



# Ensemble optimisation, multiple constraints and overconfidence: a case study with future Australian precipitation change

Nadja Herger<sup>1</sup> · Gab Abramowitz<sup>2</sup> · Steven Sherwood<sup>1</sup> · Reto Knutti<sup>3</sup> · Oliver Angéllil<sup>1</sup> · Scott A. Sisson<sup>4,5,6</sup>

Received: 27 May 2018 / Accepted: 18 February 2019 / Published online: 1 April 2019  
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

## Abstract

Future climate is typically projected using multi-model ensembles, but the ensemble mean is unlikely to be optimal if models' skill at reproducing historical climate is not considered. Moreover, individual climate models are not independent. Here, we examine the interplay between the benefits of optimising an ensemble for the performance of its mean and the effect this has on ensemble spread as an uncertainty estimate. Using future Australian precipitation change as a case study, we perform optimal subset selection based on present-day precipitation, sea surface temperature and/or 500 hPa eastward wind climatologies. We use either one, two, or all three variables as predictors. Out-of-sample projection skill is assessed using a model-as-truth approach (rather than observations). For multiple variables, multi-objective optimisation is used to obtain Pareto-optimal subsets (an ensemble of model subsets), to gauge the uncertainty in optimisation arising from the multiple constraints. We find that the spread of climate model subset averages typically under-represents the true projection uncertainty (overconfidence), but that the situation can be significantly improved using mixture distributions for uncertainty estimation. The single best predictor, present-day precipitation, gives the most accurate results but is still overconfident—a consequence of calibrating too specifically. It is only when all three constraints are used that projection skill is improved and overconfidence is eliminated, but at the cost of a poorer best estimate relative to one predictor. We thus identify an important trade-off between accuracy and precision, depending on the number of predictors, which is likely relevant for any subset selection or weighting strategy.

**Keywords** Multi-objective optimisation · Pareto optimality · Constraint · Multi-model ensemble · Prediction · Model-as-truth experiments

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s00382-019-04690-8>) contains supplementary material, which is available to authorized users.

✉ Nadja Herger  
hergernadja@gmail.com

- <sup>1</sup> Climate Change Research Centre, ARC Centre of Excellence for Climate System Science, UNSW Australia, Sydney, NSW 2052, Australia
- <sup>2</sup> Climate Change Research Centre, ARC Centre of Excellence for Climate Extremes, UNSW Australia, Sydney, NSW 2052, Australia
- <sup>3</sup> Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland
- <sup>4</sup> School of Mathematics and Statistics, UNSW Australia, Sydney, NSW 2052, Australia
- <sup>5</sup> ARC Centre of Excellence for Climate Extremes, UNSW Australia, Sydney, NSW 2052, Australia
- <sup>6</sup> ARC Centre of Excellence for Mathematical and Statistical Frontiers, UNSW Australia, Sydney, NSW 2052, Australia

## 1 Introduction

Multi-model ensembles such as that from the Climate Model Intercomparison Project phase 5 (CMIP5) (Taylor et al. 2012) are an indispensable tool for projections of climate change. The CMIP5 archive is often referred to as an *ensemble of opportunity* given that its composition is determined by the ability of international modelling centers to contribute to them (Knutti et al. 2010a; Tebaldi and Knutti 2007). There is a lack of systematic sampling of the uncertainties affecting their projections, and the multi-model ensemble thus does not necessarily span what we think the true uncertainty range of projections is. Due to a lack of agreed-on alternatives, the equally-weighted multi-model mean is most often used as a best estimate if a set of climate models is available (Knutti et al. 2010a). Even though this ensemble mean usually outperforms most or all of the individual models due to the cancelling of non-systematic errors, this

*one-model-one-vote* approach has increasingly been criticised (Knutti et al. 2010b). Different research groups share ideas, literature, parameterisations, observational datasets for model evaluation, and sometimes even code of whole model components. Individual model simulations therefore do not represent truly independent estimates (Abramowitz 2010). A failure to address this issue can likely lead to overconfidence in our results (Knutti et al. 2010a), biases in the ensemble mean and variance, and potentially spurious correlations in the archive due to model replication (Caldwell et al. 2014; Sanderson et al. 2015b).

Different approaches have recently been proposed to account for these issues by either weighting or subsampling climate models (Bishop and Abramowitz 2013; Sanderson et al. 2015a, 2017; Leduc et al. 2016; Annan and Hargreaves 2017; Knutti et al. 2017; Herger et al. 2018). Most of these approaches optimise a single cost function, resulting in a single best weighted ensemble or subset. However, often multiple sets of variables or functional forms of cost functions can plausibly be useful for constraining the future climate response. This is also the case for future Australian precipitation change, which is characterised by notable model disagreement and is potentially influenced by a large range of climate processes. The CMIP5 models do not agree on the sign and magnitude of precipitation change over Australia (IPCC 2013), which makes it an interesting case study for us.

Despite being imperfect approximations of reality (Box and Draper 1987), climate models can be used to constrain future projections based on how well they agree with observations in the instrumental period. A recent paper by Langenbrunner and Neelin (2017) (hereafter referred to as LN17) applied this idea to constrain end-of-century California wet season precipitation change, which is characterised by large inter-model disagreement. Multi-objective optimisation was conducted based on the CMIP5 models' historical performance in tropical Pacific sea surface temperature, upper level zonal winds in the midlatitude Pacific, and California precipitation. Using multi-objective optimisation (optimising simultaneously on three separate cost functions), they identified a set of Pareto-optimal subsets (an ensemble of subset means), which built the basis for the constraint. The set was found to narrow the range of projected California precipitation, increasing confidence in a positive mean precipitation change.

Multi-objective optimisation is a step towards more generalised calibration. In the related emergent constraints literature, calibration is most commonly implemented based on a single cost function. The aim is to identify a well-observed metric in the current climate that correlates well with end-of-century projections of a variable of interest across different climate models to end up with more reliable projections (Boé et al. 2009; Knutti et al. 2017). However, often multiple relevant metrics should be minimised or maximised at the

same time (rather than their sum) to constrain a given target projection. In such cases, multi-objective optimisation can be used to quantify performance trade-offs between those predictors. Multiple observational constraints can be used to reduce the large disagreement across models.

Apart from identifying a decreased ensemble spread of projected California precipitation, no out-of-sample testing was conducted by LN17. In-sample skill, calculated based on the same time period in which the calibration was performed, does not guarantee out-of-sample skill for projected climate change. One way to account for this is the use of model-as-truth experiments (Sanderson et al. 2017; Knutti et al. 2017; Abramowitz and Bishop 2015). Such experiments give us pseudo-observations in the historical period and for the projections in form of a model considered as the "truth". This allows us to test our ensemble subselection approach in a period where we have no real observations.

Moreover, in LN17, the sensitivity of the constraint to the number of predictors was not tested. Sanderson et al. (2017) highlight the difficulty of choosing the metrics most suitable to constrain a particular projection. Compared to targeted metrics, multi-variate metrics were found to be more robust to changes in the spatial domain. Borodina et al. (2017) studied future changes in high-latitude temperature variability by finding relationships with the models' present-day performance in sea-ice related metrics. Following the emergent constraint procedure, a reduction in spread compared to the full ensemble was found for many metrics. A robust and physically meaningful constraint was found when combining multiple metrics across all seasons (a so-called "broad constraint"). Having such an aggregated constraint leads to higher probability of capturing the relevant processes for the future climate. Narrow constraints, which only consider one metric or season, were found to potentially lead to overfitting. In contrast to LN17, Sanderson et al. (2017) and Borodina et al. (2017) both considered multiple predictors but combined them in a single cost function rather than conducting multi-objective optimisation.

Here, we build on the idea of LN17 by assessing the potential of Pareto-optimal estimates to constrain future Australian precipitation change, which is characterised by large model disagreement. Different to LN17, no observational constraint is provided as the goal is not to show actual projections for Australia precipitation but rather to explore the potential and caveats of using a model subset-selection approach to narrow constraints of future climate change and how it affects our uncertainty estimates. The constraint is based on models' ability to appropriately simulate present-day total precipitation in Australia, Pacific sea surface temperature, and high latitude 500 hPa eastward wind climatologies of a given model as "truth". Those predictors are chosen based on their physical connection to Australian precipitation. We use either one, two, or all three variables

as predictor(s). When using more than one predictor, we apply the concept of Pareto optimality, which results in an ensemble of model subsets. Out-of-sample skill of the calibrated subsets is then tested using a range of model-as-truth experiments. As the spread of an ensemble consisting of subset averages is typically too narrow (overconfidence), we apply mixture distributions for better uncertainty estimation. Skill is then assessed by comparing the calibrated subsets to the individual CMIP5 models and by studying the accuracy and precision of the projected change.

