AGU ADVANCING EARTH AND SPACE SCIENCE

Geophysical Research Letters

RESEARCH LETTER
10.1029/2020GL091152

Do CMIP6 Climate Models Simulate Global or Regional Compound Events Skillfully?

Nina N. Ridder[1], **Andrew J. Pitman**[1], **and Anna M. Ukkola**[2]

[1]ARC Centre of Excellence for Climate Extremes and Climate Change Research Centre, University of New South Wales, Sydney, NSW, Australia, [2]ARC Centre of Excellence for Climate, Australian National University, Canberra, ACT, Australia

Key Points:

- Some CMIP6 models reproduce observed return periods of co-occurring rain and wind extremes and co-occurring heat waves and droughts well
- CMIP6 models simulate these compound events over North America, Europe, or Eurasia with similar levels of skill
- CMIP6 models simulate these compound events over Australia with lower skill than the other regions analyzed

Supporting Information:

- Supporting Information S1

Correspondence to:
N. N. Ridder,
n.ridder@unsw.edu.au

Citation:
Ridder, N. N., Pitman, A. J., & Ukkola, A. M. (2021). Do CMIP6 climate models simulate global or regional compound events skillfully?. *Geophysical Research Letters*, *48*, e2020GL091152. https://doi.org/10.1029/2020GL091152

Received 12 OCT 2020
Accepted 5 DEC 2020

Abstract Compound events have the potential to cause high socioeconomic and environmental losses. We examine the ability of the sixth phase of the Coupled Model Intercomparison Project (CMIP6) models to capture two bivariate compound events: the co-occurrence of heavy rain and strong wind, and heat waves and meteorological drought. We evaluate the models over North America, Europe, Eurasia, and Australia using observations and reanalysis data set spanning 1980–2014. Some of the CMIP6 models capture the return periods of both bivariate compound events over North America, Europe, and Eurasia surprisingly well but perform less well over Australia. For heavy rain and strong wind, this poor performance was particularly clear in northern Australia which suggests limits in simulating tropical and extratropical cyclones, local convection, and mesoscale convective systems. We did not find higher model resolution improved performance in any region. Overall, our results show some CMIP6 models can be used to examine compound events, particularly over North America, Europe, and Eurasia.

Plain Language Summary Compound events, such as the co-occurrence of heavy rain and strong wind or heat waves and drought, can have major economic, social, and environmental consequences. We therefore ask the question whether the new generation of climate models represented by the sixth phase of the Coupled Model Intercomparison Project (CMIP6) can simulate the occurrence of these important events. We found that some of the CMIP6 models do simulate these compound events surprisingly well over North America, Europe, and Eurasia. Unfortunately, they perform less well over Australia which is likely associated with the problem of simulating extratropical cyclones, local convection, and mesoscale convective systems. Our results suggest that some CMIP6 models can be used to examine these two compound events particularly over North America, Europe, and Eurasia.

1. Introduction

Global climate models are used to simulate climate and weather extremes, including extreme rainfall, high and low temperatures, droughts, and winds. Analyses of observations, historical simulations, and projections of extremes (Alexander, 2016; Bindoff et al., 2013; Sillman et al., 2013) have provided major advances in understanding how the statistics of extremes respond to natural variability and global warming. Many analyses of extremes focus on single hazards, such as how hot is the hottest day each year, or how much rain fell during the rainiest 5-day stretch of the year. An evaluation of models included in the fifth phase of the Coupled Model Intercomparison Project (CMIP5) highlights that extremes are generally more difficult to represent realistically than the average (Flato et al., 2013; Sillman et al., 2013b). For instance, Flato et al. (2013) note that CMIP5 models generally capture observed trends in temperature extremes, but rainfall extremes are more challenging, although this might be partly due to higher observational uncertainty. Since this assessment, extensive literature has emerged demonstrating the improved skill of climate models in simulating temperature (e.g., Di Luca et al., 2020) and rainfall extremes (Bador et al., 2020), particularly in hot and cold extremes (Di Luca et al., 2020) and the intensity of heavy rainfall (Kim et al., 2020). The evaluation of wind extremes is more limited, but Kumar et al. (2015) noted that CMIP5 models simulated the multimodel mean (MMM) of spatial patterns of extreme winds with 25–100-year return periods (RPs) well.

In the last decade, compound events (CEs) have emerged as a focus for understanding the link between changes in weather and climate and impacts on vulnerable systems (Leonard et al., 2014; Seneviratne et al., 2012; Zscheischler et al., 2020). Despite their potential to cause socioeconomic impacts, CEs have

received little attention in climate model evaluations. As distinct to single hazards, CEs are caused by the joint occurrence of two or more hazards and/or drivers (Zscheischler et al., 2018). Impacts from weather and climate are commonly the result of CEs. For example, rainfall that might not constitute as treacherous in isolation, co-occurring in time and space with winds that are not technically extreme can combine to generate significant hazards (De Luca et al., 2020). Here, we focus on CEs defined as the joint exceedance of two climate variables or indices with predefined thresholds.

A key issue in evaluating CEs in climate models was the lack of a suitable climatology as baseline. Ridder et al. (2020) provided a first step in resolving this by using a combination of high-quality station data, supplemented by reanalyses, to provide a global climatology of multivariate CEs for 27 hazard pairs based on temperature, rainfall, wind, streamflow, and storm surge data. We use these data to evaluate the skill of CMIP6 models in capturing the probability of bivariate CEs. This is timely as studies are now emerging using climate models to study CEs for individual variable pairs (e.g., Vogel et al., 2020). We first explore the skill of the CMIP6 models for two kinds of CEs for which model outputs were available (heavy rainfall co-occurring with strong wind, and heat waves (HWs) co-occurring with meteorological drought). We then examine how consistent model skill is across four relatively data-rich regions: North America, Western Europe, Eurasia, and Australia.

2. Materials and Methods

2.1. Overview

We evaluate the performance of a subset of CMIP6 models in reproducing observed bivariate CEs. Following Ridder et al. (2020), we adopt the definition of CEs introduced by Zscheischler et al. (2018). We focus exclusively on events in the category called "multihazard CEs" (hereafter CEs) in which two (or more) hazards simultaneously affect one region (Zscheischler et al., 2020).

We were unable to evaluate the full 27 hazard combinations reported by Ridder et al. (2020) with CMIP6 data because many drivers of CEs including daily storm surge and streamflow data are not directly available from the CMIP6 models. Therefore, we consider two bivariate combinations, namely daily precipitation sums and daily mean 10-m wind speed (WP-CEs hereafter), and Excess Heat Factor (EHF) and Standardized Precipitation Index (SPI) calculated at a 3-month scale as a measure for HWs and the occurrence of a meteorological drought (HD-CEs hereafter). The focus on these two combinations was based on data availability, as well as the recognition of their potential to cause socioeconomic and environmental impacts and their relative importance in many geographic regions (Ridder et al., 2020). These variables also provide a first test of the models for three key variables simulated by CMIP6 models: wind, precipitation, and temperature.

