Calibrating climate model ensembles for assessing extremes in a changing climate

Nadja Herger[1,2], Oliver Angélil[1,2], Gab Abramowitz[1,3], Markus Donat[1,2], Dáithí Stone[4,5], Karsten Lehmann[6]

[1]Climate Change Research Centre, UNSW Sydney
[2]ARC Centre of Excellence for Climate System Science, Australia
[3]ARC Centre of Excellence for Climate Extremes, Australia
[4]Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[5]Global Climate Adaptation Partnership, Oxford, U.K.
[6]Satalia, Berlin, Germany

Key Points:

- Out-of-sample testing when introducing bias correction approaches is critical
- Biases in trends can dominate uncertainty in estimates of anthropogenic effect on extreme weather
- Calibration on distribution shapes does not guarantee improved skill of attribution statements

Corresponding author: Nadja Herger, `nadja.herger@student.unsw.edu.au`

Abstract

Climate models serve as indispensable tools to investigate the effect of anthropogenic emissions on current and future climate, including extremes. However as low dimensional approximations of the climate system, they will always exhibit biases. Several attempts have been made to correct for biases as they affect extremes prediction, predominantly focused on correcting model-simulated distribution shapes. In this study, the effectiveness of a recently published quantile-based bias correction scheme, as well as a new subset selection method introduced here, are tested out-of-sample using model-as-truth experiments. Results show that biases in the shape of distributions tend to persist through time, and therefore correcting for shape bias is useful for past and future statements characterising the probability of extremes. However, for statements characterised by a ratio of the probabilities of extremes between two periods, we find that correcting for shape bias often provides no skill improvement due to the dominating effect of bias in the long-term trend. Using a toy-model experiment, we examine the relative importance of the shape of the distribution versus its position in response to long-term changes in radiative forcing. It confirms that the relative position of the two distributions, based on the trend, is at least as important as the shape. We encourage the community to consider all model biases relevant to their metric of interest when using a bias correction procedure and to construct out-of-sample tests that mirror the intended application.

1 Introduction

Observations and climate models show an increase in the frequency and intensity of hot and wet extremes and a decrease in the frequency and intensity of cold extremes, as associated regional mean temperatures increase [*Alexander et al.*, 2006; *Seneviratne et al.*, 2012; *Hartmann et al.*, 2013; *Collins and Knutti*, 2013; *Lewis and King*, 2015]. These changes coincide with a period of rapid increase in atmospheric CO_2 concentrations as a consequence of anthropogenic industrialisation. Given the current state of rapid change, the climate science community, governments, the public, and news media have become interested in how human interference with the climate system has affected various characteristics of extreme weather. This includes current changes in occurrence probability [*Peterson et al.*, 2012, 2013; *Herring et al.*, 2014, 2015, 2016]—a field known as 'event attribution'—as well as 21st century (and beyond) projections of extremes [*Sillmann et al.*, 2013] and the impacts associated with them [*Patz et al.*, 2005]. Since the 2015 Paris Agreement, which aims to pursue efforts to limit warming to 1.5°C above pre-industrial levels, and hold the increase in the global average temperature to well below 2°C, studies comparing projections of future extremes between 1.5°C and 2°C worlds have grown in popularity [*King and Karoly*, 2017; *King et al.*, 2017; *Perkins-Kirkpatrick and Gibson*, 2017; *Lewis and King*, 2017; *Sanderson et al.*, 2017a].

For both event attribution and projections of extremes, climate model simulations are widely used as they encapsulate our understanding of how human interference might affect the climate system. Because models exhibit a range of biases [*Ehert et al.*, 2012] including their ability to reproduce the observed frequency distribution of extreme events and/or long-term trends [*Sippel et al.*, 2016; *Angélil et al.*, 2016; *Bellprat and Doblas-Reyes*, 2016], the accuracy of model-derived statements pertaining to extremes is not always clear. This has been demonstrated in sensitivity studies where attribution results can change in their sign depending on the model, observational dataset, or method used. For example the likelihood of occurrence of specific rainfall extremes can either be found to be more likely (positive attribution statement), less likely (negative statement), or hardly changed (neutral statement) as a consequence of anthropogenic emissions depending on the approach taken [*Angélil et al.*, 2017b; *Hauser et al.*, 2017]. For temperature extremes, the sign of the attribution statement may not change, but the actual attribution statement in terms of the quantification of how much anthropogenic climate change has altered the likelihood of the event, can vary by an order of magnitude [*Angélil et al.*, 2017b]. Furthermore, model-simulated extremes may be systematically biased across various models compared to observations/reanalyses

[*Christensen et al.*, 2008; *Wang et al.*, 2014; *Angélil et al.*, 2016; *Donat et al.*, 2017; *Bellprat and Doblas-Reyes*, 2016], and therefore taking the median or mean of the metric of interest across ensemble members can be unreliable [*King and Karoly*, 2017; *King et al.*, 2017; *Perkins-Kirkpatrick and Gibson*, 2017; *Lewis and King*, 2017]. Such biases are not necessarily reduced after the poorest performing models have been removed from an ensemble, indeed this process can reinforce model biases if metrics are not carefully chosen, since the best performing models might have common biases due to shared model development history (so-called model interdependence) [*Herger et al.*, 2017].

One way to mitigate some of these issues is to constrain the regional changes in frequency and intensity of hot temperature extremes by the shape of the model's present-day temperature distribution [*Borodina et al.*, 2017]. Other studies have developed statistical bias correction schemes, the vast majority focusing on correcting for distribution shapes when they are not representative of the distribution shapes of observational data. Many of these studies involve a procedure in which a 'transfer function' is derived by matching percentiles between simulated and observed cumulative distribution functions (sometimes also referred to as 'quantile mapping' or 'histogram equalisation') and have been expanded on and refined in the last decade [*Piani et al.*, 2010b; *Li et al.*, 2010; *Piani et al.*, 2010a; *Hempel et al.*, 2013; *Sippel et al.*, 2016; *Jeon et al.*, 2016]. The aim of such methods is to also improve 'out-of-sample' results (a term used throughout this paper to describe time periods which have not been used to apply bias corrections and will be used to test their effectiveness).

A fundamental issue with most of these bias correction techniques is that they are often applied and tested on the same data ('in-sample'), but not in the period of their intended application (for example, because no observational data exist in the later 21st century). There is the risk that while the correction works perfectly in-sample (where observations are available), it may actually degrade predictability out-of-sample. This may be because not all relevant model biases for the metric of interest were considered in the calibration. In statistics, the equivalent might be that when we see success at interpolation, it by no means guarantees success at extrapolation.

A solution is out-of-sample testing using long observational records or model-as-truth experiments, which are common in some areas of climate science [*Abramowitz and Bishop*, 2015; *Sanderson et al.*, 2017b; *Knutti et al.*, 2017; *Herger et al.*, 2017] but appear to be sparse in others such as in the extremes community where they are critically needed. In this study we test one quantile-based bias correction method [*Jeon et al.*, 2016] (hereinafter referred to as 'the Jeon method'). Their bias correction was applied to the standard event attribution method, which utilises two model-simulated distributions of weather, each forced under a different climate scenario: a counter-factual 'natural' world without industrialisation (commonly termed 'NAT') and the 'real world' forced with all known natural and anthropogenic boundary conditions (commonly termed 'ALL' or 'RW'). Of the bias correction methods already mentioned [*Piani et al.*, 2010b; *Li et al.*, 2010; *Piani et al.*, 2010a; *Hempel et al.*, 2013; *Sippel et al.*, 2016], the Jeon method is the most simple. It adjusts the event magnitude which is being attributed, by ensuring its percentile (relative to the simulated distribution) equals the percentile of the observed event (relative to the observed distribution). For example if the simulated tail is longer than the observed tail (as is the case in their study), the observed event magnitude is shifted further out into the tail until the two percentiles (each relative to their own distributions) are equal. However, such a correction, although perfect in-sample by definition, may not reduce biases out-of-sample which also depends on the probability of extremes in a world with different forcings. We test for this possibility below.

Apart from testing the out-of-sample skill of the Jeon method, we also detail a new method to correct for biased model distribution shapes in multi-model ensembles. The technique selects the subset of climate simulations from a multi-model ensemble that reduces distribution biases (when compared to a model-as-truth), following the flexible approach introduced in *Herger et al.* [2017]. Here, a modelled distribution is obtained by pooling data from a collection of climate models. Similarly to previous methods [*Hempel et al.*, 2013; *Sippel et al.*, 2016], it corrects for the entire distribution shape, allowing it to be used for any distribution-based problem of interest, rather than just exceedance probabilities (which the Jeon method is limited to). The two methods (Jeon and the subset selection approach introduced here) can also be used in combination, providing a third bias correction option. Using long model runs (1870–2100), we test and compare the effectiveness of these three approaches for assessing the probability of extremes in a changing climate, relative to a baseline where no correction is performed. We then compare the relative influence of tail bias on attribution statements versus another relevant source of uncertainty—the bias of response to changes in long-term radiative forcing. Finally we discuss what type of bias correction and model evaluation strategies should be prioritised to determine whether models are fit for purpose in assessing extremes in a changing climate.

2 Data

We use one Coupled Model Intercomparison Project Phase 5 (CMIP5) [*Taylor et al.*, 2012] simulation per modelling institute (21 simulations). The simulations cover the 1870–2100 period (RCP8.5 after 2005) and can be found in Table S1 in the Supplementary Information (SI). We split the 231 years into seven 33-year periods to explore out-of-sample testing. The seven Time Periods (TPs) are hereinafter referred to as TP1 (1870–1902), TP2 (1903–1935), TP3 (1936–1968), TP4 (1969–2001), TP5 (2002–2034), TP6 (2035–2067), and TP7 (2068–2100).