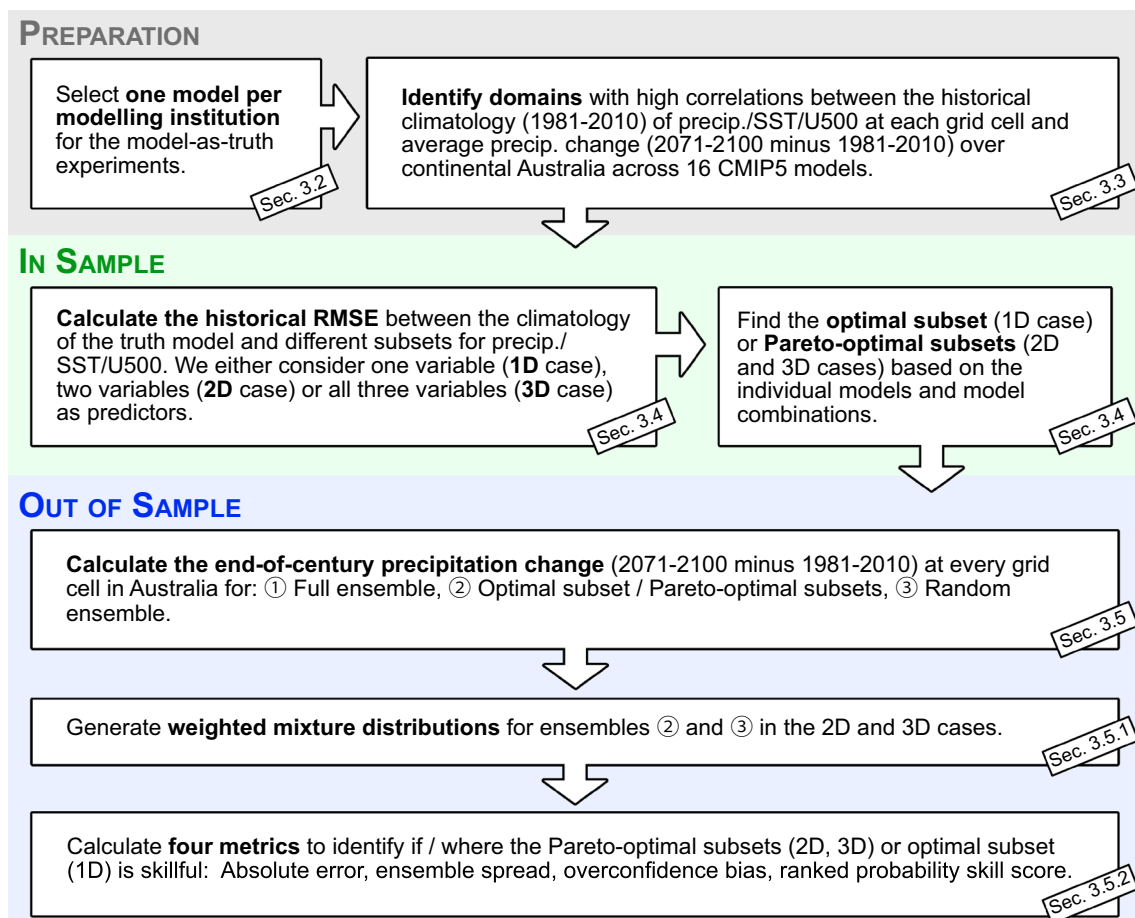
## 2 Brief overview

Our goal is to examine the potential of constraining future Australian precipitation change using climate model subsets, which are optimally chosen based on each models' historical performance. More specifically, we are trying to address the following questions: is a model subset that performs well in-sample still skillful out-of-sample? How much real skill can be gained from optimal subset

selection relative to random sampling or the full ensemble? How sensitive are the results to the chosen number of predictors and what those predictors are? How can we interpret the spread of an ensemble consisting of subset averages?

Different from LN17, no observations and reanalyses were used to come up with a best possible way to constrain end-of-century Australian precipitation change. This is a methodological study first and foremost in which the out-of-sample testing of calibration approaches is encouraged, and multi-objective optimisation is explored. However, assessing the full extent of projected changes that applying this approach has to Australian precipitation change is a potential future study.

We use the flowchart shown in Fig. 1 to guide us through the process. Generally, the process can be split into three parts: data preparation (grey); calibrating ensembles in-sample based on historical performance (green); and applying those subsets out-of-sample to climate change projections (blue). The whole study is based on model-as-truth, or "perfect model" experiments which are introduced in Sect. 3.2.



**Fig. 1** Flowchart of the data preparation, in-sample and out-of-sample steps. More information about the specific steps can be found in the sections referred to in the small white boxes

Section 3 contains a description of the method used. This includes an introduction to the model data and variables used throughout this study (Sect. 3.1), model-as-truth experiments (Sect. 3.2), and the identification of the relevant spatial domains for each variable separately (Sect. 3.3). We also discuss how we calibrate ensembles in-sample for a varying number of predictors (Sect. 3.4) before we apply them to constrain future Australian precipitation change in Sect. 3.5. This is also where we discuss the use of mixture distributions to obtain a more meaningful estimate of ensemble spreads and the metrics used to assess skill. In Sect. 4 we present the results, explained in terms of both accuracy and precision. Finally, Sect. 5 contains a discussion and Sect. 6 the conclusions.

### 3 Method

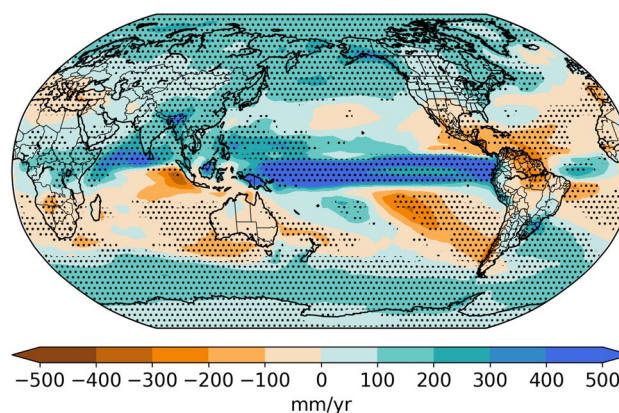
#### 3.1 Model data

We use 16 CMIP5 model runs, each from a different modelling institution, which cover the historical period (1981–2010; RCP8.5 after 2005) and the RCP8.5 future projection scenario (2071–2100); see Table S1 in the Supplement. We motivate why one model per modelling institution was chosen in Sect. 3.2. Our study is based on absolute values of gridded monthly total precipitation (Precip.; model variable: pr), sea surface temperature (SST; model variable: tos), and 500 hPa eastward wind (U500; model variable: u500). Historical climatologies (1981–2010) of those variables were used to constrain end-of-century precipitation change (2071–2100 relative to 1981–2010) over Australia. The model output was regridded to a common  $2.5^\circ \times 2.5^\circ$  grid using bilinear interpolation. Maps of historical climatologies for those three variables are presented in the Supplementary Fig. S1.

Figure 2 shows the end-of-century precipitation change based on the periods 2071–2100 and 1981–2010 for the multi-model mean of 16 CMIP5 models. Stippling indicates that at least 80% of the models agree on the sign of the change. The multi-model mean projects a general decrease in total precipitation for Australia, but it is evident that there is no strong model agreement (no stippling). When looking at the maps of end-of-century precipitation change for the individual models (Supplementary Fig. S2) the extent of this disagreement becomes evident. No observational products were used for this study as model-as-truth experiments were conducted instead (see Sect. 3.2).

#### 3.2 Model-as-truth experiment

As we will explain in later sections, model subsets are selected based on the models' historical performance. To



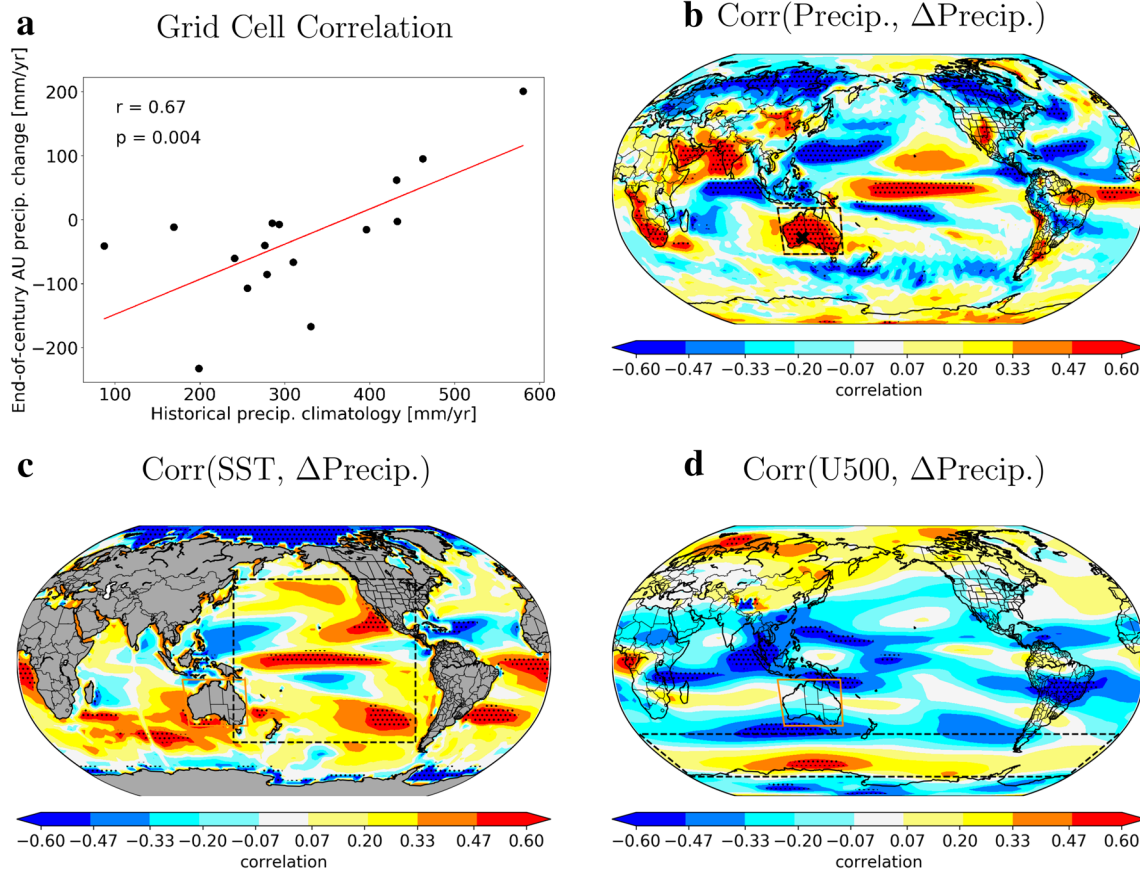
**Fig. 2** End-of-century precipitation change (2071–2100 relative to 1981–2010) averaged across all CMIP5 models. Stippling indicates that at least 80% of the models agree on the sign of the precipitation change