To identify the occurrence of bivariate CEs, each of the four meteorological variables was assigned a fixed threshold to identify potentially hazardous conditions. We used the thresholds from Ridder et al. (2020) to allow a direct comparison between their observational results and the CMIP6 models. Accordingly, daily precipitation sums and mean 10-m winds are considered extreme if they exceed their respective 99th percentile, a HW occurs if the EHF exceeds $0^{\circ}C^{2}$, and a meteorological drought arises when SPI ≤ -1.3 for the base period spanning 1980–2014. We define a CE as the simultaneous exceedance of the two components in a variable pair at the same location, that is, in the same grid cell, on the same day (see Section 2.3). Due to the extended nature of EHF and SPI, that is, at least three consecutive days of above-average temperatures and a 3-month period of low precipitation, HD-CEs are likely to occur over multiple consecutive days. Nevertheless, we treat them as daily events to be consistent with the analysis of WP-CEs.

2.2. Data

Observed joint RP maps for the time period from 1980 to 2014 for HD-CEs were taken directly from Ridder et al. (2020). The climatology for WP-CEs based on daily mean wind speed and precipitation sums was generated following the method applied by Ridder et al. (2020), who used daily maximum instead of mean wind speeds. HWs were calculated from HadGHCND daily maximum and minimum temperatures (Caesar et al., 2006). The two precipitation-related hazards, heavy precipitation and meteorological drought, are

based on the REGEN data set (Contractor et al., 2020). Wind speed was calculated from the meridional and zonal components of 10-m wind of the ERA-Interim reanalysis data set provided by the European Center for Medium-Range Weather Forecasts (Berrisford et al., 2011, p. 23).

We include 23 CMIP6 models for WP-CEs and 16 for HD-CEs—all the models with daily data available for the considered variable pairs (Table S1). We used daily data from the historical run spanning 1980–2014 utilizing one ensemble member (r1i1p1f1) per model. EHF was calculated for each model using the model's daily maximum and minimum near-surface temperatures on the model's native grid where available or one of the CMIP6 target grids following Perkins and Alexander (2013) using python code developed by Tammas Loughran (https://github.com/tammasloughran/ehfheatwaves). SPI was calculated from monthly precipitation at the 3-month scale using the ClimPact2 software package on the same grid as EHF (https://github.com/ARCCSS-extremes/climpact2). ClimPact2 relies on the SPEI R package (Vicente-Serrano et al., 2010), which uses the gamma distribution to calculate SPI. Monthly SPI values are transformed to daily data to match the EHF timeseries with days taking the value of the associated month. Both climate indices and variables were bilinearly interpolated to a 2.5° × 2.5° regular grid to match Ridder et al. (2020).

2.3. Statistical Methods

We present results as RPs, that is, the inverse of occurrence probability, following Ridder et al. (2020). These are determined for each CMIP6 model individually on a grid cell basis following the methodology of Ridder et al. (2020) using the thresholds in Section 2.1. Following Ridder et al. (2020), we omit grid cells from our analysis where percentile thresholds are below the minimum threshold for wind (<0.5 m/s) and/or precipitation (<1 mm/day) and those without CE occurrences throughout the study period (1980–2014).

A grid cell in a CMIP6 model is considered to experience a CE if both hazards in a pair simulated by the same model exceed their respective threshold on the same day. We calculate the joint occurrence probability empirically using the number of daily time steps over the 1980–2014 period (12,874 days for models with leap years and 12,775 days for those without). Note that the calculation of EHF uses a nonleap year calendar. Therefore, probabilities for HD-CEs are based on 12,775 days regardless of individual CMIP6 model calendar. To obtain RPs in years, the inverse of the probability is then divided by 365.

The RPs of CEs in each model are compared to observed RPs (Ridder et al., 2020) using the probability density function (PDF) of the RPs in four different geographic regions (North America, Western Europe, Eurasia, and Australia; Figure S1). These regions were chosen based on the availability of reliable long-term observations and statistically significant RPs identified in Ridder et al. (2020).

To assess model skill for reproducing observed RPs, we applied the skill score (S_{skill}) based on Perkins et al. (2007). S_{skill} compares similarities between the probability density of a given distribution derived from a model (Z_i^{mod}) with that of observations (Z_i^{obs}). This method divides the modeled and observed PDFs into n bins ($n \in \mathbb{N}$; here, $n = 100$ resulting in a bin width of about 4 months, that is, one season). For each bin, the modeled ($Z_i^{\text{mod}}; i \in \left[1,2,\ldots,n\right]$) and observed probability densities ($Z_i^{\text{obs}}; i \in [1,2,\ldots,n]$) are compared. S_{skill} is the minimum of this comparison summed over all bins:

$$S_{\text{skill}} = \sum_{i=1}^{n} \text{minimum}\left(Z_i^{\text{obs}}, Z_i^{\text{mod}}\right). \quad (1)$$

Consequently, S_{skill} ranges between 0 and 1 ($S_{\text{skill}} \in \{\mathbb{R}^+ \mid S_{\text{skill}} \in [0,1]\}$). A value of 1 indicates a perfect match to observations and 0 no model skill. Here, S_{skill} does not measure the skill over the whole PDF as in Perkins et al. (2007). Rather, by focusing on the RPs, that is, the probability of joint exceedance of extreme thresholds, S_{skill} highlights a model's ability to reproduce the joint tail of the variable pair. This is particularly useful when one is interested in the occurrence probability of a CE, and the absolute values for the two variables and their thresholds are less important.

For the WP-CEs, we also considered a bivariate skill score ($S_{\text{skill}}^{\text{bivariate}}$) to cover the full distribution of both variables and assess model performance in reproducing the observed correlation between the two. For this, we extended the skill score presented in Perkins et al. (2007) to include two dimensions, one for precipitation and the other for wind speed. We divided the joint distribution into $p \times m$ ($p,m \in \mathbb{N}$) two-dimensional bins.