One model per institute is chosen from the CMIP5 archive in order to reduce model interdependency. Reducing model interdependency is an important step before performing model-as-truth experiments (see e.g., *Abramowitz and Bishop* [2015] and *Sanderson et al.* [2017b]) as it helps avoid artificial skill improvements due to the 'truth' model being too similar to the remaining model simulations (increasing the risk of over-fitting). Choosing one model per institute removes multiple initial condition members of the same model as well as similar, or similarly calibrated models. By doing this the average model-to-model distances are expected to become more similar to the average model-to-observation distances [*Herger et al.*, 2017]. Indeed Figure S1a shows that for surface air temperature, the average KS test statistic between these 21 simulations and the land-only gridded observational product CRU-TS, v4.00 [*Harris et al.*, 2014] is generally smaller than the mean model-model KS value. Results for total precipitation (Figure S1b) are similar, with model-obs KS values varying slightly more within the spread of model-model KS values across regions.

Distributions of monthly mean surface air temperature (tas) and total precipitation (pr) are analysed over 58 WRAF2-v3.0 regions (see Figure 1). The regions are on average $2{\cdot}10^6$ km^2 in size. We apply the WRAF masks to the model data and calculate area-weighted monthly spatial averages over each region, covering the 231-year period. Note, the analyses could equally be performed on daily data, however this would reduce the model pool size. This work also primarily serves as a proof of concept and we thus decided against higher temporal resolution.

No observational products were used in this study, except for in Figure S1. Instead, each model is removed from the ensemble and used as if it were observations, commonly referred to as either model-as-truth experiment or perfect model setup (see section 3.1). With this, we avoid the problem with long observational records having inconsistent quality through time as a consequence of varying station density [*Macias-Fauria et al.*, 2014], yet are still able to test the fidelity of the bias correction approaches.

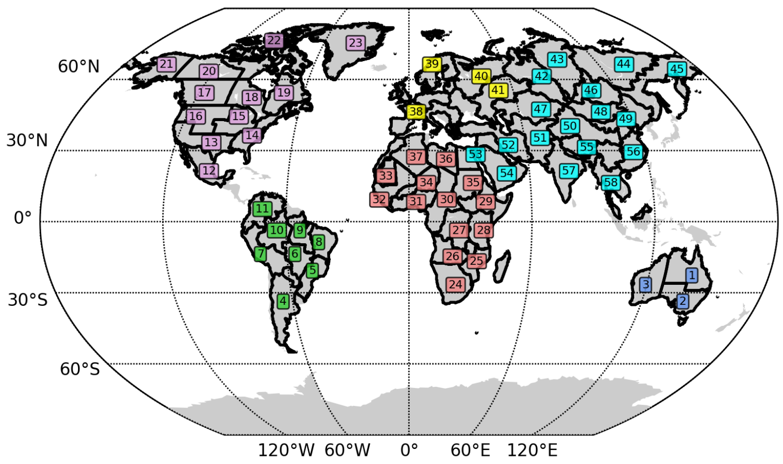

Figure 1. This map shows the 58 WRAF2-v3.0 regions used in this study. Each region is roughly $2{\cdot}10^6$ km^2 on average. The regions are colour-coded according to their continents.

3 Methods

In this study we define extreme events as the 1-in-1-year and 1-in-5-year return value based on monthly temperature and precipitation data. Even though extremes are often analysed on a daily time scale, the concept itself can be well demonstrated using 1-in-1-year and 1-in-5-year thresholds using monthly averages as done here. Furthermore, the sensitivity of extremes metrics such as the Probability Ratio (PR; looked at in this study and discussed later) to the temporal scales of the events (daily, 5-day, and monthly) have already been documented [*Angélil et al.*, 2017a]. The 1-in-1-year return value is the 91.67 percentile for warm and wet months, and the 8.33 percentile for cold months from the distribution of 33 years (12 x 33 = 396 points) in the middle time period (TP4). Note, that this is roughly (but not exactly) the climatology of the locally warmest/coldest/wettest month in the year. The 1-in-5 year return value is the 98.33 percentile for warm and wet months, and the 1.67 percentile for cold months. Given that results for 1-in-1-year events are 'cleaner' than those for 1-in-5-year events (for the latter, exceedance probabilities of zero were frequent enough to render results indistinguishable between some TPs) and since key findings are similar between both, results for 1-in-5-year extremes are shown in the SI. Results for 1-in-1-year and 1-in-5-year wet months are also only shown in the SI.

3.1 Models-as-truth experiment

Model-as-truth experiments as conducted in this study involve removing one of the ensemble members and treating it as if it were observations, or 'truth'. The remaining ensemble is then calibrated (using either the Jeon method or the subset selection method introduced in section 3.3) to try to better estimate the truth member, using data from the middle TP (TP4). The calibrated ensemble can then be tested out-of-sample in the remaining six TPs against the 'truth' member. The ability of each technique to offer an improvement over the default ensemble (the 20 remaining ensemble members) is then assessed. The process is repeated with each of the 21 models playing the role as 'truth', and results aggregated to provide an uncertainty estimate of the ability of each bias correction approach. Ensuring that model-model distances are at least that of model-observation distances (as explained in Section 2 above) gives us some confidence that success in model-as-truth experiments should translate to effective application of these techniques when adjusting climate projections.

3.2 Jeon method

As briefly mentioned in the introduction, the 'Jeon method' [*Jeon et al.*, 2016] accounts for the discrepancy between the probabilities of extreme weather events derived from the 'truth' and the model dataset by mapping the 'truth' quantile to the modelled quantile. We then calculate temperature and precipitation thresholds in the model-as-truth and remaining 20 model datasets in TP4 (simply using the same percentile in the 'truth' and model distributions to define thresholds is the essence of the Jeon method), rotating through each of the 21 models-as-truth and for each region separately.

For a real application, we usually start with an observed event which can be described as a certain percentile of the observational record. Here, however, we start with a given percentile (e.g., 91.67 percentile for warm events or 8.33 percentile for cold events) and calculate a model-derived threshold using that percentile. Exceedance Probabilities (EPs; for warm or wet events) or Probabilities of Falling Below (PsFB; for cold events) are computed relative to this threshold. When applying the Jeon method, the threshold is obtained from the pooled model distribution rather than from the model-as-truth. For a graphical representation of the Jeon method we refer to Figure 3 in their paper.

3.3 Ensemble-based subset-selection method

In *Herger et al.* [2017], an optimal subset of model runs is chosen to minimise the Root Mean Squared Error (RMSE) of global temperature or precipitation fields between a 'truth' (either observational product or model-as-truth) and an ensemble average for a given subset size. Here, we tailor the method to extremes by finding the optimal subset of CMIP5 model runs that when pooled (i.e. not averaging but rather concatenating all the data into one long vector) minimises the two-sample Kolmogorov-Smirnov (KS; *Stephens* [1970]) test statistic compared to a given 'truth' (model-as-truth in this studx). Different to the subset selection in *Herger et al.* [2017], here we are pooling rather than averaging model runs and we are minimising the KS test statistic for temperature and rainfall distributions over regions rather than the global RMSE.

We also note that the meaning of 'optimal' is not general and can vary depending on the specific application. When we refer to an optimal subset we are talking about the subset that minimises the cost function for a specific variable, region, TP, model-as-truth, metric and so on. A globally optimal subset does not exist and would not be very meaningful.

The KS test statistic is defined as the maximum vertical distance between the 'true' Empirical Survival Functions (ESF) and the ESF of the pooled model runs. The maximum vertical distance is the same as the maximum vertical distance between two Empirical Cumulative Distribution Functions (ECDFs; ECDF=1-ESF). Examples of ESFs are shown in Figure 3. Since there can be any number of members (between 1 and 20) in the optimal subset, we use K to denote the number of pooled model runs found to minimise the KS test statistic.

We note that the Anderson Darling (AD; *Anderson and Darling* [1954]) test presents an alternative metric that is more sensitive to the tails of distributions than the KS test [*Heo et al.*, 2013]. We attempted to select a subset to minimise the AD test statistic; however the optimisation was not feasible due to computational constraints, given the more complex cost function which had to be rewritten for the mathematical solver.

A workflow of the novel methodology is shown in Figure 2, illustrated for one particular region and one model-as-truth. The same procedure is then repeated for the remaining WRAF regions and models-as-truth.

As noted above, the optimal subset is only calculated using TP4. Each implementation of the optimisation approach finds an optimal subset for a given ensemble size K, so in addition to selecting an optimal subset, we need a mechanism to choose the ensemble size best suited across different TPs. To do this, we use a cross-validation approach using the middle three 33-year TPs (TP3–TP5). We optimally select ensemble members for all ensemble sizes using one of these TPs and test the skill of these optimal ensembles on the other two periods. This process is repeated for all three TPs, and results averaged to find the best out-of-sample cross-validated optimal subset size K_{CV}—see Figure 2. We refer to the period we train on (that is, derive the optimal ensemble) as 'in-sample' and the periods we test on—periods never seen by the subset-selection algorithm—as 'out-of-sample'. The advantage of this approach is of course that we have models-as-truth both in- and out-of-sample and we can thus test the degree to which our bias correction methods degrade out-of-sample. We can also go much further out-of-sample had we just relied on long observational records. We use the term 'optimal ensembles' to denote the ensembles that are selected for a given ensemble size. 'Optimal subset' is used for the overall best (lowest KS test statistic) subset across all ensemble sizes.

Consider case 1 in Figure 2 (red rectangle), where we train on TP4 and test on TP3 and TP5. For each ensemble size between 1 (single best simulation) and 20 (all runs pooled), we find the subset of ensemble runs which when pooled minimise the KS test statistic in the in-sample period (TP4) compared to the model-as-truth—see ECDF inset Figure 2a. This is a non-trivial task as there are for example 184756 possible ensembles of size 10. Due to time-constraint issues, a 'brute-force' approach is therefore simply not possible for each model-as-truth, over each of the 58 regions, for three TPs, and two variables. Instead, we use the state-of-the-art mathematical programming solver Gurobi [*Gurobi*, 2015] to minimise the KS test statistic for a given ensemble size. Details, including a link to a simplified Python script used to do this can be found in the SI. Note that Gurobi is only ever used to obtain the optimal ensembles in the training periods. We end up with a curve similar to the schematic in Figure 2a: the KS test statistic of the optimal ensemble as a function of ensemble size.