test whether a given subset selection strategy is skillful for projections (when no observations are available), we conduct a series of model-as-truth experiments (often also referred to as perfect model tests). Such experiments are sometimes used in climate science (e.g., Abramowitz and Bishop 2015; Knutti et al. 2017; Sanderson et al. 2017) to ensure that a particular pattern found *in-sample*—where we are calibrating our ensemble—persists into the future, which we refer to as *out-of-sample*. Cross-validation could be considered as the equivalent technique in statistics. For the model-as-truth experiments, one simulation per modelling institution is used as the “truth”, as if it were observed data. The calibration is then conducted on the remaining 15 models based on the in-sample period 1981–2010 and a given set of variables (see Sect. 3.4). The skill of the calibrated ensemble can subsequently be tested out-of-sample given that “pseudo-observations” are now available until the end of the 21st century (see Sect. 3.5). This experiment is repeated for all models as truth and results are averaged across all experiments. We use one simulation per modelling institution from the original ensemble as initial condition members from one model, or simulations from closely related model versions of the same institution (Knutti et al. 2013; Boé 2018) are likely to be much closer to each other than to a real observational product. This step is required if we want our results to be transferable to a real world situation. This is consistent with suggestions by Leduc et al. (2016) and Sanderson et al. (2017). Note, that model-as-truth experiments should be considered to be a necessary but not sufficient condition for true out-of-sample skill. One reason being that shared assumptions in models (which adds to the dependence issue) are likely not detected by assessing model-model similarity relative to model-observation similarity during the model pre-selection process. Other studies account for model dependence by

downweighting models that have similar biases (Sanderson et al. 2015b; Knutti et al. 2017; Lorenz et al. 2018). Another aspect to be aware of when using model-as-truth experiments is that a climate model could potentially simulate climate states that may not be observable in the real world. So, calibrating towards such a truth might be erroneous. While model-as-truth experiments are useful out-of-sample tests for long-term climate change, it could in some cases be sufficient to make use of long observational records. This is particularly the case when the intended application of the weighting or subset selection approach is not characterised by significantly different forcing conditions compared to what has been observed in the past. The issue with such an approach is the limited availability of long observational records (for most variables and regions) and the quality of those observational products might change over time. In any case, ensuring that the out-of-sample test mimics as well as possible to intended application is essential.

### 3.3 Identifying the spatial domains

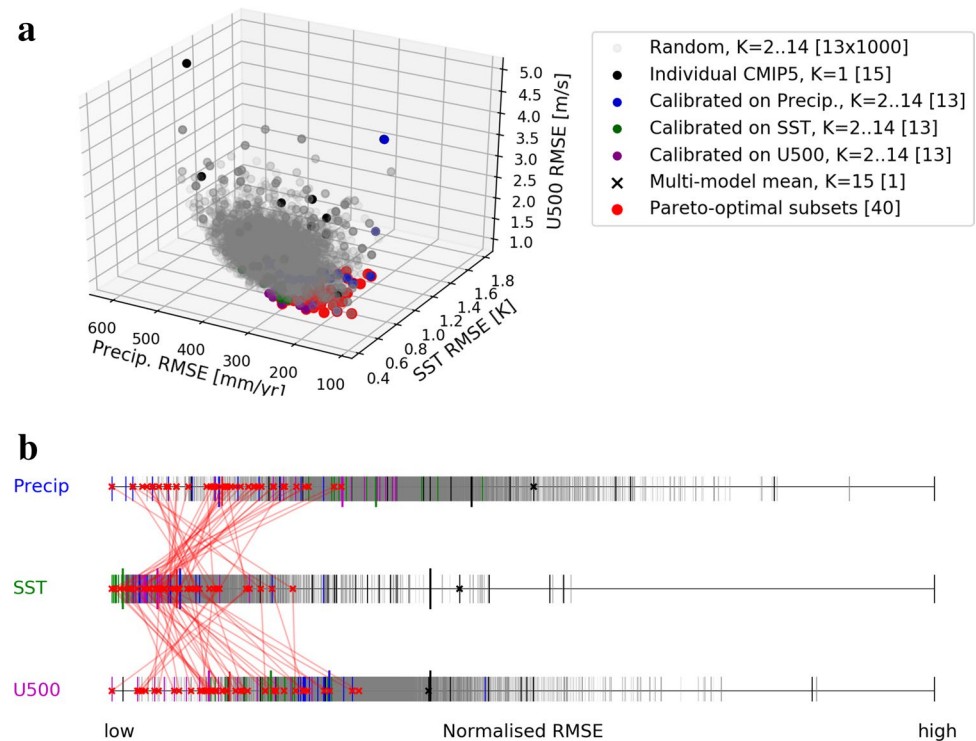
We identify a spatial domain separately for each variable which is used to constrain future precipitation change over Australia (grey box in Fig. 1). Precipitation, sea surface temperature and 500 hPa eastward wind fields are potentially relevant for future precipitation change over the Australian domain. Grid point correlation maps across models as shown in Fig. 3 were used to help select the appropriate spatial domains. The idea is shown in Fig. 3a for a grid cell over Australia (marker X in b). There are 16 markers, one for each climate model. A high correlation between historical CMIP5 model precipitation climatology at this grid cell and end-of-century precipitation change averaged over Australia (38.75° S–11.25° S and 111.25° E–153.75° E; orange rectangle on the maps) is shown. In Fig. 3b we repeat this idea for every grid cell separately. This is equivalent to what LN17 did in their Fig. 4a, b. Here, significant Pearson correlation coefficients



**Fig. 3** a Correlation across climate models between the historical precipitation climatology at the grid cell highlighted with a black cross in **b** and future precipitation change averaged over Australia. Correlation is computed across 16 CMIP5 models (black circular markers). The remaining panels show correlation maps used to identify the spatial domains for precipitation (**b**), sea surface temperature (**c**), and 500 hPa eastward wind (**d**). Correlations are calculated

between historical climatologies at every grid cell and end-of-century precipitation change averaged over the Australian continent (orange box). As in **a**, Pearson correlation coefficients are computed across 16 different models. Stippling is used to highlight significant correlations ( $p < 0.05$ ). Black dashed boxes show the identified domains which are used during the calibration process of our subsets

**Fig. 4 a** Three dimensional space of RMSE values for total precipitation, SST and U500 calculated from the individual models and model averages with the CESM1-CAM5 (r1i1p1) model as truth. Results are based on historical climatologies (1981–2010) using the spatial domains found in Sect. 3.3. Different markers are explained in the legend together with the subset size  $K$  and the number of models/subsets in square brackets. **b** The three axes separately. Low (left) and high (right) normalised RMSE values of the individual CMIP5 models and subsets for the 3D case are displayed. Colours are equivalent to what has been used in **a**



( $p < 0.05$ ) are highlighted with stippling. Generally, regions of high (positive or negative) correlations are our domains of interest. For precipitation, we decide to use continental Australia as our domain. This is the same spatial domain as we use later on for precipitation change. Grid point correlation maps are also created for the other two variables. For sea surface temperature (Fig. 3c) we select the whole Pacific ( $48.75^{\circ}\text{S}$ – $48.75^{\circ}\text{N}$  and  $141.25^{\circ}\text{E}$ – $281.25^{\circ}\text{E}$ ; black dashed rectangle) and for 500 hPa eastward wind (3d) we decide to use the high latitude band in the Southern Ocean as our domain ( $71.25^{\circ}\text{S}$ – $43.75^{\circ}\text{S}$  and  $1.25^{\circ}\text{E}$ – $358.75^{\circ}\text{E}$ ; black dashed rectangle).

Note that statistical relationships found between a present-day metric and projections with some plausible underlying physical mechanism are more likely to prove robust, whereas the relationship is more likely to be spurious for those with a lack of physical explanation. Perkins and Pitman (2009) found that omitting weak climate models in terms of simulating the observed probability density function of precipitation helps to better explain future climate changes over Australia. It leads to changes in the projected patterns of precipitation change and the frequency of rainfall (Pitman and Perkins 2008). They argue that a model that is unable to simulate the present day climate is not suited to simulate the future, as it likely cannot simulate the drivers and associated feedbacks appropriately. This argument motivates the use of Australian precipitation climatology as a predictor in our study. Using Pacific SSTs as a predictor for Australian precipitation change is reasonable, given that

the correlation map (Fig. 3c) shows a pattern that resembles the El Niño–Southern Oscillation (ENSO). The relationship between ENSO and Australian climate in both models and observations has been studied extensively (e.g., Power et al. 2006; King et al. 2015). ENSO drives rainfall changes over Australia, with predominantly dry conditions during El Niño years and wet conditions during La Niña years. Finally, we have chosen the 500 hPa eastward wind field south of Australia to incorporate the effect of changing wind speed and position on future Australian precipitation change. The Southern Annular Mode describes variability in the westerly jet that circles Antarctica, dominating the middle to higher latitudes of the Southern Hemisphere. A change in the position of this westerly wind belt can influence Australia’s rainfall variability (Risbey et al. 2009).

Despite deciding on these domains based on correlation maps (as is commonly done in empirical prediction studies) with the potential of spurious correlations, we have reasonable confidence they are appropriate, given that there are physical links present (as described above). Other methods would likely come up with slightly different domains. Here, the goal is not to come up with the best possible constraint of Australian precipitation change, but rather to introduce a novel method based on Pareto-optimal subsets and model-as-truth experiments that can help us to quantify how prediction accuracy and precision are affected by calibration approaches (such as subset selection or model weighting). We generally choose large domains to avoid the risk of overfitting and thus making the result sensitive to the ensemble

at hand, as motivated by LN17. If the goal was to constrain future Australian precipitation change with real observations and provide a best estimate of the change that can be expected, then a more rigorous approach of domain and variable selection could and should be considered. However, this is not the aim of this study and identifying the best possible variables and corresponding spatial domains should be the focus of a separate study.

### 3.4 Calibrating ensembles in-sample

After identifying a spatial domain for each predictor variable, we start the ensemble calibration process using the model-as-truth setup (green box in Fig. 1). Similar to Herger et al. (2018), we calculate the historical performance in terms of area-weighted root mean square error (RMSE) between the climatology (1981–2010) of the truth model and different models/subsets for each variable. The area-weighted RMSE of each variable is calculated over the domains identified in the previous step. This is done separately for each model-as-truth.