Here, p and m are chosen depending on the variable based on a trade-off between the recommended bin size following Scott's Normal Reference (Scott, 1979) and computational cost (400 for precipitation and 300 for wind speed). The skill score is then derived as

$$S_{\text{skill}}^{\text{bivariate}} = \sum_{j=1}^{m}\sum_{i=1}^{p}\text{minimum}\left(Z_{j,i}^{\text{obs}}, Z_{j,i}^{\text{mod}}\right) \quad (2)$$

This measure assesses the full individual distribution of both variables that make up the CE and simultaneously measures their correlation. For example, if a model underestimates rainfall, the $S_{\text{skill}}^{\text{bivariate}}$ will be lower even if the correlation between rainfall and wind is modeled correctly. If the correlation between strong wind and heavy rainfall is not captured properly $S_{\text{skill}}^{\text{bivariate}}$ likewise decreases. As such, $S_{\text{skill}}^{\text{bivariate}}$ is a useful tool to determine model performance when one is interested in changes in the absolute values of RPs.

Note that theoretically a high S_{skill} can be obtained despite a low $S_{\text{skill}}^{\text{bivariate}}$ since the two skill scores assess different parts and characteristics of the joint distribution and due to the independence of S_{skill} from model biases in the individual variables involved in a CE.

3. Results

3.1. Bias in Global RPs: Strong Winds and Heavy Precipitation

The MMM bias for RPs of WP-CEs suggests that there is a tendency to overestimate RPs over Europe, Australia, North Africa, India, and China and underestimate RPs over much of Eurasia, southern South America, and southern Africa (Figures 1a and 1b). However, the models show relatively little consistency in the sign of the bias (Figure 1b).

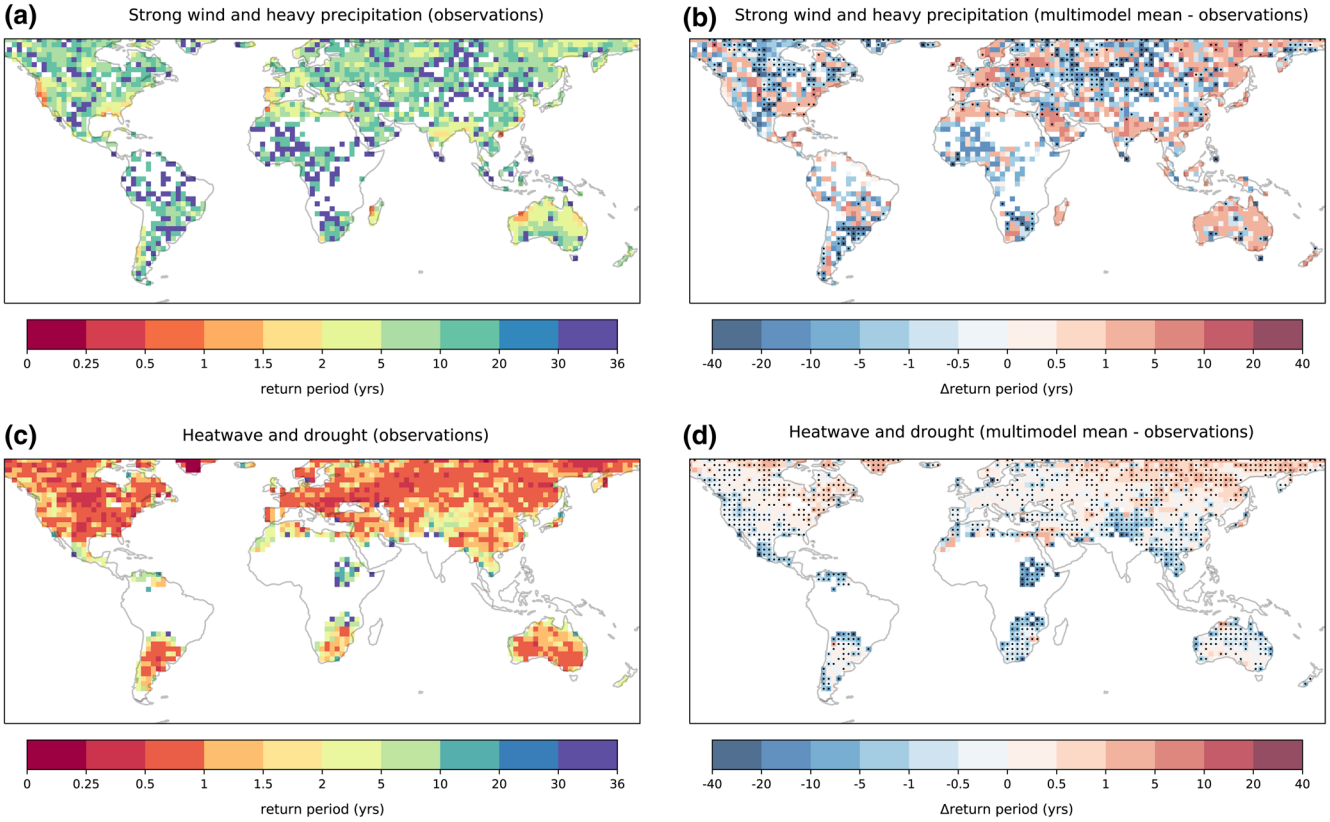

Figure 1. (a) Joint return periods of compound events consisting of strong daily mean winds and high daily precipitation sums in observations. (b) Multimodel mean bias in joint return periods for the co-occurrence of strong winds and heavy precipitation. (c and d) As (a and b) but for compound events consisting of heat waves and meteorological drought. In panels (b and d), stippled regions indicate grid cells where at least 75% of the CMIP6 models agree on the sign of bias. CMIP6, sixth phase of the Coupled Model Intercomparison Project.

The observations suggest that the RP for WP-CEs is mostly >5 years (Figure S2). Details on the mechanisms causing the spatial pattern of RPs are discussed in Ridder et al. (2020). Some models seem to capture this well (INM-CM4-8, INM-CM5-0, and IPSL-CM6-LR) while many models estimate the RPs < 1 year in many regions. This suggests that the models overestimate the correlation between precipitation and wind. Nevertheless, given previous assessments that found CMIP6 models simulate heavy rainfall relatively poorly (Bador et al., 2020; Kim et al., 2020) and the relative lack of analyses of strong winds, many CMIP6 models capture WP-CEs surprisingly well. However, there are areas where individual models fail to reflect the observations well. For example, TaiESM1, MRI-ESM2.0, and GFDL-ESM4 all underestimate RPs over North America while BCC-ESM1, BCC-CSM2-MR, GFDL-CM4, FGOALS-f3-L, GFDL-ESM4, and SAM0-UNICON all underestimate RPs over India and south-east Asia.