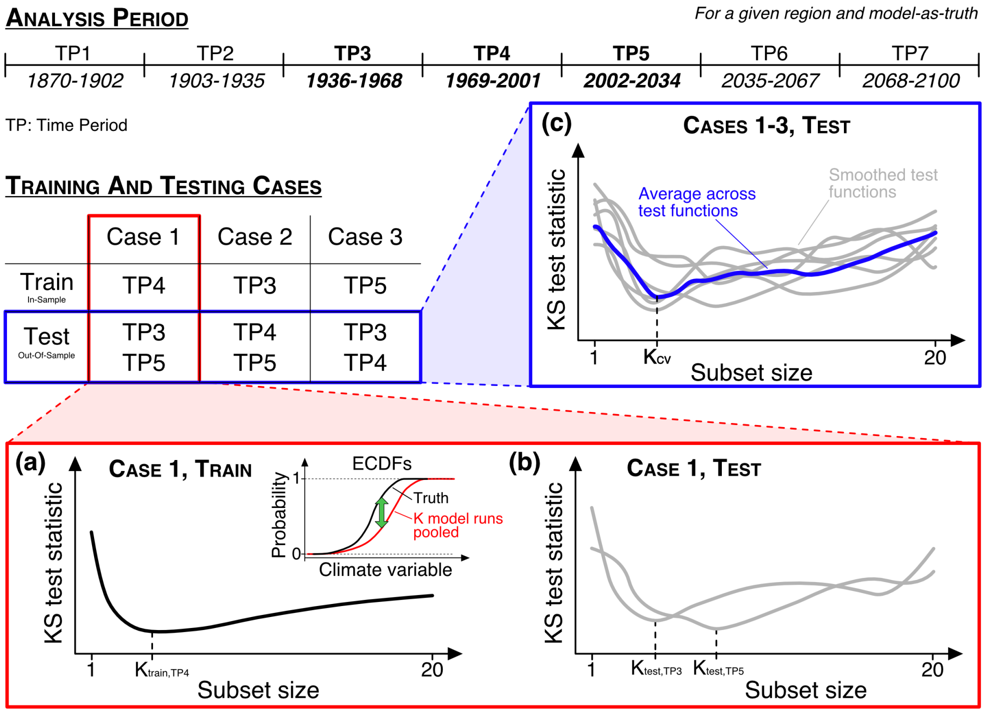

Figure 2. Methodological workflow of the study. The analysis period is split into seven 33-year periods (TP1–TP7). Only TP3–TP5 are used to obtain the cross-validated optimal subset size. **(a)** For a given model-as-truth (could equally be observations in practice), we obtain the optimal ensembles in the training set (case 1) for subset sizes 1–20. Those ensembles are then tested out-of-sample (in TP3 and TP5), see **(b)**. Performance of the optimal ensembles are tested out-of-sample in a total of six test periods (grey lines in blue box **(c)**). To account for noise generally at small ensemble sizes, these functions are smoothed using a running mean of three ensemble sizes. To obtain the cross-validated optimal subset size (K_{CV}), we average across all six smoothed test cases (blue line in **(c)**). The subset size at the minimum of this function for a particular region and model-as-truth is then used for the remainder of this study. A different size is obtained depending on the chosen region and model-as-truth.

Note that the KS test statistic can vary between 0 and 1. Here, $K_{train,TP4}$ is the number of simulations in the optimal subset for TP4.

Using only $K_{train,TP4}$ to go out-of-sample may be risky, as we do not know if the members of this optimal subset are still optimal in the two testing periods when climate forcing is different. It is possible that a different value of K would be best out-of-sample. The next step in the process is therefore to use the in-sample ensembles for each K found in (a), to calculate the KS test statistics in the two out-of-sample periods (see Figure 2b). Those KS values will likely be higher than the in-sample values. For each TP that we test on out-of-sample, we obtain a slightly different curve. Ideally we want the K with the minimum KS value for those curves ($K_{test,TP3}$ and $K_{test,TP5}$) to be close to the K with the minimum KS value found in-sample ($K_{train,TP4}$), but this is not always the case. To avoid overfitting we search for the optimal K across all three cases (termed 'cross-validation' (CV) in the literature).

We repeat the steps described above for cases 2 and 3, where the training and testing periods are changed. The curves for the six out-of-sample tests are shown in Figure 2c. Grey curves illustrate the smoothed functions using a moving window that averages the KS test statistics across three ensemble sizes. The reason we smooth those curves is because the grey lines can be very noisy at small ensemble sizes. Failure to address this might lead to overfitting in an ensemble subset size that is small.

Next, we average across the six grey curves to obtain the blue one. The cross-validated ensemble size, K_{CV}—the size used for the remainder of the study (for a given region and model-as-truth), is the subset size with the overall smallest KS test statistic across these six out-of-sample tests. We refer to it as 'cross-validated optimal subset size'. An example of in- and out-of-sample KS values for WRAF region 38, the Southern European Economic Area (EEA) and CSIRO-Mk3.6.0 r2i1p1 as the truth, can be found in the SI (Figures S3 and S4). This is an example where it is particularly important to execute the smoothing step. Without it we would end up with a small subset size, where the curves are noisy. For this region, we end up with a K_{CV} subset size larger than the in-sample optimal subset sizes. The optimal ensemble in TP4 for K_{CV} then becomes the 'CV optimal subset'. A larger ensemble size means that we are relying on a wide range of climate models rather than betting on a small subset of models to perform well out-of-sample. Note that TP3 and TP5 may now not be considered as truly out-of-sample for testing the ability of our bias correction approaches, since they are used to find the optimal cross-validated subset size K (this is why they are in boldface in Figure 2).

The pooling of model runs from the CMIP5 archive for each 33-year period mitigates the effect of internal variability (each run being in a different state of internal variability). What remains is therefore primarily the forced response, being the main difference between the TPs.

3.4 Calculation of extremes metrics

After correcting for shape bias, whether it be with the Jeon or sub-selection approach, we calculate EPs (for warm and wet events), PsFB (for cold events), and PRs—the ratio of two EPs or PsFB characterising the change in probability of the event between two periods of different forcings, in TP1–TP3 and TP5–TP7.

The PR is typically used by the event attribution community between ALL and NAT forced climates to characterise the anthropogenic contribution to the chance of an extreme, but is unconventionally used in this study between two 33-year periods within the 1870–2100 period. This allows out-of-sample testing forward and backward in time and so includes a broader range of forcing changes with which to test the bias correction techniques. The EPs, PsFB, and PRs obtained from the reference distribution of all 20 models pooled when using the truth to define the threshold (in TP4) are shown against EPs, PsFB, and PRs (again in the distribution of all 20 models pooled) obtained when using the Jeon method to calculate the threshold (light and dark green markers in Figures 5 and 6). The same procedure is also applied to the CV optimal subset (yellow and orange markers in Figures 5 and 6). The skill of both methods is gauged by comparing them to the 'true' EPs, PsFB, and PRs derived from using each model as truth.

4 Results

4.1 Obtaining the cross-validated optimal subset

Cross-validated optimal subsets for each of the 58 WRAF regions are obtained as described in Section 3.3. Here, we illustrate the ensemble-based subset-selection method in TP4 using WRAF region 38, which is the Southern EEA. ESFs and normalised histograms are shown in Figure 3. The 'truth' (CSIRO-Mk3.6.0, r2i1p1) is shown in black and the remaining 20 CMIP5 simulations in grey. The model run closest to the 'truth' in terms of the KS test statistic is shown in cyan. Note that the warm tails of most of the CMIP5 runs are too short relative to the 'truth'. This tail bias persists in other TPs (seen in Figure 5a and discussed later). The ESFs for precipitation are shown in Figure S2.

Simply pooling all 20 model runs will not solve this problem, as shown with the light green line. This is where the subset-selection comes into play. The red line is the optimal subset in the in-sample period (here: TP4), with $K = 7$. The cross-validated optimal subset is shown in yellow, with $K_{CV} = 9$. Both the red and yellow lines are closer to the observations than the green line. Note, that any subset selection approach can only be successful if the original ensemble spans the entire distribution of the 'true' conditions, as it does in this case.

The horizontal dashed lines show the 1-in-1-year warm and cold month events (91.67 and 8.33 percentiles respectively). The vertical lines refer to the corresponding thresholds of the different distributions. The thresholds for the optimal subset and cross-validated optimal subset are now positioned closer to the 'true' thresholds, which is not guaranteed in all cases since we are optimising for the shape of the entire distribution, not specifically the tails. Thresholds for the '20 runs pooled' distribution (light green) and the 'CV optimal subset' (yellow) are later used for the Jeon method.

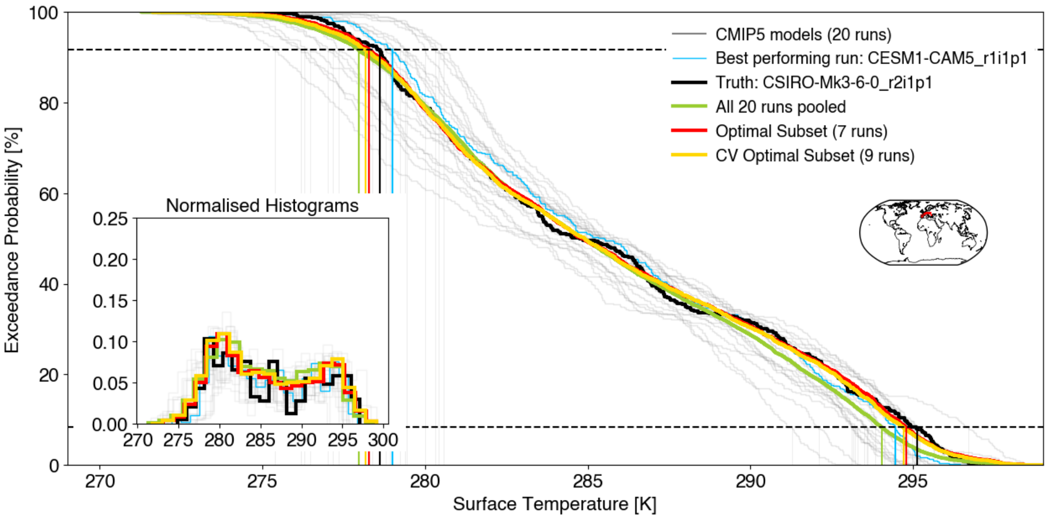

Figure 3. Empirical survival function of monthly surface temperature in period TP4 over WRAF region 38 (Southern EEA) for CSIRO-Mk3.6.0 r2i1p1 as truth. The raw (no correction for mean bias) individual CMIP5 model distributions are shown in grey, and the truth in black, each distribution consisting of 396 (33 years × 12 months) points. The cyan curve is the single best performing run (in terms of the lowest KS-test statistic compared to the model-as-truth). The green curve is the 20 CMIP5 runs pooled. The red curve is the optimal subset of CMIP5 runs which results in the lowest KS-test statistic compared to the truth derived within TP4 (happens to be $K = 7$), and the yellow curve is the optimal subset when $K = 9$, being the subset size best suited across TP3–TP5 (tuned via cross-validation). Vertical lines show the 1-in-1 year cold (8.33th percentile) and warm (91.67th percentile) thresholds derived from the various distributions.