#### 3.4.1 Results based on one variable (1D case)

We first consider the case where we try to constrain precipitation change using each of the three predictor variables separately. We refer to this as the *1D case*. For a given model-as-truth, we compute the historical RMSE between that truth and all 15 individual CMIP5 model estimates based on the domain identified in Fig. 3 and Sect. 3.3. In addition to the performance of the individual models and the multi-model mean of those 15 models, we compute performance of the multi-model mean of subsets, with subset sizes  $K = 2, \dots, 14$  and ensure that the RMSE between the subset average and the truth is minimised. So, for each subset size  $2, \dots, 14$  we obtain a single best subset. This is done using the state-of-the-art mathematical programming solver Gurobi (Gurobi 2015). Rather than going through all possible combinations of  $K$  models, computing the RMSE for each combination and then finding the one with the smallest RMSE compared to the truth, Gurobi finds a much quicker solution using a branch-and-cut algorithm (Mitchell 2002). We refer to this as a mixed integer quadratic programming problem because the decisions are binary (model simulation is part of the subset or not), the cost function is quadratic, and the constraint is linear. This idea of optimising a single cost function is very similar to what has been done by Herger et al. (2018). Indeed, we use the same code to find the subsets as in their study.

The *optimal subset* is identified as the subset (or individual model) with the overall smallest RMSE compared to the truth across all subset sizes for that particular variable and spatial domain. Evidently, a different optimal subset will

likely be found for each model-as-truth and variable. Given that we identified three variables to be potentially important for constraining end-of-century Australian precipitation change, we end up with three different 1D cases (one per variable). By finding the model subset of a given size that minimises the RMSE compared to a model-as-truth, we end up with an optimal subset typically consisting of models that are more independent of one another. The reason for this is that regional biases are more likely to differ within the subset, and hence cancel in the subset mean if the models are more independent and therefore closer to a random sample (Herger et al. 2018).

#### 3.4.2 Results based on two or three variables (2D and 3D cases)

We can now expand on the idea discussed above by adding one (*2D case*) or two additional variables (*3D case*) as constraints in the subset selection process. We consider the 2D case first.

Suppose we want to use the models' ability to reproduce precipitation and sea surface temperature climatology fields to constrain future Australian precipitation change. There are at least two approaches to doing this. The first is to combine the skill metrics (in our case RMSE of both variables) into a single weighted cost function. The problem with this is that we would have to come up with an arbitrary normalisation factor or weighting of the individual terms in this cost function. Moreover, it is challenging to separately investigate the uncertainties coming from those two variables. An alternative option is *multi-objective optimisation* (Deb 2014). Rather than combining those two objective functions into one, we optimise for both at the same time, and consider the trade-off between the two variables. What we end up with is not a single best solution, but rather an ensemble of "good" solutions (Pareto 1906). This ensemble of solutions is referred to as the *Pareto front* consisting of a non-inferior set of solutions. A subset (or individual model) is part of the Pareto front if it is impossible to improve on one variable (e.g., reduce precipitation RMSE) without making the performance of the other variable worse (e.g., increase SST RMSE).

For the 2D case, there are three combinations of two variables. We first illustrate the idea of multi-objective optimisation using historical precipitation and sea surface temperature as predictors (see Fig. S3). The procedure for the other two variable combinations is equivalent. For a given model-as-truth and variable, we first compute area-weighted RMSE values of 15 individual CMIP5 models and the multi-model mean consisting of the 15 models for the relevant spatial domain, as for the 1D case. In addition to that, we find the best subset based on precipitation climatology for subset sizes  $2, \dots, 14$  using Gurobi (one subset per size) and

compute the corresponding RMSE values for sea surface temperature in order to obtain a dot in an imaginary two-dimensional performance space. We then repeat this procedure the other way around by calibrating on sea surface temperature fields and calculating RMSE values of those 13 subsets ( $K = 2, \dots, 14$ ) for precipitation. As we only find the single best subset for each subset size ( $K = 2, \dots, 14$ ), we also add 1000 subsets of randomly selected ensemble members (with replacement) per subset size to potentially sample the Pareto front better. This is slightly different from what LN17 did. In their study, they calculated all possible combinations of models for subset sizes one through five, rather than just the single best combination per size. They decided not to go to any larger subset sizes due to the computational cost of calculating all possible combinations for larger sizes. We avoid this problem by using the solver Gurobi.

Once the two-dimensional space with RMSE values of precipitation climatology (for the Australian domain) against the truth on one axis and RMSE of sea surface temperature climatology (over the Pacific) on the other axis is populated with models and model averages, the next step is to identify the Pareto front. We identify the Pareto front based on the data points mentioned above (first five legend entries in Fig. S3) using a Python implementation of the Simple Cull algorithm. The number of subsets that are part of the Pareto front changes depending on the model-as-truth and variable combination. On average, approx. 11 subsets were found to be part of the Pareto front with an average subset size of 4 climate models (see Fig. S4b–d). We refer to the subsets part of the Pareto front as *Pareto-optimal subsets* (Pareto-optimal subensembles in LN17).

The procedure for the 3D case is similar to the 2D case, just with an additional variable on the third axis, see Fig. 4a. This space is filled with RMSE values of the 15 individual models ( $K = 1$ ; black markers), the multi-model mean ( $K = 15$ ; black cross), 13 subsets calibrated on precipitation ( $K = 2, \dots, 14$ ; blue dots), 13 subsets calibrated on sea surface temperature ( $K = 2, \dots, 14$ ; green dots) and 13 subsets calibrated on 500 hPa eastward wind ( $K = 2, \dots, 14$ ; purple dots). In addition to that, we add 1000 subsets consisting of randomly chosen ensemble members for each subset size between 2 and 14 (with replacement; grey dots). The algorithm which finds the subsets that are part of the Pareto front uses all the above mentioned points in the cloud (first six legend entries) as an input. We do not expect the readers to see every single point in this cloud but have chosen this 3D graph as an illustration to show which models and model subsets go into the algorithm. The resulting Pareto-optimal subsets in this three-dimensional space are highlighted with red circular markers in Fig. 4a. Note that a perfect model or subset mean would be positioned at the origin of this space (closest corner with respect to the viewer). For the 3D case, around 50 subset means are part of the Pareto front across all

models-as-truth with an average subset size of 4.7 (see Fig. S4a). In LN17, only the three-dimensional Pareto front was used for constraining Californian precipitation change. Here we are interested in comparing the Pareto-optimal subsets of the 3D case with the subsets of the 2D case and the optimal subset of the 1D case.

Figure 4b confirms that the subsets that make up the Pareto front behave as expected in the 3D case: there is a performance trade-off between the three variables. The three horizontal strips represent normalised in-sample RMSE values, sorted from low (left) to high (right), for the three predictor variables using the CESM1-CAM5 model as the truth as an example. Those strips are essentially equivalent to the three axes in the 3D space shown in a. Blue line segments indicate subsets ( $K = 2, \dots, 14$ ) calibrated on precipitation, green is used for subsets calibrated on sea surface temperature and purple for subsets calibrated on 500 hPa eastward wind. Black line segments are used for individual CMIP5 models, whereas the thick black cross shows the multi-model mean ( $K = 15$ ). Random subsets are shown in grey. Thicker and slightly longer line segments indicate average RMSE values for the respective colours. Lastly, red crosses are used to highlight ensemble members which are part of the Pareto front. Equivalent Pareto-optimal subsets are connected across the three strips using thin red lines.

As expected, the subset with the overall best performance (lowest RMSE) for precipitation is a subset calibrated on precipitation (blue line segment furthest to the left). The same is true for sea surface temperature and 500 hPa eastward wind (green/purple line segment is furthest to the left). Also unsurprisingly, the subset averages clearly show better performance compared to the individual models (Knutti et al. 2010b).

### 3.5 Applying the calibrated ensembles to future precipitation change

In the previous section we identified the optimal subset in the 1D case and Pareto-optimal subsets in the 2D and 3D cases based on historical performance (in terms of in-sample RMSE). As a next step (blue box in Fig. 1), we calculate end-of-century precipitation change (2071–2100 relative to 1981–2010) at every grid cell in Australia using the subsets identified in the previous sections. For each model as truth, our goal is to assess out-of-sample skill of the following three ensembles:

*Full ensemble* This ensemble consists of the original 15 CMIP5 models.

*Optimal subset and Pareto-optimal subsets* These consist of subsets or individual models which are part of the Pareto front (2D and 3D cases) or the optimal subset (1D). The subset sizes (and number of subsets) differ depending



on the model considered as truth and variable(s) used as predictor(s).

**Random ensemble** In the 1D case, the random ensemble consists of 1000 randomly chosen subsets (with replacement) with the same ensemble size as the optimal subset. For the 2D and 3D cases, we have 1000 randomly chosen subsets for each subset size that is part of the Pareto front, chosen so that the relative occurrence of ensemble sizes is equivalent to that in the Pareto-optimal subsets. Note that this random ensemble is different from the random ensemble (grey dots) shown in Fig. 4, which was generated to more densely populate the cloud in-sample.

LN17 assessed the skill of their Pareto-optimal subsets through comparison of the spread of subset means with those of the original CMIP5 models (full ensemble in our case) and the means of all subensembles (all possible subsets of sizes  $K = 1-5$ ). This is however not a fair comparison given that a distribution of ensemble averages (as is the case for the Pareto-optimal subsets) will be narrower than a distribution which contains individual models (e.g., the “Original CMIP5 models” and more importantly, the “All subensembles” case in LN17) as a result of regional differences and internal variability being reduced more when averaging over models. We have overcome this problem by generating the random ensemble, which consists of the same relative occurrence of subset sizes as the Pareto-optimal subsets. This enables us to make a fair comparison between the Pareto-optimal subsets and the random ensemble, without having to deal with the issue of varying subset sizes.

However, a comparison between those two distributions and the distribution of the full ensemble (or a scalar truth) still cannot be made due to the cancellation of errors when averaging model patterns together (Knutti et al. 2010b; Annan and Hargreaves 2011; Pincus et al. 2008). The spread of an ensemble consisting of subset averages would therefore typically under-represent the true uncertainty, so this interpretation represents *overconfidence*. An improved

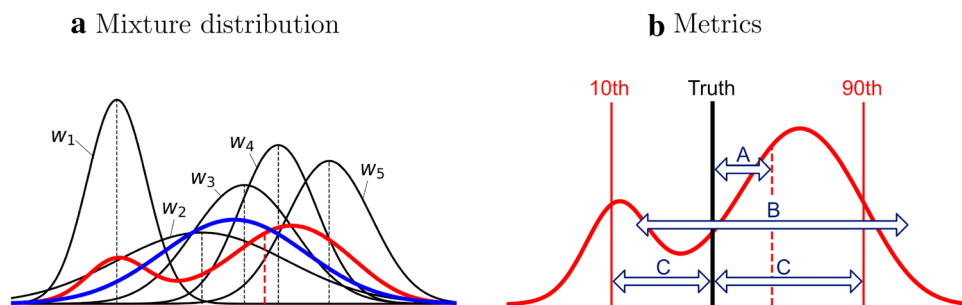
uncertainty estimation is achieved using weighted mixture distributions, as described below.

