Clim Dyn (2014) 43:2297–2308
DOI 10.1007/s00382-014-2054-3

On the generation of climate model ensembles

Ned Haughton · Gab Abramowitz · Andy Pitman · Steven J. Phipps

Received: 12 June 2013 / Accepted: 12 January 2014 / Published online: 31 January 2014

Abstract Climate model ensembles are used to estimate uncertainty in future projections, typically by interpreting the ensemble distribution for a particular variable probabilistically. There are, however, different ways to produce climate model ensembles that yield different results, and therefore different probabilities for a future change in a variable. Perhaps equally importantly, there are different approaches to interpreting the ensemble distribution that lead to different conclusions. Here we use a reduced-resolution climate system model to compare three common ways to generate ensembles: initial conditions perturbation, physical parameter perturbation, and structural changes. Despite these three approaches conceptually representing very different categories of uncertainty within a modelling system, when comparing simulations to observations of surface air temperature they can be very difficult to separate. Using the twentieth century CMIP5 ensemble for comparison, we show that initial conditions ensembles, in theory representing internal variability, significantly underestimate observed variance. Structural ensembles, perhaps less surprisingly, exhibit over-dispersion in simulated variance. We argue that future climate model ensembles may need to include parameter or structural perturbation members in addition to perturbed initial conditions members to ensure that they sample uncertainty due to internal variability more completely. We note that where ensembles are over- or under-dispersive, such as for the CMIP5 ensemble, estimates of uncertainty need to be treated with care.

N. Haughton (✉) · G. Abramowitz · A. Pitman · S. J. Phipps
Climate Change Research Centre Level 4, Mathews Building, University of New South Wales, Sydney, NSW 2052, Australia
e-mail: ned@nedhaughton.com

Keywords Climate model ensembles · Ensemble generation · Ensemble uncertainty

1 Introduction

The probabilistic interpretation of climate model ensemble distributions is well established (Meehl et al. 2007; Tebaldi and Knutti 2007; Knutti et al 2010). However, there has been surprisingly little discussion about how best to generate an appropriate ensemble, which metrics might define an appropriate ensemble, or indeed what ensemble spread actually represents (Knutti et al. 2010). Climate model ensembles can be created in several ways, including by perturbing initial conditions (e.g. Phipps et al. 2013), perturbing model parameters (e.g. Murphy et al. 2004) or by using multiple model structures (i.e. a multi-model ensemble). The ensembles in the Coupled Model Intercomparison Project (CMIP) are typically referred to as "ensembles of opportunity" (e.g. Tebaldi and Knutti 2007; Annan and Hargreaves 2010). These are created by inviting modelling groups to submit model simulations, and then combining them under the implicit assumption that they provide a meaningful and appropriate representation of uncertainty. While the experiments run by CMIP have been strategically designed, contributions are dictated largely by the number of participating groups and the capacities of each group. Some may submit a single simulation, others an entire ensemble, while larger groups may submit several ensembles using different variants of their model. These ensembles are used collectively by researchers and the Intergovernmental Panel on Climate Change (IPCC) to assess how well climate models capture observed climate (Randall et al 2007) and how the climate may change in the future (Solomon et al 2007). This is typically done by

considering all simulations with equal weight. It is therefore of considerable importance that we understand the extent to which spread in ensembles such as CMIP3 and CMIP5 is representative of internal climate system variability, and the extent to which it relates to the uncertainty in creating models of the climate system.

A key step in deciding on appropriate ensemble generation techniques is an explicit acknowledgement of the ensemble interpretation paradigm upon which any analysis will be based. By this, we mean the assumptions regarding: (i) the relationship between observed values and the multi-model mean, and (ii) the meaning of the ensemble spread. This is an essential step in distinguishing between internal system variability and model uncertainty within the ensemble. To date, we are aware of three paradigms, all of which are mutually conceptually incompatible:

The Truth-plus-error paradigm (Knutti et al 2010): Suggests that the discrepancy between a model—at least one without significant biases—and observations is essentially noisy. The multi-model mean averages out the 'noise' from different models, such that if we had enough very good models, the discrepancy between observations and the multi-model mean should be very small. Model error is essentially viewed as a random variable.