Figure 4a confirms that the sub-selection is working in-sample (TP4) for all regions, showing the in-sample KS test statistic values based on absolute surface temperature. The marker colours are consistent with what was used in Figure 3. Region 38 is highlighted in grey as this is the region used to illustrate results in (b) and subsequent panels. The smaller the KS test statistic, the closer the corresponding distribution is to the 'truth'. There are even some regions where all the model distributions are significantly different ($p < 0.05$) from the 'true' distribution (black border around markers).

We observe that simply pooling all 20 available model runs (green marker) already seems to bring the distribution closer to the 'truth'. It is usually better than most individual model runs. However, choosing ensemble members optimally can improve our pooled distribution even further. As before, the red marker is the subset which is optimal in-sample (here: TP4) and the yellow marker is the optimal subset in TP4 for size K chosen across TP3–TP5. Results for precipitation are similar (Figure S5a).

The CV optimal subset size is usually larger than the in-sample subset size (not shown). This tendency towards larger ensemble sizes is consistent with findings by *Reifen and Toumi* [2009] who suggest that having a 'portfolio' of climate models is better than relying on a small subset when making predictions as there is a risk associated with small ensemble sizes.

In Figure 4b, which shows results only for WRAF region 38, we test whether the subset selection improves skill, measured as the KS test statistic, in the remaining six TPs. Here, each model is used as the 'truth', so there are 21 points in each of the boxplots. By definition, the bias correction improves skill in-sample (TP4) relative to the case where no correction is performed (all runs pooled). We note that it also improves skill out-of-sample as far as TP1 and TP7 (biases in the shape tend to persist), although the skill gradually diminishes (yellow and red boxplots form a V-shape) the further away in time (and forcing) we move from the training period. Results in this format for the other 57 regions are similar (not shown here), as well as for precipitation (Figure S5b). Given that skill of the optimal subset and CV optimal subset are fairly similar, we only show results using the CV optimal subset in the remainder of the study.

4.2 Application of bias correction to extremes

Now that we have confirmed that the ensemble-based subset-selection successfully improves the shape of the distribution in- and out-of-sample, we can focus our attention on extreme events. We start with EPs and PsFB (section 4.2.1) for warm and cold events respectively before we test its skill on PRs (section 4.2.2). For extremes, we are of course only interested in the tails of the distribution even though we calibrated the whole distribution to be similar to the 'truth'. However, calibrating on the whole distribution still makes sense as we are not fixing usage to a particular extreme and can thus explore a range of thresholds for extremes in a consistent way. Moreover, we ensure that the mean climate (i.e., the bulk of the distribution) is right, and avoid an unrealistically truncated distribution (by e.g. solely optimising the tail of the distribution).

4.2.1 Probabilities of exceeding or falling below a threshold

Calibrating on the shape of the distribution in-sample does not guarantee that we subsequently get better estimates of PsFB or EPs. This is an assumption of *metric transitivity*, meaning that we expect an improvement in one metric—the shape of the distribution, to increase skill of another metric—EPs or PsFB—as though they were dependent. If this were not the case, testing the metric out-of-sample on anything other then what it was calibrated on in-sample would likely give poor results. In this section we test if metric transitivity holds for temperature extremes. Results for wet events can be found in the SI.

Panels 5a and b show the probabilities of exceeding the 91.67 percentile in TP4 (1-in-1-year warm events; left column) or falling below the 8.33 percentile in TP4 (1-in-1-year cold events; right column) over Southern EEA using CSIRO-Mk3.6.0 as the 'truth'. Results for 1-in-5-year warm and cold month events are shown in Figure S6. We see that the probability of warm events decreases towards earlier TPs and increases towards later TPs (vise versa for cold events). We do not see such clear changes in EPs for precipitation (Figure S7a for 1-in-1-year events and Figure S8a for 1-in-5-year wet month events). For warm events, the increase in EPs towards TP7 is significantly larger than the decrease in EPs towards TP1, indicating the stronger change in forcing towards the end of the 21st century. There are two additional markers compared to Figure 4. Dark-green markers refer to the case when all 20 runs are pooled and the threshold was based on this pooled distribution in TP4 (Jeon method) rather than the truth distribution. Orange markers refer to the CV optimal subset with threshold derived from this subset itself in TP4 (again Jeon method) rather than the truth. The closer the coloured markers are to the truth (black marker with horizontal line) outside of TP4, the more skillful the given bias correction procedure. Both the Jeon and subset selection methods appear to improve EPs and PsFB relative to when no correction is performed (light green marker).

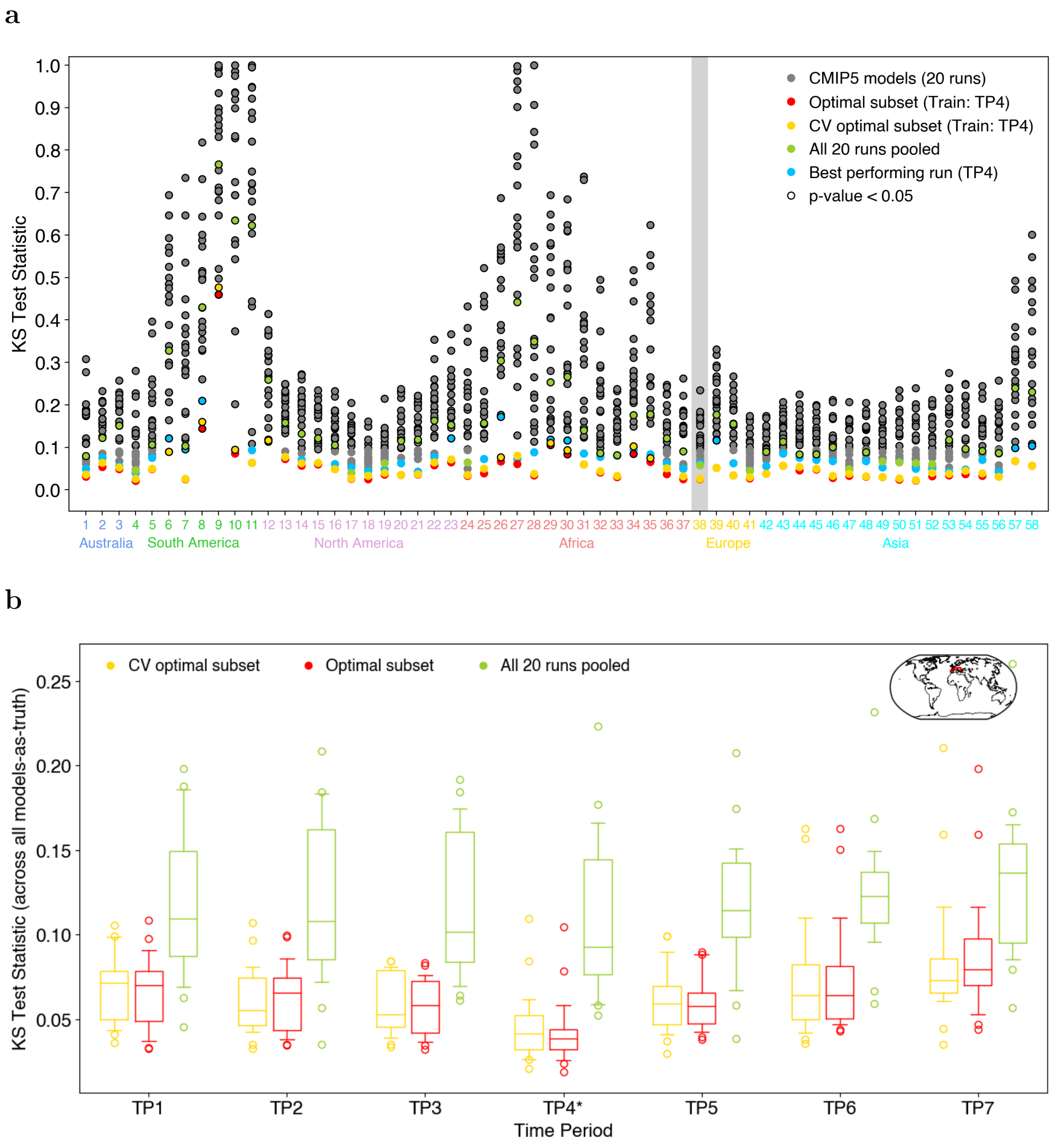

Figure 4. **(a)** The in-sample KS (TP4) test statistics for all WRAF regions are shown based on CSIRO-Mk3.6.0 r2i1p1 as truth. TP4 is used as our training period, and the KS-test statistics (compared to the model-as-truth) of the individual and pooled runs are shown within the same period. We show results of absolute surface temperature from the individual CMIP5 simulations (grey), the single best run (cyan), all 20 runs pooled (green), the optimal subset (red), and the cross-validated optimal subset (yellow). Markers have a black border if the corresponding distribution is significantly different ($p < 0.05$) from the distribution of the 'truth'. WRAF region 38 (Southern EEA), is highlighted in grey. **(b)** Results for WRAF region 38 are aggregated across all models-as-truth and for the seven time periods. In all cases, the subset is obtained in TP4 and applied to the other time periods. Boxplots for the optimal subset (red), CV optimal subset (yellow) and all 20 runs pooled (green) are shown. For the boxplots, the centerline is the median, the box spans the 25th–75th percentile range, and the whiskers span the 10th–90th percentile range.