### 3.5.1 Mixture distributions

As mentioned above, the subsets part of the Pareto front in the 2D and 3D cases (and therefore the random ensemble) consist of subset averages of varying subset sizes. Naively interpreting the spread of an ensemble consisting of subset means as uncertainty (as shown with the blue distribution in Fig. 5a) is not appropriate due to the reduced variability that is a consequence of the averaging process. We therefore need to come up with a way to make them comparable. This transformation is done using a weighted *mixture distribution*.

As an example, assume that our Pareto front consists of five Pareto-optimal subsets, each with a different subset size (e.g.,  $K = 10, 4, 7, 9, 5$  members). As a first step, we fit a Gaussian distribution to estimates of precipitation change (at a given grid cell) separately to each of those five subsets. We end up with five Gaussian distributions with different means and standard deviations (black distributions in Fig. 5a). We assign different mixture weights  $w_i$  to those Gaussian distributions (often referred to as mixture components) which are proportional to the subset size (in the example above:  $w_i = 0.29, 0.11, 0.20, 0.26, 0.14$ ). From those five distributions we can generate a weighted mixture distribution (red distribution in Fig. 5a) by randomly sampling estimates of precipitation change from each distribution. This random sampling is done proportional to the weights  $w_i$ . So, the higher  $w_i$ , the higher the chance that we sample from the corresponding Gaussian distribution to obtain values for the mixture distribution. We have now successfully transformed an ensemble of subset averages into a single mixture distribution, which can now be compared to a scalar truth.

This procedure is then repeated for each grid cell in Australia (120 in our case), each model-as-truth, and each ensemble of model subsets.



**Fig. 5 a** Schematic of the concept of a mixture distribution (red) based on five Gaussian distributions (black) with their respective weights  $w_i$ . The blue distribution is simply the distribution of the mean of draws from each black distribution. **b** Three out of the four metrics used in this study to assess skill. A: absolute error, B: ensemble spread, C: overconfidence bias. Note that the length of arrow A correctly depicts the error metric whereas this is not the case for the arrows B and C, which are only there for illustration. We refer to Sect. 3.5.2 for more details on those metrics

ble spread, C: overconfidence bias. Note that the length of arrow A correctly depicts the error metric whereas this is not the case for the arrows B and C, which are only there for illustration. We refer to Sect. 3.5.2 for more details on those metrics

### 3.5.2 Metrics used to assess skill

The following four metrics are used to compare out-of-sample results across 1D, 2D and 3D cases and the three ensembles listed above (see Fig. 5b and last step in Fig. 1):

**Absolute error** For any given grid cell, we calculate the absolute distance between the truth and the ensemble mean, and then average across models-as-truth to gauge how close the distribution is from the truth. We refer to this as *accuracy*. The goal is of course to maximise accuracy. See arrow A in Fig. 5b for a graphical representation of this metric.

**Ensemble spread** For any given grid cell and model-as-truth, we calculate the standard deviation of the precipitation change distribution to get an estimate of the ensemble spread. This is often referred to as *precision*. We are aiming to increase precision by reducing the ensemble spread (without being overconfident). See arrow B in Fig. 5b for a graphical representation of this metric.

**Overconfidence bias** For a given grid cell, we compute the 10th and 90th percentiles of the distribution. We then count how often the truth is within this range across all models-as-truth. We would of course expect the truth to be on average 80% of the time within the 10th–90th percentile range if the truth was a random draw from the same distribution. This metric helps us determine if our distribution is overconfident, if the truth is less than 80% of the time within the 10th–90th percentile range. This statistical concept is also called *coverage*. Arrows C in Fig. 5b illustrates this idea.

**Ranked probability skill score** The ranked probability skill score (RPSS) is a metric often used in weather forecasting, but not commonly used in climate science (Weigel et al. 2007). It is essentially a combination of accuracy and precision. We define the RPSS between two distributions  $d1$  and  $d2$  as  $RPSS = 1 - RPS_{d1}/RPS_{d2}$ . Here,  $RPS_{d1}$  is defined as the integral of the squared difference (and not area) between the empirical cumulative distribution (ECDF) of  $d1$  and the scalar truth (whose ECDF is a step-function). An example is shown in Fig. S5 in the supplementary information. RPS therefore penalises distributions with low precision and accuracy. A negative RPSS value indicates higher skill of  $d2$  compared to  $d1$ . Higher skill of  $d1$  relative to  $d2$  means that  $RPS_{d1} < RPS_{d2}$ , which leads to positive RPSS values. RPSS values in the results section are shown as maps of Australia averaged across all 16 models-as-truth experiments.

In the following results section, maps of these metrics are shown to compare the full ensemble consisting of 15 individual models with (Pareto-)optimal subsets (after converting them to mixture distributions). Results with the random ensemble as the reference product are very similar and are thus not shown. Note however that this similarity in results is not guaranteed for different uses cases or ensembles and we therefore recommend to perform a comparison to random

sampling when applying this approach to a different situations (or when using real observations).

## 4 Results

As discussed in the previous section, the skill of different ensembles is assessed based on four different metrics. For each metric, we first present a summary figure comparing the different cases with varying numbers of predictors (3D, 2D and 1D cases) and then show maps of Australia for three of those cases to investigate the spatial pattern of the constraint. All results are based on end-of-century Australian precipitation change predicted via the ensembles described at the beginning of Sect. 3.5. For ensembles consisting of model subsets, we first transform the distribution consisting of subset averages into a weighted mixture distribution, as described in Sect. 3.5.1. The goal is to investigate if multi-objective optimisation leads to more accurate (reduced bias) and precise (reduced variance) probabilistic projections compared to single-objective optimisation. Those two terms were introduced in the previous section.

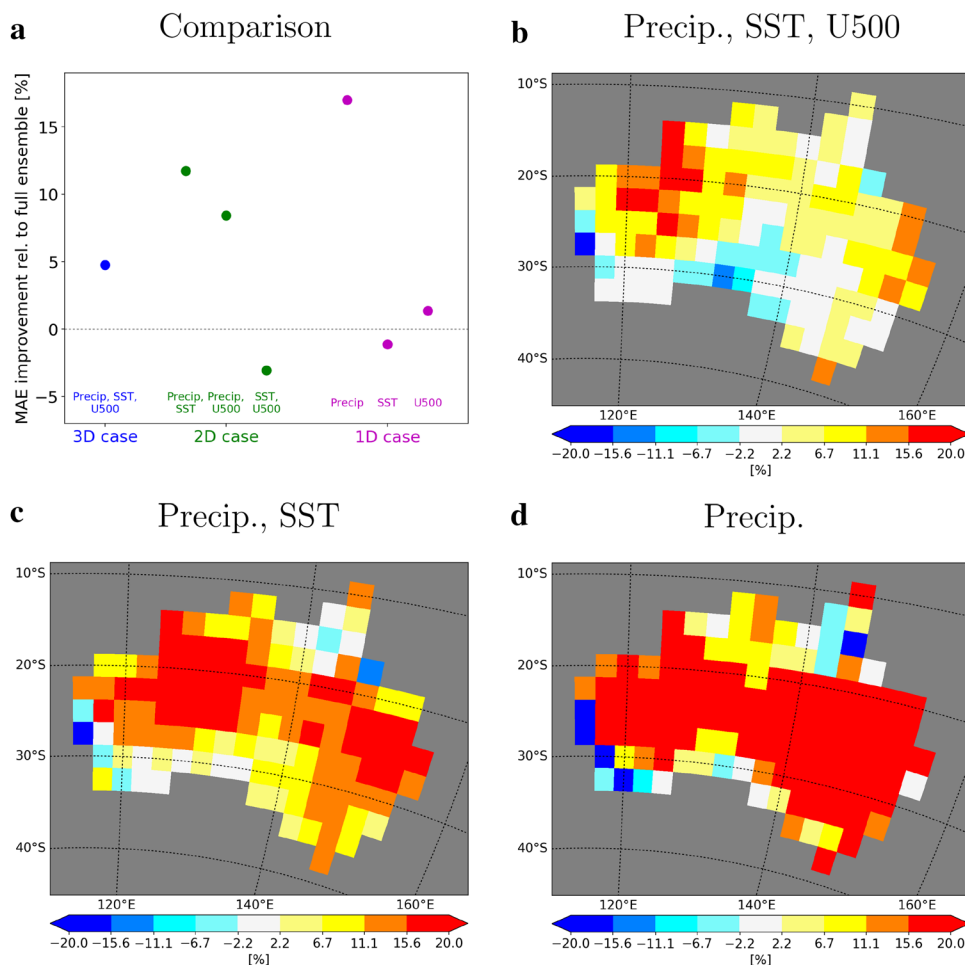
### 4.1 Accuracy

We first consider how close the means of Pareto-optimal subsets (3D and 2D) or optimal subset (1D) are on average from the truth compared to the full ensemble. As noted above, this metric is closely related to accuracy. Here and in the remaining results section when comparing two ensembles  $A$  and  $B$ , we refer to improvement in  $B$  relative to  $A$  if  $(metric_A - metric_B)/metric_A > 0$ , where  $metric_A$  is a measure of the skill of ensemble  $A$  compared to a given truth.

Figure 6a shows a plot of the mean absolute error (MAE) improvement for the 3D case in blue, the 2D cases in green and the 1D cases in magenta. Values shown are area-averages across Australia and all model-as-truth experiments. The variables used as predictors are given at the bottom of the plot. We observe the largest MAE improvement for historical precipitation as single predictor. The other two variables do not seem to be of large importance. The general decrease in MAE improvement when moving towards more predicting variables makes sense as we include “worse” variables (see the relatively low skill of SST and U500 alone and their combination in the 2D case).

