7	9	4	10	10	18	21	4	12	7	10	7
CAU	CMIP6	18	6	23	20	5	21	16	29	28	22	24	17	26	10
	CMIP5	10	2	16	14	2	14	7	21	22	12	13	10	20	9
MWA	CMIP6	23	8	27	21	3	24	17	30	31	21	25	18	28	7
	CMIP5	16	5	14	13	2	15	7	22	21	15	16	10	17	3

444

445 **5 Discussion and conclusions**

446 This study examines the performance of the newly released CMIP6 models in simulating
447 the 30-year climatologies and time series of extreme temperature indices over Australian regions.
448 Using two observational datasets, AWAP and BEST, as the verification, the historical
449 simulations from 31 CMIP6 models are compared with 26 models from CMIP5. Since extreme
450 temperatures are defined based on TX and TN, we also use Perkins' Skill Score (PSS) to
451 evaluate the models' abilities in simulating the probability distributions of TX and TN, for which
452 we expect more robust conclusions to be obtained.

453 Similar to previous studies, some differences between AWAP and BEST are found to be
454 substantial, implying that multiple observations or reanalysis datasets are needed for the

evaluation studies on climate models (e.g., Alexander & Arblaster, 2017; Herold et al., 2017; Kim et al., 2020; Sillmann et al., 2013; Srivastava et al., 2020). For example, compared to AWAP, the spatial pattern of DTR shown in BEST is smoother, which suggests that the process to minimize the square of the local weather term in the algorithm differs to that of AWAP. Moreover, while AWAP and BEST use a comparable amount of stations in their calculations (Jones et al., 2009; Rohde, Muller, Jacobsen, Muller, et al., 2013; Rohde, Muller, Jacobsen, Perlmutter, et al., 2013), the interpolation procedures in BEST are more complex (Rohde, Muller, Jacobsen, Muller, et al., 2013; Rohde, Muller, Jacobsen, Perlmutter, et al., 2013). Thus, due to different underpinning methods it is not surprising that these observational products yield different ETCCDI values and highlights why multiple datasets should always be used when evaluating climate models.

Although the performance of CMIP6 and CMIP5 models in simulating extreme temperatures are comparable, there are some improvements in CMIP6. For TXx, TNn and SU, the multi-model means and medians of RMSEs in CMIP6 are generally lower. In terms of model ranges in CMIP6, the interquartile model ranges of RMSEs, for some cold extremes (e.g., TNn, TN90p, TN10p, CSDI and FD), are usually narrower; and there are narrower spreads and interquartile model ranges for the temporal trends as well. However, it is noted that the full range of model results should not be considered as uncertainty, and to know whether the model uncertainty is reduced depends on our understanding of the physical processes and feedbacks (Meehl et al., 2020).

With the results from PSS, the RMSEs for some individual models need to be interpreted with caution. For example, as the model MIROC-ES2L is much better at simulating TX than TN, the relatively lower RMSEs of some cold extremes for MIROC-ES2L are doubtful. Moreover,

the lower PSSs and the higher RMSEs for the model NorCPM1confirm that its performance in simulating extreme heat is among the worst performers.

Over the regions SEA, TA and SSA, both CMIP ensembles usually show relatively large deficiencies in simulating temperature extremes. As documented in previous studies, TA can be influenced by the South Pacific convergence zone, tropical cyclones and ENSO (Perkins et al., 2015; Vincent et al., 2011); over southeast Australia, the SAM and the Madden Julian Oscillation are two important factors related to extremes (Parker et al., 2014; Perkins et al., 2015), and in southern Australia, it is generally assumed that there exists a positive relationship between the Indian Ocean Dipole (IOD) and extreme events (White et al., 2014). Moreover, with finer resolution in climate models, which can better represent localized processes (e.g., land surface influences) and topography, the models' performance in simulating extreme temperatures can be further improved over all Australian regions.

In this regional study, it seems that the higher ECS in CMIP6 models does not lead to regional warmer trends in the historical simulations. However, as suggested by Meehl et al. (2020), in order to reproduce the historical temperature response, it is likely that the improved aerosol-cloud interactions in CMIP6 produced large negative radiative forcing, making the ECS in some CMIP6 models larger. A study on future projections over Australia is needed to further investigate if higher ECS leads to regional warmer trends.

This study provides an assessment of the CMIP6 models' ability in simulating extremes, first analyzing the probability distributions of daily-scale weather variables and then calculating the extreme indices, for which more robust conclusions are expected. However, it should be recognized that with more CMIP6 models available, the conclusions may be changed to some

extent. Also, in the future, remote sensing data may be assimilated into the observations, so that robust conclusions over the data-sparse regions like western Australia can be obtained.