Panels (c) and (d) show the absolute error between each of the coloured markers and the 'truth', still over Southern EEA, using each model-as-truth, allowing us to present a range of skill. By definition, the absolute error for the methods based on the Jeon method are zero in the in-sample period (TP4). Again, we observe that both methods improve EPs and PsFB as far from the training period as TP1 and TP7. Both methods also improve skill in the EP for precipitation events going back to TP1 and forwards to TP7 (Figures S7b and S8b). The significant reduction in the size of the absolute error in panel (d) towards the end of the 21st century is due to the reduction in the probabilities of cold extremes in a rapidly warming climate.

Panels (e) and (f) show results averaged within the six continents: absolute errors for each model-as-truth are averaged across all WRAF regions that fall within a given continent. We summarise results by only showing results for TP1, TP4 (in-sample), and TP7 for a given continent. As for the Southern EEA, the bias correction strategies generally improve skill out-of-sample. The exception being the Jeon method in TP7 over South America and Africa for warm events (panel (e)), where the absolute error of the dark green marker is higher than for the light green marker. Applying the Jeon method on top of optimally selecting ensemble members usually leads to marginal improvements in skill beyond only optimally selecting ensembles members. Similar conclusions can be made for precipitation (Figures S7c and S8c) and 1-in-5-year temperature events (Figure S6c).

So, calibrating on the shape of the distribution leads to improved EPs and PsFB, even when training and testing periods are several decades apart. These findings are consistent with a study by *Borodina et al.* [2017], who found a strong correlation between the modelled present-day temperature distribution and the projected frequency of warm extremes (defined as future exceedance of today's 95th percentile), which they then use to constrain changes in the intensity of warm extremes in various regions. The reason the Jeon method and our subset selection method are successful (relative to no bias correction) is because shape bias tends to persist through time, as already mentioned, and EPs are strongly influenced by the shapes of the tails which can be strikingly biased in many cases. Although EPs improve substantially with the bias correction methods, they are still imperfect, one reason likely being another model bias which is discussed next.

4.2.2 Probability ratios

In the event attribution community, it is not the EP or PFB but rather the PR that is of interest, being the ratio of two EPs or PsFB, typically between NAT and ALL forced simulations. The NAT scenario would refer to the same TP but under a forcing scenario representative of a world without anthropogenic influences. However, here the PR is calculated by dividing the EP or PFB in each TP, by the EP or PFB in TP4 (PRx = EP(TPx)/EP(TP4)). In Figure 6 (same as Figure 5 but for PRs) we test the effectiveness of the different bias correction strategies on the PR by comparing the ratio of two EPs (warm events; left column) or PsFB (cold events; right column) in the bias corrected distributions against the 'true' PR. Results for 1-in-5-year warm and cold month events are shown in Figure S9.

Panel (a), over Southern EEA using CSIRO-Mk3.6.0 as the 'truth', indicates that the EP in TP1–TP3 is lower than in TP4 ($PR < 1$; log2(PR)<0); and the EP in TP5–TP7 is higher than in TP4 (log2(PR)>0), when defining the threshold in TP4. The bias correction strategies appear to help as we move towards TP7: dark green, yellow, and orange markers lie closer to the black marker than the light green marker does. However, the bias correction methods do not appear to help going back to TP1, which can be considered most similar to what would be done in event attribution. In panel (b), we see that cold events in TP1–TP3 are more common than in TP4 (log2(PR)>0) and cold events in TP5–TP7 are much less common than in TP4 (log2(PR)<0). It appears (going back to TP1 or forwards to TP7) that the bias correction strategies hardly help.

Panels (c) and (d) provide more complete results for Southern EEA, as they show the spread when using each model-as-truth (each boxplot consisting of 21 points). Arrow-up markers in (d) indicate PRs of infinity as cold events defined in TP4 never occur in TP7 where the forcing conditions are very different. Again, we see that it is only for warm events going into the future that the Jeon and subset selection methods help, which is even more apparent for 1-in-5-year warm events (Figure S9c). The reason for this is most likely because

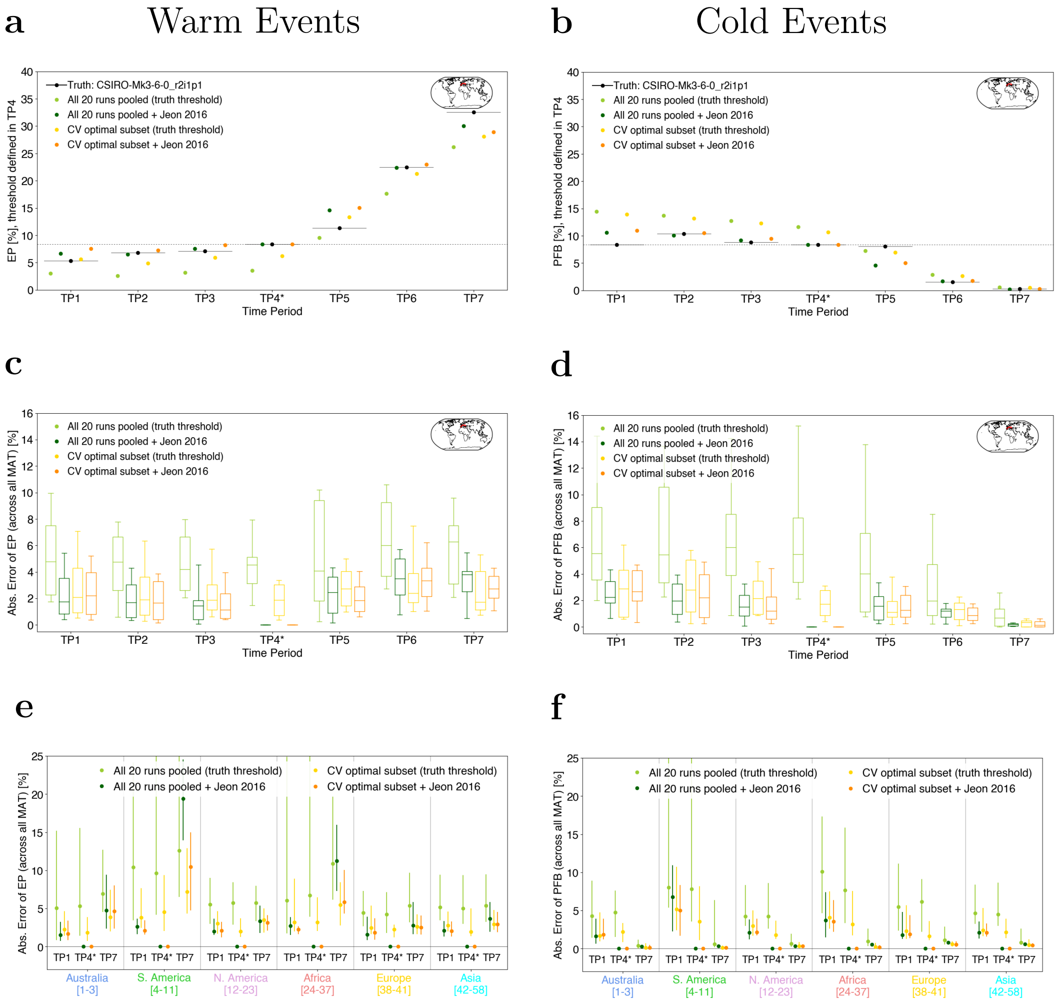

Figure 5. EPs for 1-in-1-year warm month thresholds are shown in the left column and PsFB for 1-in-1-year cold month thresholds are shown in the right column. **(a)** EPs for CSIRO-Mk3.6.0 r2i1p1 as truth and WRAF region 38 (Southern EEA) are shown for TP1–TP7. The threshold is defined in TP4 and its EP is plotted for the remaining time periods. EPs of the truth (black dot and line) are compared to the distribution of all 20 runs pooled without (light green dot) and with applying the Jeon method (dark green); and the cross-validated optimal subset without (yellow) and with applying the Jeon method (orange). **(b)** is the same as (a) but for cold events. **(c)** For the same WRAF region 38, we aggregate absolute errors of EP across all models-as-truth. The errors are obtained by calculating the absolute distances between the truth and the remaining ensembles. For the boxplots, the centerline is the median, the box spans the 25th–75th percentile range, and the whiskers span the 10th–90th percentile range. **(d)** is the same as (c) but for cold events. **(e)** aggregates the results shown in (c) across six continents by averaging results within those continents. Absolute errors of EP in TP1, TP4 and TP7 are shown. The lines span from the 10th to the 90th percentile and the dot indicates the median. **(f)** is the same as (e) but for cold events.

warm events are very well-sampled as we move towards TP7; far more than cold events going towards TP7 or warm events going back to TP1 (both of which decrease in likelihood). Even though cold events going back to TP1 increase in likelihood, the effect of the Jeon and subset selection methods is not as strong as for warm events as we move to TP7; the reason being that anthropogenic climate change is non-linear. Therefore, for warm events going forwards, we are essentially no longer in the tails, but rapidly moving towards the centre of the distribution, increasing the importance of the shape of the distribution, which we have optimised for. Error in the PR becomes increasingly larger for warm events as we move back to TP1 (similar to what is done in event attribution), or cold events as we move to TP7, since the events become poorly sampled. Therefore, correcting for the shape of the distribution does not appear to improve skill in the PR, allowing for the influence of another bias (long-term regional temperature response to changing CO_2 concentrations) to begin to dominate (discussed later).

Panels (e) and (f) reinforce that what we found for Southern EEA is valid over other regions too: the bias correction methods mostly improve skill in the PR for warm events as we move towards TP7. There also appears to be a noticeable improvement in the skill of the PR for cold events going back to TP1. This finding is consistent with our reasoning discussed in the previous paragraph: as for warm events going towards TP7, cold events going back to TP1 become more frequently sampled, increasing the importance of the shape of the distribution as opposed to just the poorly sampled tails. Arrow-up markers in panels (e) and (f) indicate errors of infinity. The effectiveness of the bias correction approaches on the PR for precipitation vary depending on the continent (Figure S10c for 1-in-1-year events and Figure S11c for 1-in-5-year events).