The remaining panels in Fig. 6 show maps of Australia for the 3D case (b), the 2D case with precipitation and SST as predictors in (c) and precipitation as the single predictor in (d). Red colour indicates improved skill of the Pareto-optimal subsets (3D or 2D cases) or optimal subset (1D) compared to the full ensemble. Results are averages across all model-as-truth experiments. It supports what has already been shown in (a). The fewer variables we use to constrain

**Fig. 6** **a** A comparison of mean absolute errors averaged over the Australian continent for all cases (1 × 3D case, 3 × 2D case and 3 × 1D case) based on the mixture distributions. The marker indicates the mean across all model-as-truth experiments. The maps of Australia show improvement in mean absolute error of the Pareto-optimal subsets (3D and 2D cases) or optimal subset (1D case) compared to the full ensemble. **b** Results for the 3D case, where we simultaneously calibrate on precipitation, sea surface temperature and 500 hPa eastward wind, **c** is one of the 2D cases, where we drop 500 hPa eastward wind. After additionally dropping sea surface temperature as a predictor, we end up with the 1D case as shown in **d**. Maps are based on averages across all 16 models-as-truth



future precipitation change, the closer we are on average to the truth. In the 3D case, we observe the largest improvement in the Northern part of Western Australia. This area of highest improvement expands further east when removing additional predictors. Results are very similar if we use the random ensembles as the reference instead of the full ensemble (not shown). Results for the remaining variable combinations are shown in Fig. S6.

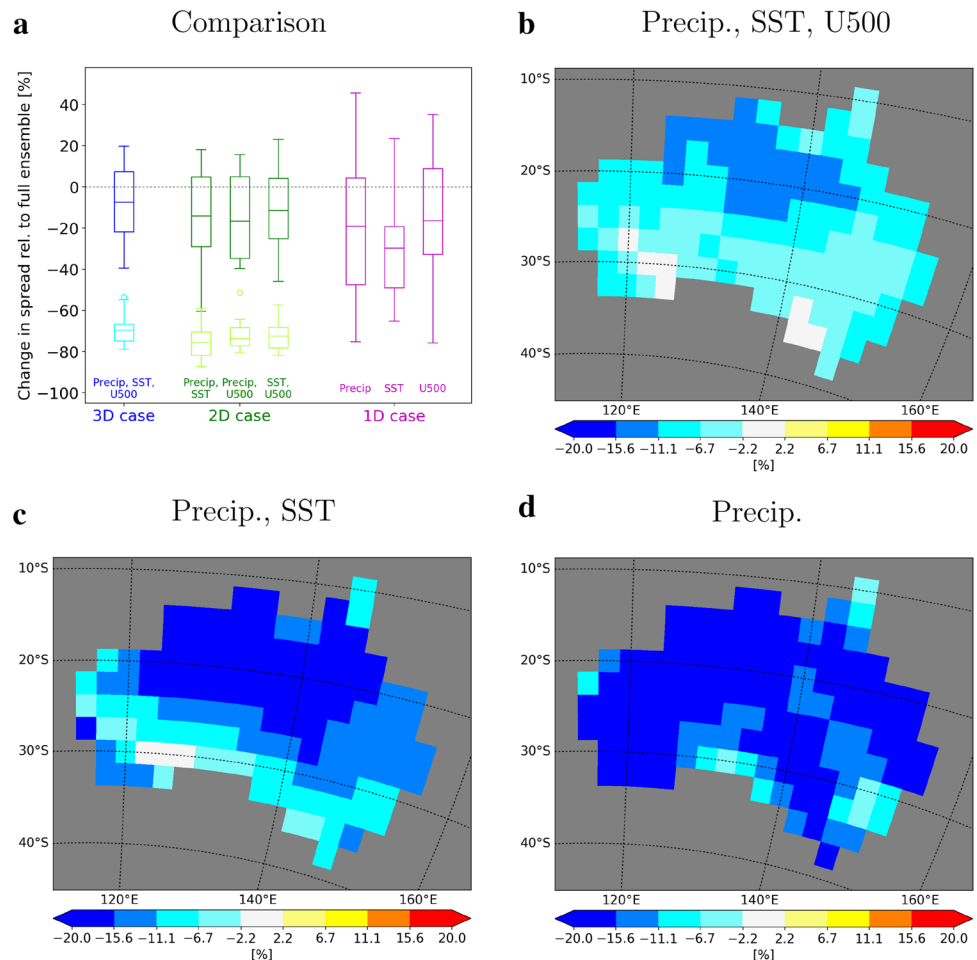
### 4.2 Precision

As motivated earlier, the spread of a distribution consisting of subset averages cannot be interpreted as a measure of uncertainty. Such a distribution would automatically be too narrow due to the effect of cancellation of errors. The mixture model (as explained in Sect. 3.5.1) was introduced as a way to overcome this problem. A comparison between the ensemble spread of those two distributions and the full ensemble is shown in Fig. 7a. Light colours are used for the distribution of the subset averages and darker colours for the mixture distributions. The 1D case only shows results based on the “corrected” spread (via mixture model) as a spread

estimate cannot be obtained from a single optimal subset mean. Figure 7a shows a general decrease in spread relative to the full ensemble for all cases. As expected, the ensemble spread consisting of subset averages is significantly smaller than the spread of the mixture distribution. The largest reduction in spread occurs in the 1D case. This is also illustrated by the maps in Fig. 7b–d, where blue colours indicate a decrease in spread of the underlying mixture distribution compared to the full ensemble. The spread decreases with fewer predictors. We observe a mean decrease in spread of up to 20% in the case when precipitation acts as the sole predictor for future precipitation change. Overall, the areas of largest spread decrease are consistent with areas of largest improvement in mean absolute error. We also generally observe a more homogeneous pattern as we move to fewer predictors. Results for the remaining cases are shown in Fig. S7. If we use the random ensemble rather than the original ensemble as our reference, we obtain very similar results (not shown).

Keep in mind that a spread decrease is generally desirable, but only if the distribution decreases while centred near the truth. We therefore next consider a metric which

**Fig. 7** Similar to Fig. 6, but for change in spread (standard deviation of the underlying distribution). **a** Boxplots containing all 16 model-as-truth experiments. The box represents the interquartile range (IQR; extends from Q1 to Q3 of the data), with a horizontal line for the mean estimate. The whiskers show the range of the data and are at 1.5 IQR. Results based on the mixture model distributions are shown in darker colours and results based on the subset averages in lighter colours. **b–d** Spatial maps of changes in spread for three different cases based on the mixture model results. Blue colours indicate a decrease in spread of the Pareto-optimal subsets or optimal subset compared to the original ensemble

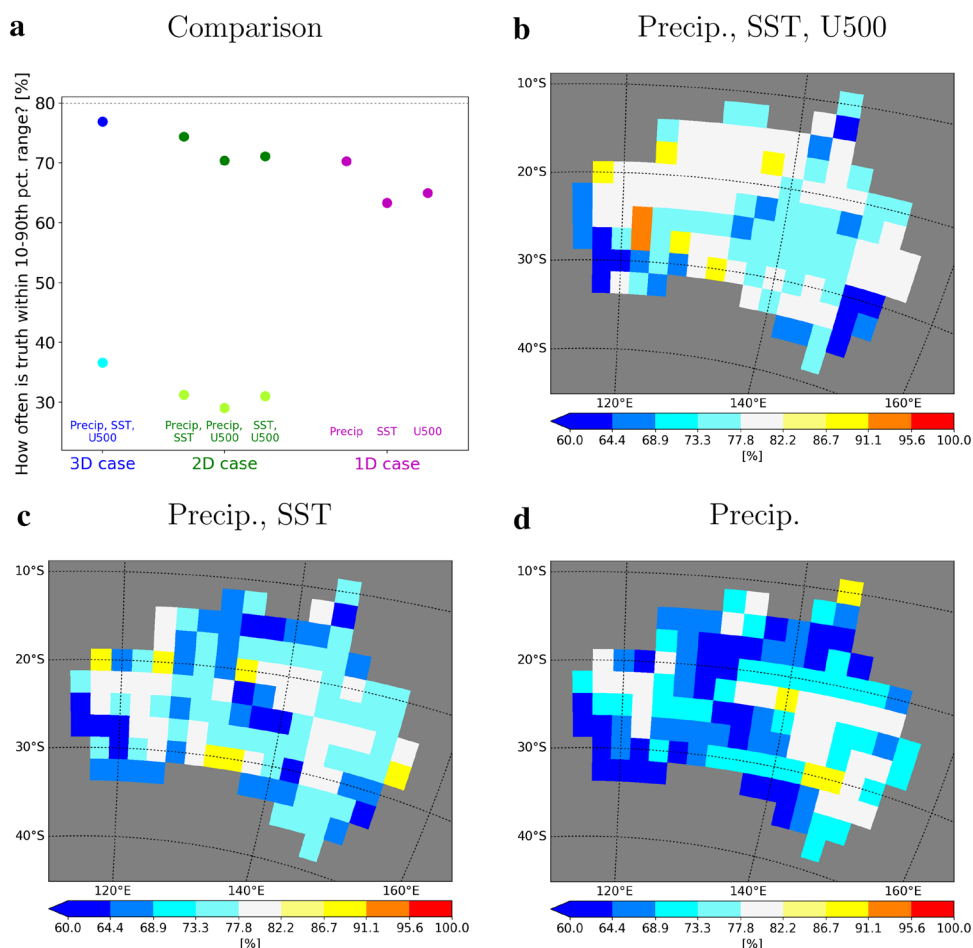


additionally takes the position of the truth into account. In Fig. 8 we test how often the truth is within the 10th–90th percentile range of the Pareto-optimal subsets (3D and 2D cases) and the optimal subset (1D case). If the truth was a draw from the same distribution, then we would expect it to be within the 10th–90th percentile range approximately 80% of the time. In Fig. 8a we see that all the markers are below the horizontal 80%-line, even for the 3D case. For the mixture distributions (darker markers), this might be related to the fact that we assumed Gaussianity when computing the percentile range. The estimated coverage probability for the distribution based on subset averages (lighter markers) are significantly smaller, as expected from the previous figure. This shows how effective the mixture model is in nearly eliminating the overconfidence bias. For both distributions, coverage probability decreases as fewer variables are included as predictors. Note that for Fig. 8a, no uncertainty bars were plotted, as information from all 16 model-as-truth experiments were required to obtain the percentage of a truth being within the 10th–90th percentile range. Coverage probabilities are slightly higher for the case when we use the random ensemble as our reference (not shown).