3.2. Bias in Global RPs: HW and Meteorological Drought

The MMM for HD-CEs reproduces observations relatively well in most regions with a positive bias in RPs largely <0.5 years at midlatitudes to high latitudes (Figures 1c and 1d). In the Northern Hemisphere, the MMM displays a consistent positive bias with considerable consistency across models (Figure 1d). However, the magnitude of the bias is relatively low with observed RPs (Figure 1c) mostly in the range of 0.5–1.5 years and the model bias mostly within 0.5 year. In the Southern Hemisphere, there are two large regions with a systematic negative bias over tropical southern South America and southern Africa corresponding to regions with longer observed RPs. Several models simulate high RPs for HD-CEs (BCC-CSM2-MR, INM-CM4-8, and NorESM2-MM) but this is strongly influenced by patterns in the tropics and subtropics where observations are lacking (Figure S3). While there are therefore major differences between CMIP6 models over Africa, India, and South America, with RPs ranging from <1 to >5 years, it is impossible to determine which is consistent with observations.

3.3. Comparison Between Regions: Strong Winds and Heavy Precipitation

A regional analysis (see Figure S1 for regions) of WP-CEs highlights a strong tendency for many models to overestimate low RPs (Figure 2; left column) and thereby underestimate the occurrence of events with RPs of >15 years particularly over North America and Eurasia. The bar charts in Figure 2 (right column) quantify the skill of the CMIP6 models across the whole joint PDF ($S_{\text{skill}}^{\text{bivariate}}$; crosses) and over the upper tail of the joint distribution, that is, the joint RP (S_{skill}; colored bars). In general, $S_{\text{skill}}^{\text{bivariate}}$ indicates relatively high skill for many CMIP6 models, with $S_{\text{skill}}^{\text{bivariate}} > 0.6$ over virtually all models in every region. The MMM $S_{\text{skill}}^{\text{bivariate}}$ is 0.81 in North America, Western Europe, and Eurasia, and 0.79 in Australia. Thirteen models achieve $S_{\text{skill}}^{\text{bivariate}} > 0.8$ over North America, 15 over Europe, 14 over Eurasia, and 9 over Australia. Only one model simulates the $S_{\text{skill}}^{\text{bivariate}} < 0.6$ (TaiESM1 over Australia). The performance of models in terms of S_{skill} (bars in Figure 2) varies strongly. For example, MPI-ESM1-2-HR reproduces observed RPs quite well in North America, Western Europe, and Eurasia with S_{skill} around 0.8 (Figure 2). CMIP6 models that capture both $S_{\text{skill}}^{\text{bivariate}}$ and $S_{\text{skill}} \gtrsim 0.8$ include MPI-ESM-1-2-HAM over North America, Europe, and Eurasia, and MPI-ESM-1-2-HR over North America and Eurasia. However, both these models perform poorly over Australia.

3.4. Comparison Between Regions: HW and Meteorological Drought

Over all four regions, CMIP6 models produce slightly shorter RPs for HD-CEs than observations (Figure 3, left column). However, differences in PDF maxima between models and observations are generally <1 year. This is in line with the low MMM bias in joint RPs (Figure 1d). Consequently, it is not surprising to see more models with $S_{\text{skill}} > 0.8$ than for WP-CEs (Figure 3; right column). For North America and Western Europe, only one model shows $S_{\text{skill}} \leq 0.6$, namely NorESM2-MM and INM-CM4-8, respectively. Both these models show $S_{\text{skill}} < 0.6$ in Eurasia and Australia. As for WP-CEs, more models perform worse over Australia than over the other regions, with the exception of BCC-ESM1, INM-CM5-0, and SAM0-UNICON. Particularly, noteworthy is the low skill of BCC-CSM2-MR over Australia ($S_{\text{skill}} < 0.1$), while showing S_{skill} is >0.8 in Western Europe and ~0.7 in North America and Eurasia.