4.3 Toy model

To test if the PR is more sensitive to the trend or the shape of the distribution, we use a toy model experiment, shown in Figure 7. Using Gaussian distributions, we calculate PRs with different shapes of the distribution (figure columns: too narrow, correct, too wide) and different trends (figure rows: underestimated, correct, overestimated). Red represents the ALL world and blue the NAT world. To illustrate the idea, we use a 1-in-1-year warm month (91.67 percentile; see black dashed line in centre panel) event threshold in panel (a), and a 1-in-1-year cold month (8.33 percentile) event threshold in panel (b) for the calculation of the PR. Results for 1-in-5-year warm and cold month events are shown in the supplementary information (Figure S13).

The standard deviation (σ) and location (μ) of these distributions are derived from the same 21 CMIP5 simulations as used for the previous figures, for WRAF region 38 (Southern EEA). Standard deviations for each run are calculated based on monthly mean surface temperature data in TP4 (January averages for cold events and July averages for warm events). The regional temperature response to changing CO_2 concentrations (and thus location difference between the red and blue distributions) was derived by regressing the regional annual average surface temperature against global annual CO_2 concentrations from 1870–2001 (TP1–TP4). We then obtain estimates of 'too narrow'/'too wide' and 'underestimated trend'/'overestimated trend' by using the 5th and 95th percentiles of distributions consisting of 21 standard deviations or trends (one value per model simulation in each of the distributions). The 50th percentile was used as our target (middle panel in both panel (a) and (b)). The difference in CO_2 between a natural world (280ppm) and a recently observed world in 2015 (400ppm) is 120ppm. We therefore multiply the slope of the regression by 120 to approximate the temperature change between the NAT and ALL distributions. Note that we make the assumption that we only observe a shift in the mean and the distributions remain Gaussian. Results for a low-latitude region (region 27; the Democratic Republic of Congo) with lower internal variability are similar and are shown in the supplementary information (Figure S12).

In addition to the traditional calculation of the PR, being the probability of exceeding the event threshold in the ALL scenario divided by the probability of exceeding the event threshold in the NAT scenario (first line of text within each panel in Figure 7), we obtain PR estimates using the Jeon method (second line within each panel). The asterisk indicates which of the two PR estimates is closer to the target PR ($10^{0.77}$ for the warm extreme and $10^{-0.38}$ for the cold extreme). Correcting for tail bias, e.g. with the Jeon method, does not always lead to an improved PR estimate.

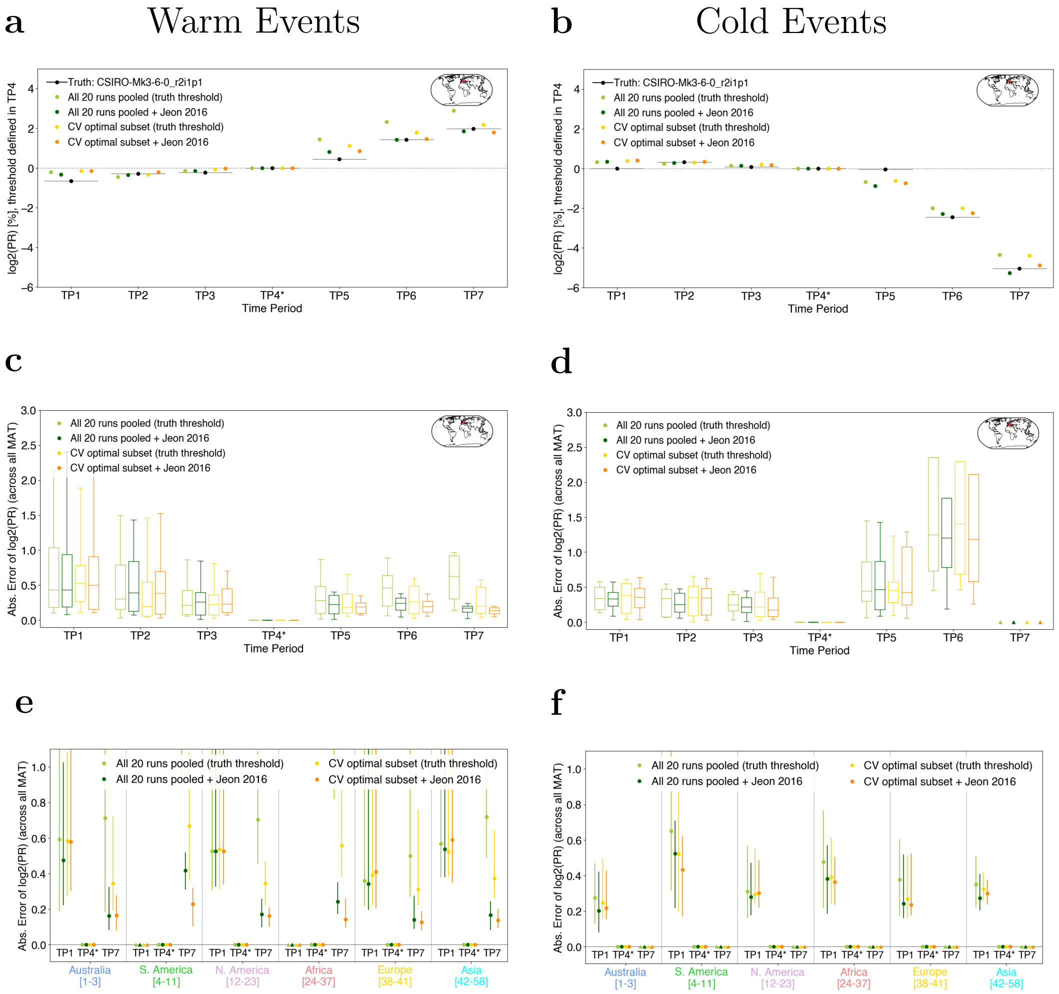

Figure 6. Probability ratios (PRs) for 1-in-1-year warm months are shown in the left column and for 1-in-1-year cold months in the right column. The threshold is defined in TP4. **(a)** log2(PRs) for CSIRO-Mk3.6.0 r2i1p1 as truth and WRAF region 38 (Southern EEA) are shown for TP1–TP7. The PR in time period x is defined as PRx = EP(TPx)/EP(TP4). PRs of the truth (black dot and line) are compared to the PRs of all 20 runs pooled without (light green dot) and with applying the Jeon method (dark green); the cross-validated optimal subset without (yellow) and with applying the Jeon method (orange). **(b)** is the same as (a) but for cold events. **(c)** For the same WRAF region 38, we aggregate absolute errors of log2(PR) across all models-as-truth. The errors are obtained by calculating the absolute distances between the truth and the remaining ensembles. For the boxplots, the centerline is the median, the box spans the 25th–75th percentile range, and the whiskers span the 10th–90th percentile range. **(d)** is the same as (c) but for cold events. **(e)** aggregates the results shown in (c) across six continents by averaging results within those continents. Absolute errors of log2(PR) in TP1, TP4 and TP7 are shown. The lines span from the 10th to the 90th percentile and the dot indicates the median. The arrow-up markers indicate that at least four out of 21 values for a given time period and continent are infinity. **(f)** is the same as (e) but for cold events.

As mentioned in Section 4.2.2, we hypothesise that the reason we hardly see an improvement in skill when calculating the PR for warm events is because the bias correction strategies only consider biases in the shapes of the distributions, without consideration of other biases such as response bias; for example, how sensitive is the regional long-term temperature to changes in global CO_2 concentrations?

In this toy model setup, the effect of response bias on the PR is roughly the same as that of shape bias for cold events. But for warm events, the effect of response bias is at least an order of magnitude larger than the effect of shape bias: given a correct trend, the PR varies from $10^{0.4}$ for 'too wide' distributions to $10^{1.17}$ for 'too narrow' distributions (factor of 6; $10^{1.17-0.4}$). However, given a correct standard deviation, the PR varies from $10^{0.25}$ for an 'underestimated trend' to $10^{2.9}$ for an 'overestimated trend' (factor of 447). For 1-in-5-year warm month extremes the PR changes by a factor of 11 when keeping the trend correct and by a factor of 1820 for a correct distribution width. So, the importance of response bias relative to shape bias increases the rarer the event.

When the standard deviation is underestimated, the sensitivity to the trend is further increased, resulting in a difference of four orders of magnitude between 'underestimated trend' and 'overestimated trend'. The toy model therefore suggests that narrower distributions exacerbate the influence of trend bias on the PR, and vice versa for overestimated standard deviations. This is because the ratio of the anthropogenic warming signal to the noise of natural variability increases or decreases as the width of the distribution decreases or increases respectively [*Angélil et al.*, 2017a].

5 Discussion and Conclusions

This study examines two bias correction approaches which account for biases in the shape of distributions of surface air temperature and total precipitation. The Jeon method artificially adjusts the threshold in the model distribution to match the percentiles in the 'true' distribution. As an approach that optimises for the whole distribution shape, we introduce a novel subset-selection method which optimally chooses ensemble members that when pooled have a distribution most similar to observations or a target 'truth' simulation. Overall results based on the Jeon method were found to be quite similar to the ensemble-based subset selection approach. This is interesting as both methods are fundamentally quite different in their underlying philosophy and technical implementation. Biases in the shape were found to persist through time based on a series of model-as-truth experiments. A subset calibrated to have a distribution shape similar to a model-as-truth in-sample was found to lead to improved out-of-sample skill when calculating EPs or PsFB, even though those probabilities are only sensitive to the tail of the distributions. This is because EPs and PsFB are strongly influenced by shape bias.