From Fig. 7 we learned that the ensemble spread decreases most for the 1D cases. However, this comes at a cost as this decrease in spread is likely to lead to overconfidence. Our optimal subset is under-dispersive and therefore fails to encompass the truth regularly. Fig. 8b–d show similar information to (a), but this time separately for each grid cell in Australia based on the mixture distributions. Blue colour indicates that the truth lies within the 10th–90th percentile range less often than 80% of the time. As expected from (a), this probability decreases when moving from the 3D case to the 2D case and eventually to the 1D case. This clearly indicates that a decrease in spread is not necessarily an indication for improved accuracy. Results for the remaining predictor combinations are shown in Fig. S8.

Sanderson et al. (2017) showed that a stronger skill weighting has a more significant effect on projected changes, but also leads to a higher risk of increased under-dispersion. We can understand this as an analogue to our 1D case, where the best estimate of projected change is adjusted the strongest, but at a cost of substantial overconfidence.

**Fig. 8** Similar to Fig. 6 but for the percentage of time the truth lies within the 10th–90th percentile range of the Pareto-optimal subsets (3D or 2D cases) or optimal subset (1D case) across all models-as-truth experiments. Darker colours in a are used for results based on the mixture distributions and lighter colours based on the distributions of subset averages. Overconfidence bias maps are based on the mixture model distributions. Different to Fig. 6, no improvement relative to the full ensemble is shown



### 4.3 Combining accuracy and precision

After looking separately at accuracy and precision, we can now combine those two metrics into a single metric to get additional insights. As introduced in Sect. 3.5.2, the Ranked Probability Skill Score (RPSS) is a metric that takes into account both the width of the distribution and its position relative to the truth.

Figure 9a shows RPSS values for the different cases across all models-as-truth experiments. Positive values indicate improved skill of the Pareto-optimal subsets (for 3D and 2D cases) or optimal subset (for 1D case) compared to the full ensemble. We observe positive mean values for the 3D case and one of the 2D cases. For the remaining cases, mean RPSS values are either zero or negative, indicating that there is no benefit of using a subset instead of the full ensemble. The fewer variables are used for the calibration, the lower the RPSS value tends to be. Interestingly, the error bars for the 1D case are the largest indicating how risky it is to trust a result solely based on one predictor. One can either obtain very large improvements or end up much worse than the original ensemble, but on average the skill will be worse than without calibration. This is consistent with the results

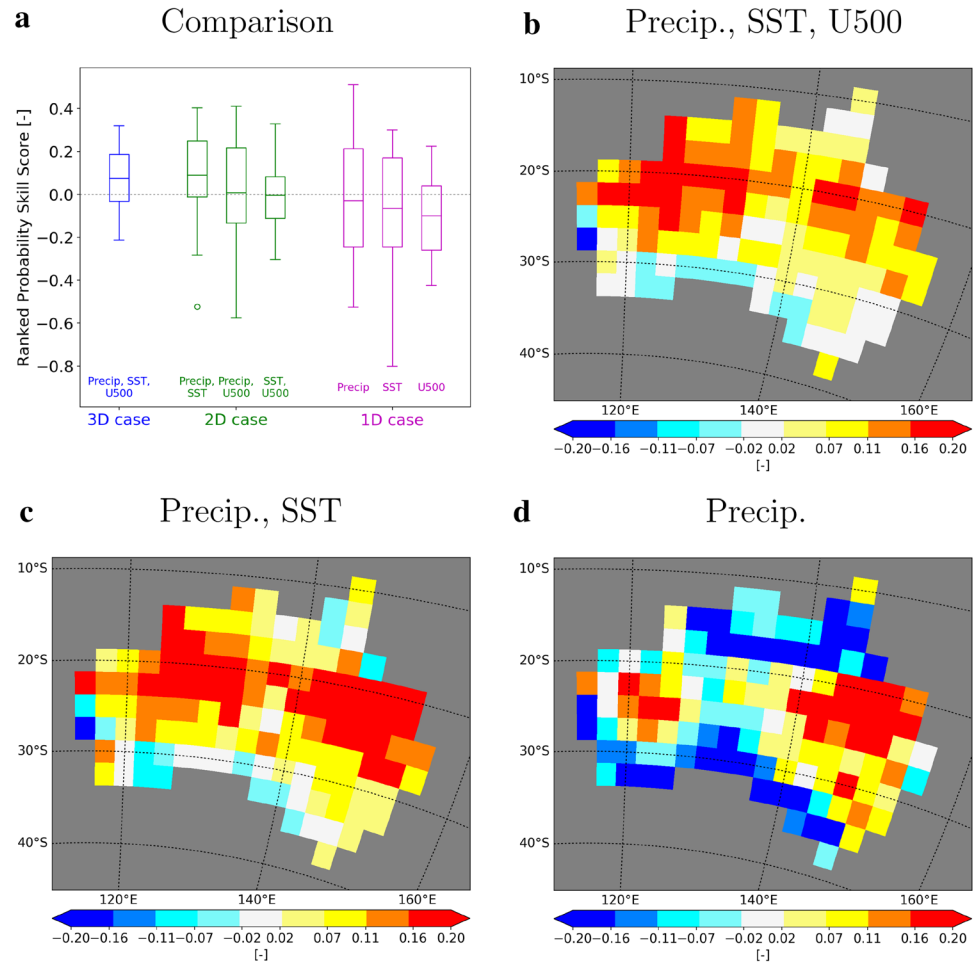
by Weigel et al. (2010). The range of outcomes for the 3D case is much smaller. When using the random ensemble as a reference instead, markers are very similar (not shown here).

The maps in Fig. 9 show RPSS values at every grid cell in Australia for the 3D case (b), one of the 2D cases (c) and one of the 1D cases (d). Results are averaged across all 16 model-as-truth experiments. Overall positive RPSS values are found in (b, c). For the 1D case in which we use historical precipitation climatology as our only predictor, we see patches of strongly positive and other patches of very negative RPSS values. The RPSS averaged over the whole continent is negative, as confirmed by Fig. 9a. This makes the 1D case a risky candidate, as it is prone to overfitting. RPSS maps for the remaining variable combinations are shown in Fig. S9.

## 5 Discussion

This study expands on the ideas introduced by LN17 and Herger et al. (2018). Despite not having the exact same experiment setup as LN17, the main conclusions made here are still largely applicable to their study. Different to the

**Fig. 9** Similar to Fig. 7 but for RPSS values of the Pareto-optimal subsets (3D or 2D cases) or optimal subset (1D case) relative to the full ensemble. Results are based on the mixture model distributions



methods of LN17, we studied a varying number of predictors and also tested the subsets' out-of-sample skill based on a series of model-as-truth experiments. As model averaging automatically reduces the ensemble spread, and thus cannot be compared to the ensemble spread consisting of the original CMIP5 models, we introduced the concept of mixture models, which improves our uncertainty estimation. Having a truth available for future precipitation change, we assess skill of our calibrated ensembles using a range of different metrics. Instead of simply drawing conclusions from a decrease in spread as implemented in LN17, we managed to study mean absolute error, overconfidence of the ensemble, and the ranked probability skill score.

We also highlight the importance of out-of-sample testing when creating subsets or introducing weighting strategies. Only once our approach passes the model-as-truth experiments and we have a handle on the risk of potentially being overconfident, should real observations be used for the constraint. This is also true in the emergent constraints literature, which has a similar aim to what we are doing here. For emergent constraints, often a single variable is used to constrain an ensemble of models with present-day

observations, which ideally leads to a narrower spread across their members than across the full ensemble. This danger of overfitting when weighting too specifically with a small number of predictors has recently also been studied by Borodina et al. (2017); Lorenz et al. (2018). This is also consistent with findings by Knutti et al. (2017), where an aggressive performance weighting leads on the one hand to higher correlation between the true and predicted September sea ice extent, but on the other hand to projection uncertainties that are too narrow. The use of Pareto-optimal subsets is a potential way to combine multiple emergent constraints that try to constrain the same target variable. This would also allow for some physical consistency across predictors. Note that just as for emergent constraints, results should ideally not change fundamentally based on the chosen ensemble, as this would indicate that the patterns found are not necessarily physical ones.

More work is required to guide the community regarding how to work with an ensemble of model subsets. The introduction of mixture models is already a first step which can help with the interpretation of an ensemble spread consisting of model means. This issue of averaging is

likely also a problem in many studies when model averages consisting of different numbers of simulations are compared. The idea of having a range of “good” solutions (and therefore potentially multiple projections) rather than a single best solution is certainly something that adds complexity. However, this can be justified in many cases where multiple variables, spatial scales, observational products and so on are of relevance for the variable of interest.

We note that the Australia-specific metrics chosen here are not exhaustive; other objective functions may be important for precipitation change and its uncertainty in this region. We have identified the role of historical precipitation to be much larger than the ones of SST and U500. For a different application, where the importance of the predictors is more even and is not dominated by a single variable, we may have come to slightly different conclusions. Apart from different variables, one could also include different seasons.

In our study, a subset is “optimal” if its RMSE with respect to a model-as-truth is minimised compared to all the other subsets of the same size. RMSE as a cost function that is being minimised is of course not always ideal. In a future study it might be interesting to include some measure of ensemble spread in the definition of optimality. One could even think of a scenario when our subset should maximise the RPSS, which combines precision and accuracy.

The choice of Gaussian distributions as building blocks for the weighted mixture distributions is an assumption that may not hold perfectly (especially for small sample sizes). However, it is still an advancement when interpreting an ensemble of subset averages.