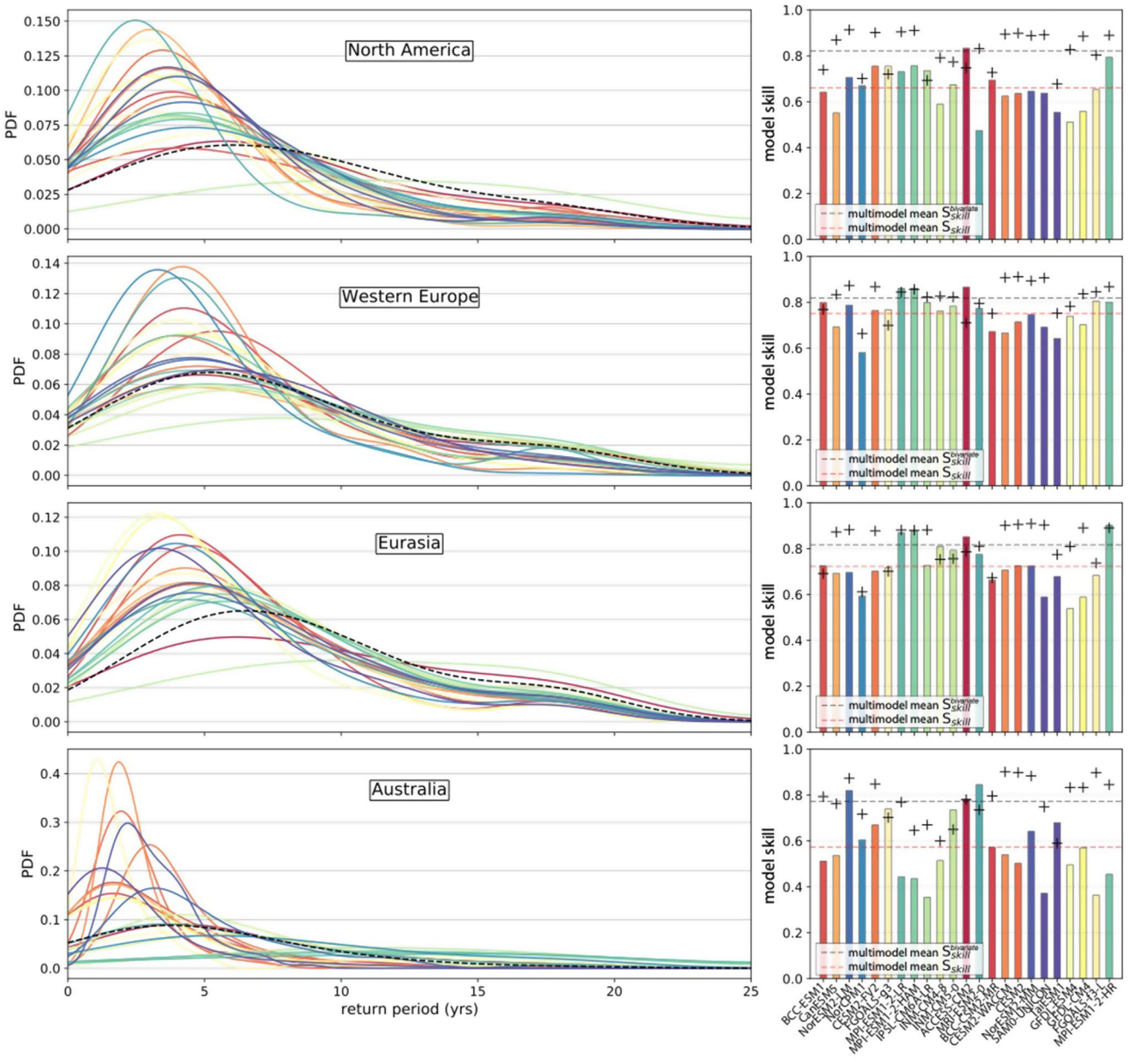

Figure 2. Comparison of model performance for the co-occurrence of strong winds and heavy precipitation. Left column: Probability density functions (PDFs) of return periods of all models (colored lines) and observations (black dashed line) in four geographic regions. Note the different *y* axis scales between the subplots. Right column: Model skill scores S_{skill} (bars) and $S_{skill}^{bivariate}$ (crosses) for all models. The color scheme in the PDFs is the same as for the bar plots. The models are ordered from lowest resolution on the left, to highest on the right. The black and red dashed lines indicate the multimodel mean value of $S_{skill}^{bivariate}$ and S_{skill}, respectively.

3.5. CMIP6 Model Skill Over Australia for Strong Wind and Heavy Precipitation

Our results suggest that WP-CEs are simulated less well in Australia than elsewhere with a MMM $S_{skill} < 0.6$ ($S_{skill}^{bivariate} < 0.8$) over Australia compared to $S_{skill} > 0.7$ ($S_{skill}^{bivariate} > 0.8$) over other regions (Figure 2). Figure 4a shows the correlation for each CMIP6 model's performance in $S_{skill}^{bivariate}$ for Eurasia, Western Europe, and Australia using North America as a reference. Some models simulate higher $S_{skill}^{bivariate}$ for Eurasia and Western Europe than North America. However, for most models, and particularly models with $S_{skill}^{bivariate} > 0.6$, CMIP6 model performance is largely comparable across these three regions. However, 16 out of the 23 models perform worse over Australia than North America, apart from BCC-ESM1, NorCPM1, ACCESS-CM2,

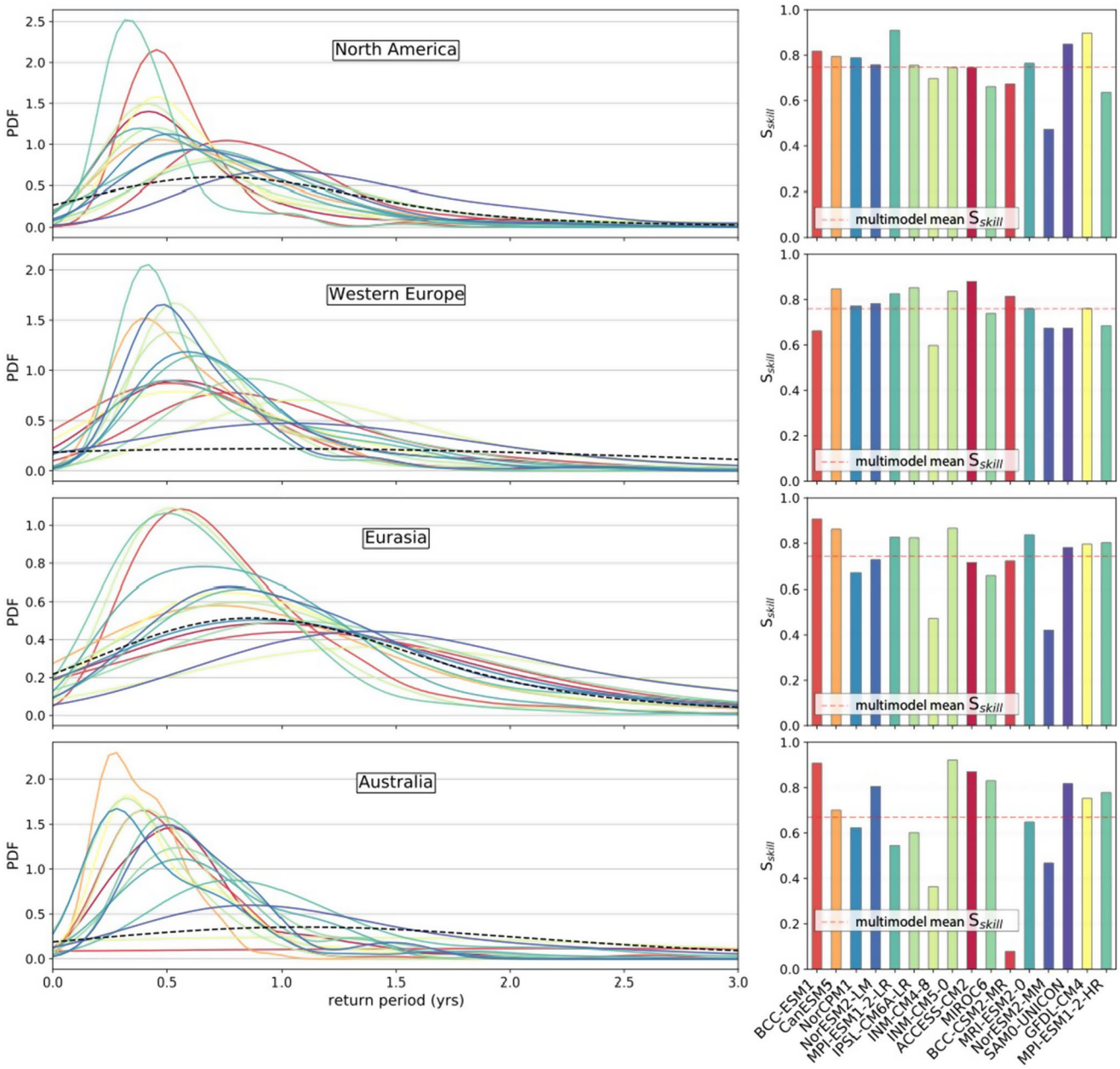

Figure 3. Comparison of model performance for the co-occurrence of heat waves and meteorological drought. Left column: Probability density functions (PDFs) of return periods of all models (colored lines) and observations (black dashed line) in the four different geographic regions. Note the different *y* axis scales between the subplots. Right column: Model skill score S_{skill} for all models. The color scheme in the PDFs is equivalent to the bar plots. The models are ordered from lowest resolution on the left, to highest on the right. The red dashed line indicates the multimodel mean value of S_{skill}.