Conflict of Interest

The authors declare no financial or other conflicts of interests that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Lisa V. Alexander for feedback and comments and Zeke Hausfather for discussions about Berkeley Earth surface temperature datasets. We also thank two anonymous reviewers, who helped us in improving the quality of the paper. This research/project was undertaken with the assistance of resources and services from the National Computational Infrastructure (NCI), which is supported by the Australian Government. We further acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP and coordinated CMIP5 and CMIP6. We thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies who support CMIP and ESGF. We thank the Bureau of Meteorology, the Bureau of Rural Sciences and CSIRO for providing the Australian Water Availability Project (AWAP) data. Berkeley Earth Data is available from the Berkeley Earth website (http:// berkeleyearth.org/data). S.E.P-K. is supported by ARC grant number FT170100106.

References

Alexander, L. V., & Arblaster, J. M. (2009). Assessing trends in observed and modelled climate extremes over Australia in relation to future projections. *International journal of climatology, 29*(3), 417-435. https://doi.org/10.1002/joc.1730

Alexander, L. V., & Arblaster, J. M. (2017). Historical and projected trends in temperature and precipitation extremes in Australia in observations and CMIP5. *Weather and Climate Extremes, 15*, 34-56. https://doi.org/10.1016/j.wace.2017.02.001

Avila, F. B., Dong, S., Menang, K. P., Rajczak, J., Renom, M., Donat, M. G., et al. (2015). Systematic investigation of gridding-related scaling effects on annual statistics of daily temperature and precipitation maxima: A case study for south-east Australia. *Weather and Climate Extremes, 9*, 6-16. https://doi.org/10.1016/j.wace.2015.06.003

Chen, C.-T., & Knutson, T. (2008). On the verification and comparison of extreme rainfall indices from climate models. *Journal of Climate, 21*(7), 1605-1621. https://doi.org/10.1175/2007jcli1494.1

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

Dey, R., Lewis, S. C., & Abram, N. J. (2019). Investigating observed northwest Australian rainfall trends in Coupled Model Intercomparison Project phase 5 detection and attribution experiments. *International journal of climatology, 39*(1), 112-127. https://doi.org/10.1002/joc.5788

Diffenbaugh, N. S., Singh, D., Mankin, J. S., Horton, D. E., Swain, D. L., Touma, D., et al. (2017). Quantifying the influence of global warming on unprecedented extreme climate events. *Proceedings of the National Academy of Sciences of the United States of America, 114*(19), 4881-4886. https://doi.org/10.1073/pnas.1618082114

Donat, M. G., Alexander, L. V., Yang, H., Durre, I., Vose, R., Dunn, R. J. H., et al. (2013). Updated analyses of temperature and precipitation extreme indices since the beginning of the twentieth century: The HadEX2 dataset. *Journal of Geophysical Research-Atmospheres, 118*(5), 2098-2118. https://doi.org/10.1002/jgrd.50150

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., et al. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development, 9*(5), 1937-1958. https://doi.org/10.5194/gmd-9-1937-2016

Fischer, E. M., & Knutti, R. (2015). Anthropogenic contribution to global occurrence of heavy-precipitation and high-temperature extremes. *Nature Climate Change, 5*(6), 560-564. https://doi.org/10.1038/nclimate2617

Grotjahn, R., Black, R., Leung, R., Wehner, M. F., Barlow, M., Bosilovich, M., et al. (2016). North American extreme temperature events and related large scale meteorological patterns: a review of statistical methods, dynamics, modeling, and trends. *Climate Dynamics, 46*(3), 1151-1184. https://doi.org/10.1007/s00382-015-2638-6

Herold, N., Behrangi, A., & Alexander, L. V. (2017). Large uncertainties in observed daily precipitation extremes over land. *Journal of Geophysical Research-Atmospheres, 122*(2), 668-681. https://doi.org/10.1002/2016jd025842

Herold, N., Ekstrom, M., Kala, J., Goldie, J., & Evans, J. P. (2018). Australian climate extremes in the 21st century according to a regional climate model ensemble: Implications for health and agriculture. *Weather and Climate Extremes, 20*, 54-68. https://doi.org/10.1016/j.wace.2018.01.001

Intergovernmental Panel on Climate Change. (2012). *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change* (C. B. Field, V. Barros, T. F. Stocker, D. Qin, D. J. Dokken, K. L. Ebi, M. D. Mastrandrea, K. J. Mach, G.-K. Plattner, S. K. Allen, M. Tignor, & P. M. Midgley Eds.). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Intergovernmental Panel on Climate Change. (2013). *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, & P. M. Midgley Eds.). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Jones, D. A., Wang, W., & Fawcett, R. (2009). High-quality spatial climate data-sets for Australia. *Australian Meteorological and Oceanographic Journal, 58*(4), 233-248. https://doi.org/10.22499/2.5804.003