The choice of predictors and cost function that is being minimised in-sample are central for the success of the approach introduced here. For a weighting approach, Weigel et al. (2010) found that the prediction can be worse than no weighting if the weights are not applied appropriately. This idea is certainly also valid for the subset selection implemented here. If the calibration is not adequate for the intended application, the resulting calibrated subset will likely be a worse predictor than the original ensemble.

The choice of the ideal number of predictors is not straight-forward. Sanderson et al. (2017) found that the effect of their weighting strategy got weaker as more variables were considered. So, their results were not too different from a naive model democracy approach. A recent paper by Lorenz et al. (2018) studied the effect of different predictors on future North American maximum summer temperatures and came to the conclusion that ideally more than one predictor should be used. Model-as-truth experiments as conducted in this study can help guide the user in a direction where skill improves without the danger of being overconfident.

## 6 Conclusion

We presented a method that constrained end-of-century Australian precipitation change based on a varying number of predictor variables and different ensembles. Constraining was implemented based on an optimal subset that minimises a single cost function if only one variable is used as predictor, or multi-objective optimisation and the resulting Pareto-optimal subsets if more than one variable was considered to be important. We introduced a mixture model approach to better assess how ensemble calibration affects projection uncertainty, and highlighted the importance of out-of-sample testing as a necessary but not sufficient condition for confidence in projections, using model-as-truth experiments.

We found that predicting future precipitation change solely based on present-day precipitation climatology led to the largest decrease in mean absolute error compared to either the original ensemble or a random ensemble. However, the ensemble spread was decreased to such a degree that the truth was too frequently outside the ensemble range, which we refer to as overconfidence. When adding more predictors and therefore dealing with Pareto-optimal subsets rather than a single best solution, the ensembles were on average further away from the truth but at the same time reduced the risk of overconfidence. This illustrates an important trade-off between accuracy (*How close is my ensemble to the truth?*) and precision (*How narrow is my ensemble?*), all controlled by the number of predictors. This is an important finding which is likely true irrespective of whether one uses the Pareto-optimal subset selection approach as done here or any other weighting strategy.

**Acknowledgements** We would like to thank Karsten Lehmann for help with the mathematical programming solver Gurobi. Thanks to Jan Sedláček for providing access to the CMIP5 archive based at ETHZ. Nadja Herger, Steven Sherwood, and Oliver Angéllil acknowledge the support of the Australian Research Council Centre of Excellence for Climate System Science (CE110001028). Gab Abramowitz and Scott A. Sisson acknowledge the Australian Research Council Centre of Excellence for Climate Extremes (CE170100023). Scott A. Sisson acknowledges the support of the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (CE140100049). ETH Zurich acknowledges support by the European Union’s Horizon 2020 research and innovation programme under Grant agreement no. 641816 (CRESCENDO) 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 US 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 (2010) Model independence in multi-model ensemble prediction. *Aust Meteorol Oceanogr J* 59:3–6 (Doi: 10.1.1.222.5811).
- Abramowitz G, Bishop CH (2015) Climate model dependence and the ensemble dependence transformation of CMIP projections. *J Clim* 28(6):2332–2348. <https://doi.org/10.1175/JCLI-D-14-00364.1>
- Annan J, Hargreaves J (2011) Understanding the CMIP3 multimodel ensemble. *J Clim* 24(16):4529–4538. <https://doi.org/10.1175/2011JCLI3873.1>
- Annan JD, Hargreaves JC (2017) On the meaning of independence in climate science. *Earth Syst Dyn* 8(1):211. <https://doi.org/10.5194/esd-8-211-2017>
- Bishop CH, Abramowitz G (2013) Climate model dependence and the replicate Earth paradigm. *Clim Dyn* 41(3–4):885–900. <https://doi.org/10.1007/s00382-012-1610-y>
- Boé J, Hall A, Qu X (2009) September sea-ice cover in the Arctic Ocean projected to vanish by 2100. *Nat Geosci* 2(5):341–343. <https://doi.org/10.1038/ngeo467>
- Boé J (2018) Interdependency in multimodel climate projections: component replication and result similarity. *Geophys Res Lett* 45(6):2771–2779. <https://doi.org/10.1002/2017GL076829>
- Borodina A, Fischer EM, Knutti R (2017) Emergent constraints in climate projections: a case study of changes in high-latitude temperature variability. *J Clim* 30(10):3655–3670. <https://doi.org/10.1175/JCLI-D-16-0662.1>
- Box GE, Draper NR (1987) *Empirical model-building and response surfaces*. Wiley, New York (ISBN 0-471-81033-9)
- Caldwell PM, Bretherton CS, Zelinka MD, Klein SA, Santer BD, Sanderson BM (2014) Statistical significance of climate sensitivity predictors obtained by data mining. *Geophys Res Lett* 41(5):1803–1808. <https://doi.org/10.1002/2014GL059205>
- Deb K (2014) *Multi-objective optimization*. Search methodologies. Springer, Boston, pp 403–449. [https://doi.org/10.1007/978-1-4614-6940-7\\_15](https://doi.org/10.1007/978-1-4614-6940-7_15)
- Gurobi Optimization, Inc. (2015) *Gurobi optimizer reference manual*. <http://www.gurobi.com>. Accessed 04 Apr 2018
- Herger N, Abramowitz G, Knutti R, Angélil O, Lehmann K, Sanderson BM (2018) Selecting a climate model subset to optimise key ensemble properties. *Earth Syst Dyn* 9:135–151. <https://doi.org/10.5194/esd-9-135-2018>
- IPCC (2013) Summary for policymakers. In: *Climate change 2013: the physical science basis*. In: Stocker TF, Qin D, Plattner GK, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, Midgley PM (eds) *Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge
- King AD, Donat MG, Alexander LV, Karoly DJ (2015) The ENSO–Australian rainfall teleconnection in reanalysis and CMIP5. *Clim Dyn* 44(9–10):2623–2635. <https://doi.org/10.1007/s00382-014-2159-8>
- Knutti R, Abramowitz G, Collins M, Eyring V, Gleckler P, Hewitson B, Mearns L (2010a) Good practice guidance paper on assessing and combining multi model climate projections. *IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections*, p 15
- Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA (2010b) Challenges in combining projections from multiple climate models. *J Clim* 23(10):2739–2758. <https://doi.org/10.1175/2009JCLI3361.1>
- Knutti R, Masson D, Gettelman A (2013) Climate model genealogy: generation CMIP5 and how we got there. *Geophys Res Lett* 40(6):1194–1199. <https://doi.org/10.1002/grl.50256>
- Knutti R, Sedláček J, Sanderson BM, Lorenz R, Fischer EM, Eyring V (2017) A climate model projection weighting scheme accounting for performance and interdependence. *Geophys Res Lett* 44(4):1909–1918. <https://doi.org/10.1002/2016GL072012>
- Langenbrunner B, Neelin JD (2017) Pareto-optimal estimates of California precipitation change. *Geophys Res Lett* 44(24):12436–12446. <https://doi.org/10.1002/2017GL075226>
- Leduc M, Laprise R, de Elia R, Šeparović L (2016) Is institutional democracy a good proxy for model independence? *J Clim* 29(23):8301–8316. <https://doi.org/10.1175/JCLI-D-15-0761.1>
- Lorenz R, Herger N, Sedláček J, Eyring V, Fischer E M, Knutti R (2018) Prospects and caveats of weighting climate models for summer maximum temperature projections over North America. *J Geophys Res Atmos*. <https://doi.org/10.1029/2017JD027992>.
- Mitchell JE (2002) Branch-and-cut algorithms for combinatorial optimization problems. *Handbook of applied optimization*, pp 65–77
- Pareto V (1906) *Manuale di economia politica*, vol 13. Societa Editrice, Milano
- Perkins SE, Pitman AJ (2009) Do weak AR4 models bias projections of future climate changes over Australia? *Clim Change* 93(3–4):527–558. <https://doi.org/10.1007/s10584-008-9502-1>
- Pincus R, Batstone CP, Hofmann RJP, Taylor KE, Glecker PJ (2008) Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models. *J Geophys Res* 113:D14209. <https://doi.org/10.1029/2007JD009334>
- Pitman AJ, Perkins SE (2008) Regional projections of future seasonal and annual changes in rainfall and temperature over Australia based on skill-selected AR4 models. *Earth Interact* 12(12):1–50. <https://doi.org/10.1175/2008EI260.1>
- Power S, Haylock M, Colman R, Wang X (2006) The predictability of interdecadal changes in ENSO activity and ENSO teleconnections. *J Clim* 19(19):4755–4771. <https://doi.org/10.1175/JCLI3868.1>
- Risbey JS, Pook MJ, McIntosh PC, Wheeler MC, Hendon HH (2009) On the remote drivers of rainfall variability in Australia. *Mon Weather Rev* 137(10):3233–3253. <https://doi.org/10.1175/2009MWR2861.1>
- Sanderson BM, Knutti R, Caldwell P (2015a) Addressing interdependency in a multimodel ensemble by interpolation of model properties. *J Clim* 28:5150–5170. <https://doi.org/10.1175/JCLI-D-14-00361.1>
- Sanderson BM, Knutti R, Caldwell P (2015b) A representative democracy to reduce interdependency in a multimodel ensemble. *J Clim* 28(13):5171–5194. <https://doi.org/10.1175/JCLI-D-14-00362.1>
- Sanderson BM, Wehner M, Knutti R (2017) Skill and independence weighting for multi-model assessments. *Geosci Model Dev* 10(6):2379–2395. <https://doi.org/10.5194/gmd-10-2379-2017>
- Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. *Bull Am Meteorol Soc* 93(4):485–498. <https://doi.org/10.1175/BAMS-D-11-00094.1>
- Tebaldi C, Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. *Philos Trans R Soc Lond A Math Phys Eng Sci* 365(1857):2053–2075. <https://doi.org/10.1098/rsta.2007.2076>
- Weigel AP, Liniger MA, Appenzeller C (2007) The discrete Brier and ranked probability skill scores. *Mon Weather Rev* 135(1):118–124. <https://doi.org/10.1175/MWR3280.1>
- Weigel AP, Knutti R, Liniger MA, Appenzeller C (2010) Risks of climate model weighting in multimodel climate projections. *J Clim* 23(15):4175–4191. <https://doi.org/10.1175/2010JCLI3594.1>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.