BCC-CSM2-MR, CESM2-WACCM, GFDL-ESM4, and FGOALS-f3-L. Similarly, models with high S_{skill} in Eurasia and Western Europe also perform well in North America, but over Australia, all but five models perform worse than in North America (Figure 4b). Using any of the other regions as reference leads to similar conclusions about relative regional performance (not shown).

To examine the relatively poor performance over Australia for WP-CEs, we divided Australia at 26°S into the northern region dominated by tropical conditions and the southern region dominated midlatitude subtropical and temperate conditions. The MMM for $S_{skill}^{bivariate}$ in northern Australia is ~0.7 (Figure S4), lower than the 0.8 in southern Australia and the other regions (Figure 2). The CMIP6 models simulate the southern region indistinguishably from the other regions for $S_{skill}^{bivariate}$ (see regression line aligned with the other three regions

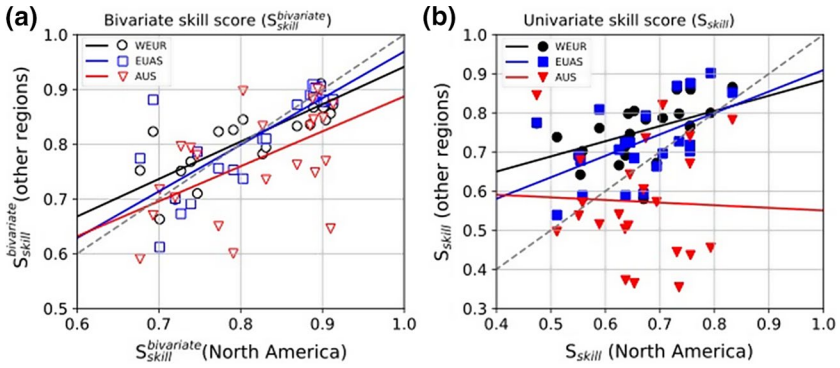

Figure 4. Comparison of model performance for WP-CEs in North America compared to Western Europe (black circles), Eurasia (blue squares), and Australia (red triangles). (a) The bivariate skill score $S_{\text{skill}}^{\text{bivariate}}$ calculated over the full joint distributions for the high precipitation, wind hazard pair. (b) The skill score over the tail of the joint distribution (S_{skill}). The solid lines show the line of best fit for each region (Western Europe: black; Eurasia: blue; Australia: red) and the gray dashed line shows the one-to-one line.

in Figure S5a). However, the CMIP6 models perform less well in general for the northern region (red line in Figure S5a). While some CMIP6 models do poorly over both northern Australia and North America (TaiESM1 with $S_{\text{skill}}^{\text{bivariate}} < 0.6$ over North Australia and with $S_{\text{skill}}^{\text{bivariate}} < 0.7$ in North America; the worst performing model there) other models show $S_{\text{skill}}^{\text{bivariate}} < 0.5$ over northern Australia while demonstrating higher skill over North America. For S_{skill} (Figures S4 and S5b), 17 out of 23 CMIP6 models demonstrate quite low skill (<0.6) over southern Australia. Two CMIP6 model groups form for northern Australia with 15 models with $S_{\text{skill}} < 0.5$, and six reaching $S_{\text{skill}} \gtrsim 0.6$ (Figure S5b). These six skilled models capture northern Australia as well as other regions. These models include CESM-FV2, FGOALS-g3, INM-CM5-0, ACCESS-CM2, MRI-ESM2-0, and NorESM2-MM (Figure S4) but do not include NorESM2-LM, NorCPM1, CESM2, FGOALS-f3-L, or INM-CM4.8 indicating that models from the same family can display very different levels of skill.

4. Discussion and Conclusions

The assessment of extremes in climate model simulations commonly finds a higher fidelity in the simulation of temperature extremes compared to rainfall or wind extremes. However, the consequences of extremes events to society, business, or the environment often emerge when several extremes co-occur in time and space. These CEs are increasingly acknowledged as important to risk assessment. Therefore, evaluating how well CMIP6 models capture their occurrence probability is a necessary step before examining how these might change in the future.

We evaluated the skill of CMIP6 models in capturing the joint probability of strong winds and heavy rainfall, and HWs and meteorological drought. Strong winds and heavy precipitation have physical connections and common meteorological phenomena can cause a correlation between the two (e.g., cyclones, midlatitude mesoscale systems, and intense convection). However, CMIP6 models usually lack the spatial resolution to resolve these synoptic-scale phenomena well. HWs and meteorological drought also have a physical connection, with meteorological drought causing evaporation to become water limited resulting in a reduction of evaporative cooling and increased sensible heating (Donat et al., 2018; Miralles et al., 2018). Conversely, HWs require several sequential hot days often associated with blocking which climate models frequently struggle to capture (Scaife et al., 2010) despite increased performance in CMIP6 compared to CMIP5 (Schiemann et al., 2020). Given these known shortcomings, low skill in CEs may be anticipated. Yet, we found some CMIP6 models captured the RPs of the two bivariate CEs well in the relatively data-rich regions of North America, Europe, and Eurasia where the estimates of observed RPs are more robust. We noted very wide disagreement between the CMIP6 models over Africa, South America, and India but could not evaluate the models in these regions due to limited local observations. We also highlighted that the skill displayed by the CMIP6 models was broadly similar between North America, Europe, and Eurasia (Figure 4) but lower over Australia.

Spatial resolution is a possible explanation for the different performance of CMIP6 models. Such a connection has been found for atmospheric blocking (Flato et al., 2013) and heat extremes (Di Luca et al., 2020). However, we found no relationship between resolution and either $S_{\text{skill}}^{\text{bivariate}}$ or S_{skill} for either CE (right columns in Figures 2 and 3). We note the highest horizontal resolution used here is ~100 km (T127 spectral grid, MPI-ESM1-2-HR) which is still below the resolution likely to enable the models to capture the weather scales that tend to characterize extremes.

The weak performance of the CMIP6 models in simulating WP-CEs over northern Australia is most likely associated with the nature of rainfall extremes in this region. Utsumi et al. (2007) examined weather systems (tropical cyclone, extratropical cyclone, etc.) associated with daily extreme precipitation. They found that, while southern Australia sourced ~50% of its extreme rainfall from fronts and ~50% from extratropical cyclones (Utsumi et al., 2007), northern Australia's extreme rainfall originates mostly from extratropical cyclones, combined with local convection and mesoscale convective systems. This suggests that the poor performance in some models in simulating WP-CEs over northern Australia may be associated with challenges in simulating these phenomena. Other regions, which receive most extreme rainfall from extratropical cyclones, include the Indian subcontinent and Mediterranean (Utsumi et al., 2007). We therefore suggest caution when exploring WP-CEs with CMIP6 models in regions where extreme rainfall is associated with extratropical cyclones or local convection and mesoscale convective systems.