Kharin, V. V., Zwiers, F. W., Zhang, X., & Hegerl, G. C. (2007). Changes in temperature and precipitation extremes in the IPCC ensemble of global coupled model simulations. *Journal of Climate, 20*(8), 1419-1444. https://doi.org/10.1175/jcli4066.1

Kharin, V. V., Zwiers, F. W., Zhang, X., & Wehner, M. (2013). Changes in temperature and precipitation extremes in the CMIP5 ensemble. *Climatic Change, 119*(2), 345-357. https://doi.org/10.1007/s10584-013-0705-8

Kim, Y. H., Min, S.-K., Zhang, X., Sillmann, J., & Sandstad, M. (2020). Evaluation of the CMIP6 multi-model ensemble for climate extreme indices. *Weather and Climate Extremes, 29*, 15. https://doi.org/10.1016/j.wace.2020.100269

King, A. D., Alexander, L. V., & Donat, M. G. (2013). The efficacy of using gridded data to examine extreme rainfall characteristics: a case study for Australia. *International journal of climatology, 33*(10), 2376-2387. https://doi.org/10.1002/joc.3588

King, A. D., Karoly, D. J., & Henley, B. J. (2017). Australian climate extremes at 1.5 degrees C and 2 degrees C of global warming. *Nature Climate Change, 7*(6), 412-416. https://doi.org/10.1038/nclimate3296

King, A. D., van Oldenborgh, G. J., Karoly, D. J., Lewis, S. C., & Cullen, H. (2015). Attribution of the record high Central England temperature of 2014 to anthropogenic influences. *Environmental Research Letters, 10*(5), 7. https://doi.org/10.1088/1748-9326/10/5/054002

Klein Tank, A. M. G., Zwiers, F. W., & Zhang, X. (2009). *Guidelines on analysis of extremes in a changing climate in support of informed decisions for adaptation*. Geneva: World Meteorological Organization.

Kumar, D., Kodra, E., & Ganguly, A. R. (2014). Regional and seasonal intercomparison of CMIP3 and CMIP5 climate model ensembles for temperature and precipitation. *Climate Dynamics, 43*(9), 2491-2518. https://doi.org/10.1007/s00382-014-2070-3

Lau, N.-C., & Nath, M. J. (2012). A model study of heat waves over North America: meteorological aspects and projections for the twenty-first century. *Journal of Climate, 25*(14), 4761-4784. https://doi.org/10.1175/jcli-d-11-00575.1

Lewis, S. C. (2018). Assessing the stationarity of Australian precipitation extremes in forced and unforced CMIP5 simulations. *Journal of Climate, 31*(1), 131-145. https://doi.org/10.1175/jcli-d-17-0393.1

Meehl, G. A., Senior, C. A., Eyring, V., Flato, G., Lamarque, J. F., Stouffer, R. J., et al. (2020). Context for interpreting equilibrium climate sensitivity and transient climate response from the CMIP6 Earth system models. *Science Advances, 6*(26), 10. https://doi.org/10.1126/sciadv.aba1981

Meehl, G. A., Tebaldi, C., Walton, G., Easterling, D., & McDaniel, L. (2009). Relative increase of record high maximum temperatures compared to record low minimum temperatures in the U. S. *Geophysical Research Letters, 36*(23), L23701. https://doi.org/10.1029/2009gl040736

Min, S.-K., Zhang, X., Zwiers, F. W., & Hegerl, G. C. (2011). Human contribution to more-intense precipitation extremes. *Nature, 470*(7334), 378-381. https://doi.org/10.1038/nature09763

Parker, T. J., Berry, G. J., Reeder, M. J., & Nicholls, N. (2014). Modes of climate variability and heat waves in Victoria, southeastern Australia. *Geophysical Research Letters, 41*(19), 6926-6934. https://doi.org/10.1002/2014gl061736

Perkins, S. E., Argueso, D., & White, C. J. (2015). Relationships between climate variability, soil moisture, and Australian heatwaves. *Journal of Geophysical Research-Atmospheres, 120*(16), 8144-8164. https://doi.org/10.1002/2015jd023592

Perkins, S. E., Moise, A., Whetton, P., & Katzfey, J. (2014). Regional changes of climate extremes over Australia - a comparison of regional dynamical downscaling and global climate model simulations. *International journal of climatology, 34*(12), 3456-3478. https://doi.org/10.1002/joc.3927