However, when calculating the PR, which is by definition the ratio of two EPs or PsFB, the bias correction methods were found to provide little to no identifiable improvement in skill (except for PRs characterising the change in probability of warm extremes into the future). When taking the fraction of two EPs or PsFB, biased tail shapes play less of a role (one can consider the tail bias present in both the numerator and denominator to cancel) and the relative importance of trend bias begins to dominate, as confirmed by the toy model experiment. It is therefore theoretically possible for a PR to be fairly close to the 'truth' even if their EPs or PsFB are not. This study explores an example where out-of-sample testing is highly beneficial and metric transitivity cannot simply be assumed. While evaluating the shapes of simulated distributions is clearly important, it is likely not the most important source of uncertainty around PR-based attribution statements and many metrics pertaining to extremes in a changing climate. Therefore, bias correction approaches that solely aim to correct for shape bias are likely to lead to only minor if not any reductions in biases in PR estimates, particularly for attribution statements pertaining to warm extremes. Note that the bias in temperature response to long-term changes in radiative forcing becomes increasingly important with increasing GHG forcing, while the shape bias is relatively static. The importance of a 'correct' distribution shape only decreases as more rare extremes are analysed, for example for 1-in-5-year events (see Figure S9 and the toy model in Figure S13 in which PRs are *even* more sensitive to the trend than the shape when compared to 1-in-1-year events).

Since evaluating response bias to long-term changes in radiative forcing using multiple observational products is not common practice in the event attribution community, we suggest that the long term response to forcing be evaluated and should be part of the optimisation process if this characteristic of the raw model output is deemed unfit for purpose (in addition to distribution properties). The difficulty here however is that the nature of the long-term temperature response in-sample does not seem to persist out-of-sample (see e.g. Figure 4 in *Herger et al.* [2017]). Note that calibrating on the PR itself will not necessarily help as the correct PR value could be obtained due to compensating errors (e.g., an appropriate combination of an overly narrow distribution and an underestimated trend). The toy model results could feed into the debate whether simulated trends should be preserved as they are (as e.g. in *Hempel et al.* [2013]) or bias corrected using observations [*Maraun*, 2016].

Future studies could test the sensitivity of results to different temporal resolutions and return periods of events. We additionally encourage the use of out-of-sample testing using long observational records and/or model-as-truth experiments to test bias correction approaches. It is critical that we identify whether there is in fact a gain in our ability to make out-of-sample predictions (i.e. does the nature of the bias being corrected persist or does it break down in the projection period? Will the bias correction performed reduce bias in the metric we are interested in, only partly, or not at all?). As we have seen here, this is not guaranteed (also see *Reichler and Kim* [2008]; *Reifen and Toumi* [2009]). Fundamentally, bias correction is a statistical calibration exercise that will work in-sample by definition. Assessing whether or not it works out-of-sample is a critical step for evaluating the nature of extremes in a changing climate.

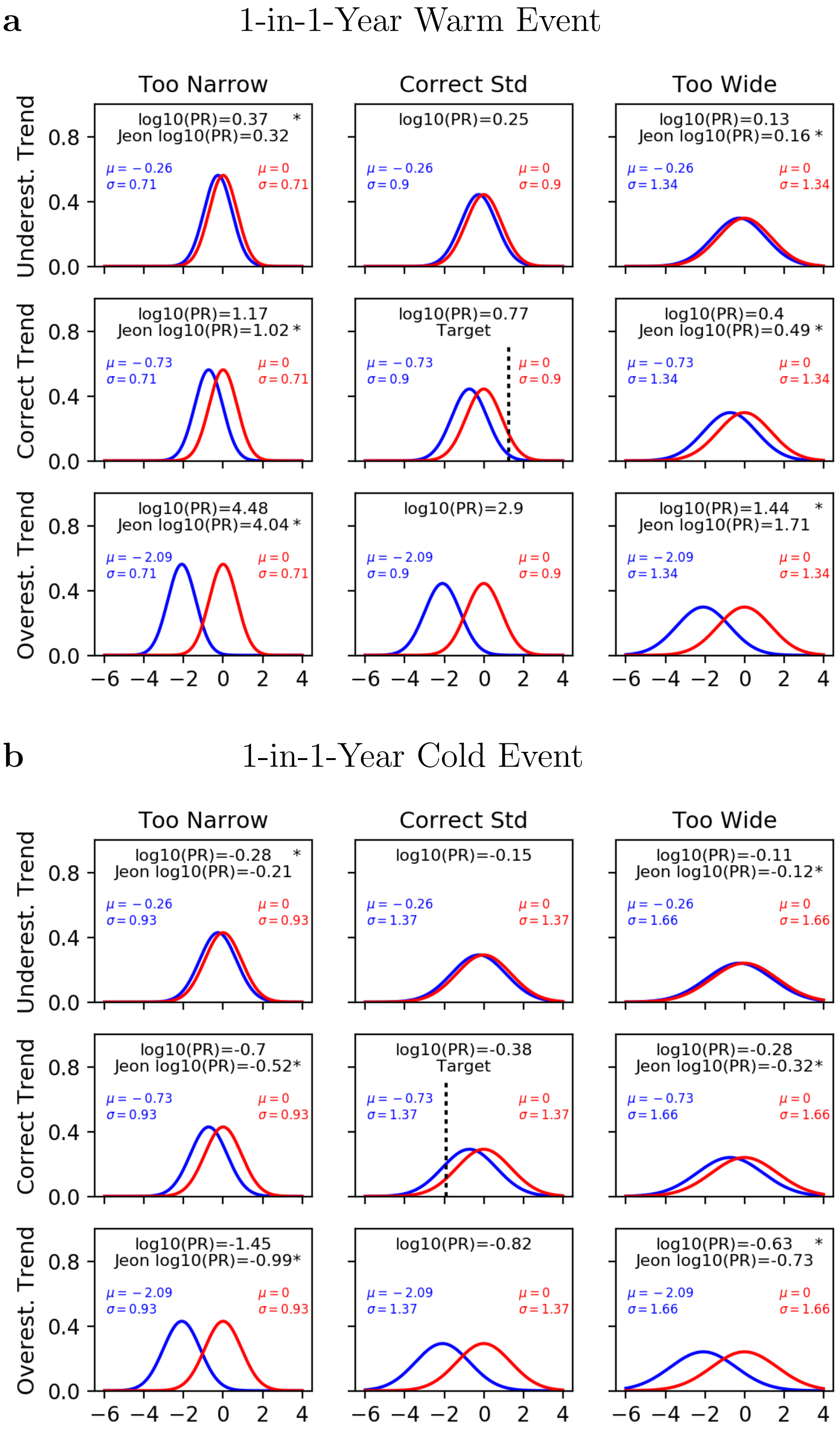

Figure 7. **(a)** Toy model experiments to demonstrate the relative importance of biases in the shape of the distribution and biases in the trend when calculating the PR. Location (μ) and shape (σ) for the Gaussian distributions were derived from 21 CMIP5 simulations for WRAF region 38 (Southern EEA). The red distributions represent the ALL forcing world and the blue distributions represent the NAT world. The 1-in-1-year warm month (91.67 percentile) of the ALL distribution in the middle panel was used as a threshold for calculating the PR. When applying the Jeon method (relevant only when distribution shapes are too narrow or too wide), the threshold is defined from each 'too narrow' or 'too wide' ALL distribution. The asterisk indicates which of the two PR estimates is closer to the target PR (middle panel). **(b)** same as panel (a) but for 1-in-1-year cold month events (8.33 percentile).

Acknowledgments

We would like to thank Jan Sedláček for providing access to the next generation CMIP5 archive based at ETHZ.

Nadja Herger, Oliver Angélil and Markus Donat acknowledge the support of the Australian Research Council Centre of Excellence for Climate System Science (CE110001028) and Gab Abramowitz the Australian Research Council Centre of Excellence for Climate Extremes (CE170100023).

We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Table S1 in the SI) for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. CMIP data can be obtained from http://cmip-pcmdi.llnl.gov/cmip5/. We use Python for data analysis and visualisation.

References

Abramowitz, G., and Bishop, C. H. (2015). Climate model dependence and the ensemble dependence transformation of CMIP projections. *J. Climate*, *28*(6), 2332–2348. doi:10.1175/JCLI-D-14-00364.1.

Alexander, L. V., Zhang, X., Peterson, T. C., Caesar, J., Gleason, B., Klein Tank, A. M. G., ... and Tagipour, A. (2006). Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research: Atmospheres*, *111*(D5). doi:10.1029/2005JD006290.

Anderson, T. W., and Darling D. A. (1970). A test of goodness of fit. *Journal of the American statistical association*, *49*(268), 765–769.

Angélil, O., Perkins, S., Alexander, L., Stone, D., Donat, M., Wehner, M., Shiogama, H., Ciavarella, A., and Christidis, N. (2016). Comparing regional precipitation and temperature extremes in climate model and reanalysis products. *Weather and Climate Extremes*, 13, 35–43. doi:10.1016/j.wace.2016.07.001.

Angélil, O., Stone, D., Perkins, S., Alexander, L. V., Wehner, M., Shiogama, H., Wolski, P., Ciavarella, A., and Christidis, N. (2017a). On the nonlinearity of spatial scales in extreme weather attribution statements. *Climate Dynamics*, 1–14. doi:10.1007/s00382-017-3768-9.

Angélil, O., Stone, D., Wehner, M. F., Paciorek, C. J., Krishnan, H., and Collins, W. D. (2017b). An Independent Assessment of Anthropogenic Attribution Statements for Recent Extreme Temperature and Rainfall Events. *Journal of Climate*, *30*(1), 5–16. doi:10.1175/JCLI-D-16-0077.1.

Bellprat, O., and Doblas-Reyes, F. (2016). Attribution of extreme weather and climate events overestimated by unreliable climate simulations. *Geophysical Research Letters*, *43*(5), 2158–2164. doi:10.1002/2015GL067189.

Borodina, A., Fischer, E. M., and Knutti, R. (2017). Potential to constrain projections of hot temperature extremes. *Journal of Climate*. doi:10.1175/JCLI-D-16-0848.1.

Christensen, J. H., Boberg, F., Christensen, O. B., and Lucas-Picher, P. (2008). On the need for bias correction of regional climate change projections of temperature and precipitation. *Geophysical Research Letters*, *35*(20), L20709. doi:10.1029/2008GL035694.

Collins, M., and Knutti, R. (2013). Chapter 12 Long-term climate change: projections, commitments and irreversibility, Climate Change 2013: The Physical Science Basis. *Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. doi:10.1017/CBO9781107415324.024.

Donat, M. G., Pitman, A. J., and Seneviratne, S. I. (2017). Regional warming of hot extremes accelerated by surface energy fluxes. *Geophysical Research Letters*, *44*(13), 7011–7019. doi: 10.1002/2017GL073733.