To simulate HD-CEs well, models must capture the individual HW and drought hazards and their interactions through land-surface feedbacks. The historical duration and frequency of seasonal-scale meteorological droughts are generally simulated well by CMIP6 models, although biases remain particularly in the tropics (Ukkola et al., 2018, 2020). These seasonal-scale droughts are relatively common in many mid-latitude regions (e.g., Europe). Meanwhile, drier regions like Australia tend to experience longer annual-scale droughts driven by interacting modes of variability (De Luca et al., 2020; Santoso et al., 2019; Wang et al., 2014). CMIP models lack the skill to simulate these longer events (Moon et al., 2018; Ukkola et al., 2020). Next, coincident with drought, the synoptic conditions need to be conducive to HWs (commonly atmospheric blocking with clear skies, high incident solar radiation) and need to be amplified by surface fluxes where appropriate (Donat et al., 2017). Land-surface models systematically underestimate latent heat fluxes during drought conditions (Trugman et al., 2018; Ukkola et al., 2016), leading to an overamplification of heat extremes in coupled models through land-surface feedbacks, particularly in humid regions (Sippel et al., 2017; Ukkola et al., 2018). This may contribute to the overestimation of the occurrence of HD-CEs in many regions (Figures 1d and S3).

We showed that CMIP6 model skill can differ strongly between the two performance metrics in this study, that is, S_{skill} and $S_{\text{skill}}^{\text{bivariate}}$. This highlights the need to base model selection on the focus of a study. For instance, a model showing high skill in terms of RPs (S_{skill}) has potential for the assessment of relative change in joint occurrence probability and the model could be useful in assessing the change in RPs of CEs. In contrast, if the main focus of an assessment is the change in the absolute value considered a joint extreme, models with a higher $S_{\text{skill}}^{\text{bivariate}}$ are likely more useful as these show combined skill in reproducing the univariate distributions and the physical correlation between the two variables in question.

Overall, we demonstrate that some CMIP6 models simulate RPs of WP-CEs and HD-CEs with useful skill over multiple regions. Inevitably, some CMIP6 models fail to capture the observations providing an opportunity to examine the parameterizations used in these models with the more skillful models to help direct model development. We only evaluated two possible CEs due to data constraints. Future work should extend the evaluation to additional CEs given their importance in driving socioeconomic risk.

Data Availability Statement

The CMIP6 outputs used in this study are available from the Earth System Grid Federation (https://esgf-node.llnl.gov). The availability of observational data sets is indicated in Ridder et al. (2020).

Acknowledgments
The research was funded by the Australian Research Council Center of Excellence for Climate Extremes (CE170100023) and was supported in part by the New South Wales Department of Planning, Industry and Environment. We are grateful to the National Computational Infrastructure at the Australian National University and the Earth System Grid Federation for making the CMIP6 model outputs available. We also thank Christian Jakob for helpful discussions on this study. We acknowledge the World Climate Research Program, which, through its Working Group on Coupled Modeling, coordinated and promoted 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 CMIP6 and ESGF.

References

Alexander, L. V. (2016). Global observed long-term changes in temperature and precipitation extremes: A review of progress and limitations in IPCC assessments and beyond. *Weather and Climate Extremes*, *11*, 4–16. https://doi.org/10.1016/j.wace.2015.10.007

Bador, M., Boé, J., Terray, L., Alexander, L. V., Baker, A., Bellucci, A., et al. (2020). Impact of higher spatial atmospheric resolution on precipitation extremes over land in global climate models. *Journal of Geophysical Research: Atmospheres*, *125*, e2019JD032184. https://doi.org/10.1029/2019JD032184

Berrisford, P., Dee, D. P., Poli, P., Brugge, R., Fielding, M., Fuentes, M., et al. (2011). *The ERA-Interim archive Version 2.0Rep*. Shinfield Park, Reading, UK: ECMWF.

Bindoff, N. L., Stott, P. A., AchutaRao, K. M., Allen, M. R., Gillett, N., Gutzler, D., et al. (2013). Detection and attribution of climate change: From global to regional. In T. F. Stocker, et al. (Eds.), *Climate change 2013: The physical science basis. Contribution of working group I to the fifth assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, UK/New York, NY: Cambridge University Press.

Caesar, J., Alexander, L., & Vose, R. (2006). Large-scale changes in observed daily maximum and minimum temperatures: Creation and analysis of a new gridded data set. *Journal of Geophysical Research*, *111*, D05101. https://doi.org/10.1029/2005JD006280

Contractor, S., Donat, M. G., Alexander, L. V., Ziese, M., Meyer-Christoffer, A., Schneider, U., et al. (2020). Rainfall estimates on a gridded network (REGEN)—A global land-based gridded dataset of daily precipitation from 1950 to 2016. *Hydrology and Earth System Sciences*, *24*(2), 919–943. https://doi.org/10.5194/hess-24-919-2020

De Luca, P., Messori, G., Pons, F. M. E., & Faranda, D. (2020). Dynamical systems theory sheds new light on compound climate extremes in Europe and Eastern North America. *Quarterly Journal of the Royal Meteorological Society*, *146*(729), 1636–16350. https://doi.org/10.1002/qj.3757

De Luca, P., Messori, G., Wilby, R. L., Mazzoleni, M., & Di Baldassarre, G. (2020). Concurrent wet and dry hydrological extremes at the global scale. *Earth System Dynamics*, *11*, 251–266. https://doi.org/10.5194/esd-11-251-2020

Di Luca, A., Pitman, A. J., & de Elía, R. (2020). Decomposing temperature extremes errors in CMIP5 and CMIP6 models. *Geophysical Research Letters*, *47*. e2020GL088031. https://doi.org/10.1029/2020GL088031

Donat, M. G., Pitman, A. J., & Angelil, O. (2018). Understanding and reducing future uncertainty in mid-latitude daily heat extremes via land surface feedback constraints. *Geophysical Research Letters*, *45*, 10627–10636. https://doi.org/10.1029/2018GL079128

Donat, M. G., Pitman, A. J., & Seneviratne, S. I. (2017). Regional warming of hot extremes accelerated by surface energy fluxes. *Geophysical Research Letters*, *44*, 7011–7019. https://doi.org/10.1002/2017GL073733

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., et al. (2013). Evaluation of climate models. In T. F. Stocker, et al. (Eds.), *Climate change 2013: The physical science basis. Contribution of working group I to the fifth assessment report of the Intergovernmental Panel on Climate Change*. Cambridge, UK/New York, NY: Cambridge University Press.