Perkins, S. E., Pitman, A. J., Holbrook, N. J., & 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*(17), 4356-4376. https://doi.org/10.1175/jcli4253.1

Perkins-Kirkpatrick, S. E., & Lewis, S. C. (2020). Increasing trends in regional heatwaves. *Nature Communications, 11*(1), 8. https://doi.org/10.1038/s41467-020-16970-7

Rohde, R., Muller, R., Jacobsen, R., Muller, E., Perlmutter, S., Rosenfeld, A., et al. (2013). A new estimate of the average Earth surface land temperature spanning 1753 to 2011. *Geoinformatics and Geostatistics: An Overview, 1*(1). https://doi.org/10.4172/2327-4581.1000101

Rohde, R., Muller, R., Jacobsen, R., Perlmutter, S., Rosenfeld, A., Wurtele, J., et al. (2013). Berkeley Earth temperature averaging process. *Geoinformatics and Geostatistics: An Overview, 1*(2). https://doi.org/10.4172/2327-4581.1000103

Seneviratne, S. I., & Hauser, M. (2020). Regional climate sensitivity of climate extremes in CMIP6 versus CMIP5 multi-model ensembles. *Earth's Future, 8*(9), e2019EF001474. https://doi.org/10.1029/2019EF001474

Sillmann, J., Kharin, V. V., Zhang, X., Zwiers, F. W., & Bronaugh, D. (2013). Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. *Journal of Geophysical Research-Atmospheres, 118*(4), 1716-1733. https://doi.org/10.1002/jgrd.50203

633 Srivastava, A., Grotjahn, R., & Ullrich, P. A. (2020). Evaluation of historical CMIP6 model simulations of extreme
634 precipitation over contiguous US regions. *Weather and Climate Extremes, 29*, 100268.
635 https://doi.org/10.1016/j.wace.2020.100268
636 Stouffer, R. J., Eyring, V., Meehl, G. A., Bony, S., Senior, C., Stevens, B., et al. (2017). CMIP5 scientific gaps and
637 recommendations for CMIP6. *Bulletin of the American Meteorological Society, 98*(1), 95-105.
638 https://doi.org/10.1175/bams-d-15-00013.1
639 Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of*
640 *the American Meteorological Society, 93*(4), 485-498. https://doi.org/10.1175/bams-d-11-00094.1
641 Tokarska, K. B., Stolpe, M. B., Sippel, S., Fischer, E. M., Smith, C. J., Lehner, F., et al. (2020). Past warming trend
642 constrains future warming in CMIP6 models. *Science Advances, 6*(12), eaaz9549.
643 https://doi.org/10.1126/sciadv.aaz9549
644 Vincent, E. M., Lengaigne, M., Menkes, C. E., Jourdain, N. C., Marchesiello, P., & Madec, G. (2011). Interannual
645 variability of the South Pacific Convergence Zone and implications for tropical cyclone genesis. *Climate*
646 *Dynamics, 36*(9), 1881-1896. https://doi.org/10.1007/s00382-009-0716-3
647 Westra, S., White, C. J., & Kiem, A. S. (2016). Introduction to the special issue: historical and projected climatic
648 changes to Australian natural hazards. *Climatic Change, 139*(1), 1-19. https://doi.org/10.1007/s10584-016-
649 1826-7
650 White, C. J., Hudson, D., & Alves, O. (2014). ENSO, the IOD and the intraseasonal prediction of heat extremes
651 across Australia using POAMA-2. *Climate Dynamics, 43*(7), 1791-1810. https://doi.org/10.1007/s00382-
652 013-2007-2
653 Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., et al. (2020). Causes of
654 higher climate sensitivity in CMIP6 models. *Geophysical Research Letters, 47*(1), e2019GL085782.
655 https://doi.org/10.1029/2019gl085782
656 Zhang, X., Alexander, L., Hegerl, G. C., Jones, P., Tank, A. K., Peterson, T. C., et al. (2011). Indices for monitoring
657 changes in extremes based on daily temperature and precipitation data. *Wiley Interdisciplinary Reviews-*
658 *Climate Change, 2*(6), 851-870. https://doi.org/10.1002/wcc.147
659 Zhang, X., Hegerl, G., Zwiers, F. W., & Kenyon, J. (2005). Avoiding inhomogeneity in percentile-based indices of
660 temperature extremes. *Journal of Climate, 18*(11), 1641-1651. https://doi.org/10.1175/jcli3366.1
661 Zwiers, F. W., Zhang, X., & Feng, Y. (2011). Anthropogenic influence on long return period daily temperature
662 extremes at regional scales. *Journal of Climate, 24*(3), 881-892. https://doi.org/10.1175/2010jcli3908.1
663