Ehret, U., Zehe, E., Wulfmeyer, V., Warrach-Sagi, K., and Liebert, J. (2012). HESS Opinions "Should we apply bias correction to global and regional climate model data?". *Hydrology and Earth System Sciences*, *16*(9), 3391. doi: 10.5194/hess-16-3391-2012.

Gurobi Optimization (2015). Inc., Gurobi Optimizer Reference Manual. http://www.gurobi.com.

Harris, I., Jones, P. D., Osborn, T. J., and Lister, D. H. (2014). Updated high-resolution grids of monthly climatic observations - the CRU TS3.10 Dataset. *Int. J. Climatol.*, 34, 623–642. doi:10.1002/joc.3711.

Hartmann, D. J., Klein Tank, A. M. G., Rusticucci, M. , Alexander, L., Brönnimann, S. , Charabi, Y. A.-R., Dentener, F. J., Dlugokencky, E. J., Easterling, D. R., Kaplan, A., Soden, B. J., Thorne, P. W., Wild, M., and Zhai, P. (2013). Observations: Atmosphere and Surface, Climate Change 2013: The Physical Science Basis. *Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, 159–254. doi:10.1017/CBO9781107415324.008.

Hauser, M., Gudmundsson, L., Orth, R., Jézéquel, A., Haustein, K., Vautard, R., van Oldenborgh, G. J., Wilcox, L., and Seneviratne, S. I. (2017). Methods and model dependency of extreme event attribution: The 2015 European drought. *Earth's Future*.

doi:10.1002/2017EF000612.

Hempel, S., Frieler, K., Warszawski, L., Schewe, J., and Piontek, F. (2013). A trend-preserving bias correctionâĂŞthe ISI-MIP approach. *Earth System Dynamics*, *4*(2), 219–236. doi:10.5194/esd-4-219-2013.

Heo, J. H., Shin, H., Nam, W., Om, J., and Jeong, C. (2013). Approximation of modified AndersonâĂŞDarling test statistics for extreme value distributions with unknown shape parameter. *Journal of hydrology*, 499, 41–49. doi:10.1016/j.jhydrol.2013.06.008.

Herger, N., Abramowitz, G., Knutti, R., Angélil, O., Lehmann, K., and Sanderson, B. M. (2017). Selecting a climate model subset to optimise key ensemble properties. *Earth Syst. Dynam. Discuss.*, doi:10.5194/esd-2017-28, in review.

Herring, S. C., Hoerling, M. P., Peterson, T. C., and Stott, P. A. (2014). Explaining Extreme Events of 2013 from a Climate Perspective. *Bulletin of the American Meteorological Society*, *95*(9), S1–S96.

Herring, S. C., Hoerling, M. P., Kossin, J. P., Peterson, T. C., and Stott, P. A. (2015), Extreme Events of 2014. *Bulletin of the American Meteorological Society*, *96*(12).

Herring, S. C., Hoell, A., Hoerling, M. P., Kossin, J. P., Schreck, C. J., and Stott, P. A. (2016). Explaining Extreme Events of 2015 from a Climate Perspective. *Bulletin of the American Meteorological Society*, *97*(12).

Jeon, S., Paciorek, C. J., and Wehner, M. F. (2016). Quantile-based bias correction and uncertainty quantification of extreme event attribution statements. *Weather and Climate Extremes*, 12, 24–32. doi:10.1016/j.wace.2016.02.001.

King, A. D., and Karoly, D. (2017). Climate extremes in Europe at 1.5 and 2 degrees of global warming. *Environmental Research Letters*. doi: 10.1088/1748-9326/aa8e2c.

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

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E., and Eyring, V. (2017). A climate model projection weighting scheme accounting for performance and interdependence. *Geophys. Res. Lett.*, 44. doi:10.1002/2016GL072012.

Lewis, S. C., and King, A. D. (2015). Dramatically increased rate of observed hot record breaking in recent Australian temperatures. *Geophysical Research Letters*, *42*(18), 7776–7784. doi:10.1002/2015GL065793.

Lewis, S. C., and King, A. D. (2017). Evolution of mean, variance and extremes in 21st century temperatures. *Weather and Climate Extremes*, 15, 1–10. doi:0.1016/j.wace.2016.11.002.

Li, H., Sheffield, J., and Wood, E. F. (2010). Bias correction of monthly precipitation and temperature fields from Intergovernmental Panel on Climate Change AR4 models using equidistant quantile matching. *Journal of Geophysical Research: Atmospheres*, *115*(D10). doi:10.1029/2009JD012882.

Lott, F. C., and Stott, P. A. (2016). Evaluating simulated fraction of attributable risk using climate observations. *Journal of Climate*, *29*(12), 4565–4575. doi:10.1175/JCLI-D-15-0566.1.

Macias-Fauria, M., Seddon, A. W., Benz, D., Long, P. R. and Willis, K., (2014). Spatiotemporal patterns of warming. *Nature Climate Change*, *4*(10), 845–846. doi:10.1038/nclimate2372.

Maraun, D. (2016). Bias correcting climate change simulations-a critical review. *Current Climate Change Reports*, *2*(4), 211–220. doi:10.1007/s40641-016-0050-x.

Pall, P., Aina, T., Stone, D. A., Stott, P. A., Nozawa, T., Hilberts, A. G. J., Lohmann, D. , and Allen, M. R. (2011). Anthropogenic greenhouse gas contribution to flood risk in England and Wales in autumn 2000. *Nature*, *470*(7334), 382–385. doi:10.1038/nature09762.

Patz, J. A., Campbell-Lendrum, D., Holloway, T., and Foley, J. A. (2005). Impact of regional climate change on human health. *Nature*, *438*(7066), 310. doi:10.1038/nature04188.

Perkins-Kirkpatrick, S. E., and Gibson, P. B. (2017). Changes in regional heatwave characteristics as a function of increasing global temperature. *Scientific Reports*, *7*(1), 12256.

doi:10.1038/s41598-017-12520-2.

Peterson, T. C., Stott, P. A., and Herring, S. C. (2012). Explaining Extreme Events of 2011 from a Climate Perspective. *Bulletin of the American Meteorological Society*, *93*(7), 1041–1067. doi:10.1175/BAMS-D-12-00021.1.

Peterson, T. C., Hoerling, M. P., Stott, P. A., and Herring, S. C. (2013). Explaining Extreme Events of 2012 from a Climate Perspective. *Bulletin of the American Meteorological Society*, *94*(9), 1–74.

Piani, C., Weedon, G. P., Best, M., Gomes, S. M., Viterbo, P., Hagemann, S., and Haerter, J. O. (2010). Statistical bias correction of global simulated daily precipitation and temperature for the application of hydrological models. *Journal of Hydrology*, *395*(3), 199–215. doi:10.1016/j.jhydrol.2010.10.024.

Piani, C., Haerter, J. O., and Coppola, E. (2010). Statistical bias correction for daily precipitation in regional climate models over Europe. *Theoretical and Applied Climatology*, *99*(1–2), 187–192. doi:10.1007/s00704-009-0134-9.

Reichler, T., and Kim, J. (2008). How well do coupled models simulate today's climate? *Bulletin of the American Meteorological Society*, *89*(3), 303–311. doi:10.1175/BAMS-89-3-303.

Reifen, C., and Toumi, R. (2009). Climate projections: Past performance no guarantee of future skill? *Geophysical Research Letters*, *36*(13). doi:10.1029/2009GL038082.

Rohde, R., Muller, R. A., Jacobsen, R., Muller, E., Perlmutter, S., Rosenfeld, A., Wurtele, J., Groom, D., and Wickham, C. (2013). A new estimate of the average Earth surface land temperature spanning 1753 to 2011. *Geoinformatics and Geostatistics: An Overview*, *1*(1), 1–7. doi: doi:10.4172/2327-4581.1000101.

Sanderson, B. M., Xu, Y., Tebaldi, C., Wehner, M., O'Neill, B., Jahn, A., Pendergrass, A. G., Lehner, F., Strand, W. G., Lin, L., Knutti, R., and Lamarque, J. F. (2017a). Community climate simulations to assess avoided impacts in 1.5 and 2° C futures. *Earth System Dynamics*, *8*(3), 827. doi:10.5194/esd-8-827-2017.

Sanderson, B. M., Wehner, M., and Knutti, R. (2017b). Skill and independence weighting for multi-model assessments. *Geosci. Model Dev. Discuss.*. doi:10.5194/gmd-10-2379-2017.

Seneviratne, S., Nicholls, N., Easterling, D. R., Goodess, C., Kanae, S., Kossin, J., Luo, Y., Marengo, J., McInnes, K., Rahimi, M., Reichstein, M., Sorteberg, A., Vera, C., and Zhang, X. (2012). Changes in climate extremes and their impacts on the natural physical environment, Managing the Risk of Extreme Events and Disasters to Advance Climate Change Adaptation. *A Special Report of Working Groups I and II of the IPCC, Annex II*, 109–230.

Sillmann, J., Kharin, V. V., Zwiers, F. W., Zhang, X., and Bronaugh, D. (2013). Climate extremes indices in the CMIP5 multimodel ensemble: Part 2. Future climate projections. *Journal of Geophysical Research: Atmospheres*, *118*(6), 2473–2493. doi:10.1002/jgrd.50188.

Stephens, M. A. (1970). Use of the Kolmogorov-Smirnov, Cramér-Von Mises and related statistics without extensive tables. *Journal of the Royal Statistical Society*, 115–122.

Sippel, S., Otto, F. E. L., Forkel, M., Allen, M. R., Guillod, B. P., Heimann, M., Reichstein, M., Seneviratne, S. I., Thonicke, K., and Mahecha, M. D. (2016). A novel bias correction methodology for climate impact simulations. *Earth Syst. Dynam.*, 7, 71–88. doi:10.5194/esd-7-71-2016.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.*, *93*(4), 485–498. doi:10.1175/BAMS-D-11-00094.1.

Wang, C., Zhang, L., Lee, S. K., Wu, L., and Mechoso, C. R. (2014). A global perspective on CMIP5 climate model biases. *Nature Climate Change*, *4*(3), 201. doi:10.1038/nclimate2118.