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*, 100269. https://doi.org/10.1016/j.wace.2020.100269

Kumar, D., Mishra, V., & Ganguly, A. R. (2015). Evaluating wind extremes in CMIP5 climate models. *Climate Dynamics*, *45*, 441–453. https://doi.org/10.1007/s00382-014-2306-2

Leonard, M., Westra, S., Phatak, A., Lambert, M., van den Hurk, B., McInnes, K., et al. (2014). A compound event framework for understanding extreme impacts. *Wiley Interdisciplinary Reviews: Climate Change*, *5*, 113–128. https://doi.org/10.1002/wcc.252

Miralles, D. G., Gentine, P., Seneviratne, S. I., & Teuling, A. J. (2018). Land–atmospheric feedbacks during droughts and heatwaves: State of the science and current challenges. *Annals of the New York Academy of Science*, *1436*, 19–35. https://doi.org/10.1111/nyas.13912

Moon, H., Gudmundsson, L., & Seneviratne, S. I. (2018). Drought persistence errors in global climate models. *Journal of Geophysical Research: Atmospheres*, *123*, 3483–3496. https://doi.org/10.1002/2017JD027577

Perkins, S., Pitman, A., Holbrook, N., & 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, S. E., & Alexander, L. V. (2013). On the measurement of heat waves. *Journal of Climate*, *26*(13), 4500–4517. https://doi.org/10.1175/JCLI4253.1

Ridder, N., Pitman, A. J., Westra, S., Ukkola, A., Do, H., Bador, M., et al. (2020). Global hotspots for the occurrence of compound events. *Nature Communications*, *11*, 5956. https://doi.org/10.1038/s41467-020-19639-3

Santoso, A., Hendon, H., Watkins, A., Power, S., Dommenget, D., England, M. H., et al. (2019). Dynamics and Predictability of El Niño–Southern Oscillation: An Australian perspective on progress and challenges. *Bulletin of the American Meteorological Society*, *100*(3), 403–420. https://doi.org/10.1175/BAMS-D-18-0057.1

Scaife, A. A., Woollings, T., Knight, J., Martin, G., & Hinton, T. (2010). Atmospheric blocking and mean biases in climate models. *Journal of Climate*, *23*, 6143–6152. https://doi.org/10.1175/2010JCLI3728.1

Schiemann, R., Athanasiadis, P., Barriopedro, D., Doblas-Reyes, F., Lohmann, K., Roberts, M. J., et al. (2020). Northern Hemisphere blocking simulation in current climate models: Evaluating progress from the Climate Model Intercomparison Project Phase 5 to 6 and sensitivity to resolution. *Weather and Climate Dynamics*, *1*, 277–292. https://doi.org/10.5194/wcd-1-277-2020

Scott, D. (1979). On optimal and data-based histograms. *Biometrika*, *66*, 605–610. https://doi.org/10.1093/biomet/66.3.605

Seneviratne, S. I., Nicholls, N., Easterling, D., Goodess, C. M., Kanae, S., Kossin, J., et al. (2012). Changes in climate extremes and their impacts on the natural physical environment. In C. B. Field, et al. (Eds.), *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 (IPCC) (pp. 109–230). Cambridge, UK/New York, NY: CUP.

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

Sillmann, J., Kharin, V. V., Zwiers, F. W., Zhang, X., & Bronaugh, D. (2013). Climate extremes indices in the CMIP5 multi-model ensemble: Part 2. Future climate projections. *Journal of Geophysical Research: Atmospheres*, *118*, 2473–2493. https://doi.org/10.1002/jgrd.50188

Sippel, S., Zscheischler, J., Mahecha, M. D., Orth, R., Reichstein, M., Vogel, M., &(2017). Refining multi-model projections of temperature extremes by evaluation against land–atmosphere coupling diagnostics. *Earth System Dynamics*, *8*, 387–403. https://doi.org/10.5194/esd-8-387-2017

Trugman, A. T., Medvigy, D., Mankin, J. S., & Anderegg, W. R. L. (2018). Soil moisture stress as a major driver of carbon cycle uncertainty. *Geophysical Research Letters*, *45*, 6495–6503. https://doi.org/10.1029/2018GL078131

Ukkola, A. M., De Kauwe, M. G., Pitman, A. J., Best, M. J., Abramowitz, G., Haverd, V., et al. (2016). Land surface models systematically overestimate the intensity, duration and magnitude of seasonal-scale evaporative droughts. *Environmental Research Letters*, *11*, 104012. https://doi.org/10.1088/1748-9326/11/10/104012

Ukkola, A. M., De Kauwe, M. G., Roderick, M. L., Abramowitz, G., & Pitman, A. J. (2020). Robust future changes in meteorological drought in CMIP6 projections despite uncertainty in precipitation. *Geophysical Research Letters*, *47*, e2020GL087820. https://doi.org/10.1029/2020GL087820

Ukkola, A. M., Pitman, A. J., Donat, M. G., De Kauwe, M. G., & Angélil, O. (2018). Evaluating the contribution of land-atmosphere coupling to heat extremes in CMIP5 models. *Geophysical Research Letters*, *45*, 9003–9012. https://doi.org/10.1029/2018GL079102

Utsumi, N., Kim, H., Kanae, S., & Oki, T. (2017). Relative contributions of weather systems to mean and extreme global precipitation. *Journal of Geophysical Research: Atmospheres*, *122*, 152–167. https://doi.org/10.1002/2016jd025222

Vicente-Serrano, S., Beguería, S., & López-Moreno, J. I. (2010). A multiscalar drought index sensitive to global warming: The standardized precipitation evapotranspiration index. *Journal of Climate*, *23*(7), 1696–1718. https://doi.org/10.1175/2009JCLI2909.1

Vogel, M. N., Hauser, M., & Seneviratne, S. I. (2020). Projected changes in hot, dry and wet extreme events' clusters in CMIP6 multi-model ensemble. *Environmental Research Letters*, *15*, 094021. https://doi.org/10.1088/1748-9326/ab90a7

Wang, S., Huang, J., He, Y., & Guan, Y. (2014). Combined effects of the Pacific Decadal Oscillation and El Niño-Southern Oscillation on global land dry–wet changes. *Scientific Reports*, *4*, 6651. https://doi.org/10.1038/srep06651

Zscheischler, J., Martius, O., Westra, S., Bevacqua, E., Raymond, C., Horton, R. M., et al. (2020). A typology of compound weather and climate events. *Nature Reviews Earth & Environment*, *1*, 333–347. https://doi.org/10.1038/s43017-020-0060-z

Zscheischler, J., Westra, S., van den Hurk, B. J. J. M., Seneviratne, S. I., Ward, P. J., Pitman, A. J., et al. (2018). Future climate risk: The challenge of compound events. *Nature Climate Change*, *8*, 469–477. https://doi.org/10.1038/s41558-018-0156-3