The indistinguishable paradigm (Annan and Hargreaves 2010): Suggests that analyses should view models and observations as draws from the same distribution, where this distribution represents our uncertainty in creating an appropriate model structure.

The replicate Earth paradigm (Bishop and Abramowitz 2013): Suggests that chaotic aspects of the climate system mean that it is only partly predictable, with the range of possible climate states given particular set of forcings or initial conditions defining a distribution of "true" climate behaviour. Samples from this imagined distribution are "Earth replicates". Our observational record is a replicate Earth in this sense, and models are viewed as imperfect attempts to create replicate Earths. The concept behind the replicate Earth paradigm is already implicitly accepted in the idea of unpredictable "internal variability" (Knutti et al. 2010; Macadam et al. 2010), although this is usually spoken about within a modelling context, rather than as a property of the natural system.

Note that the indistinguishable paradigm makes the assumption that models and observations are draws from the same distribution, whereas the replicate Earth paradigm suggests that a perfect model—or replicate Earth—would be drawn from the same distribution as observations, emphasising that CMIP models are not replicate Earth-like (see Bishop and Abramowitz 2013). All three paradigms predict that the multi-model mean should provide the best naïve estimate of observations, although for different reasons. While we subscribe to the replicate Earth paradigm in our discussion and analysis below, our results are equally applicable to the indistinguishable paradigm. In particular, we focus on the question of whether observations and ensemble values of surface air temperature are drawn from the same distribution. We therefore build on recent analyses of the CMIP3 ensemble using similar metrics (Annan and Hargreaves 2010; Bishop and Abramowitz 2013), which showed that some variables display a spread in the model ensemble that exceeds the spread in observational data sets (that is, the ensemble is over-dispersive), while others display a spread that is slightly narrower than observational ranges (that is, the ensemble is under-dispersive).

We explore the properties of ensembles generated using three common techniques: an initial conditions ensemble, a perturbed parameters ensemble, and a perturbed structure ensemble. The latter is intended to represent a simplified multi-model ensemble. We include a 122-member CMIP5 Historical ensemble in our analyses for comparison, noting that despite the different types of uncertainty sampled by CMIP5, all simulations are typically considered to be comparable statistically. We analyse the variance of these ensembles to show that the different generation techniques produce ensembles with vastly different results.

2 Methodology

2.1 Observed data and modelling strategy

We use HadCRUT3 (Brohan et al 2006) observational surface air temperature data from 1971 to 2010. All model simulations, including those from CMIP5, were re-gridded to the HadCRUT3 5° × 5° grid, using area weighted averaging. Only grid cells with more than 80 % of HadCRUT3 anomaly data in the period 1971–2010 were included in the analysis.

We use the CSIRO Mk3L climate system model version 1.2 (Phipps et al 2011, 2012, 2013) to generate three types of ensemble, each comprising 20–25 members and covering the period 1971–2010. CSIRO Mk3L is a fully-coupled general circulation model, incorporating components which describe the atmosphere, land surface, sea ice and ocean. The atmospheric component has a horizontal resolution of 5.6° × 3.2° with 18 vertical levels, while the oceanic component has a horizontal resolution of 2.8° × 1.6° with 21 vertical levels. Although CSIRO Mk3L is primarily designed for millennial-scale climate simulation, it also performs well on shorter time scales, including the twentieth century (Phipps et al. 2012). The computational efficiency of the model allows it to be used here to generate multiple ensembles.

All ensemble members were integrated over the period 1851–2010, following the CMIP5 protocol for the

Historical experiment. The model was driven with the prescribed changes in orbital parameters, atmospheric greenhouse gas concentrations, solar irradiance and stratospheric sulphate aerosols due to volcanic eruptions. The analysis period (1971–2010) allows us to make the optimal use of observational data for the purposes of evaluating each ensemble member. All simulations were based on the default configuration of CSIRO Mk3L (Phipps 2010), with the members of each ensemble being perturbed relative to that baseline.

2.2 Ensemble generation

To generate the initial conditions ensemble, restart files spaced at intervals of 100 model years (sourced from the control simulation used by Phipps et al. 2013), were used to initialise the model.

To generate the perturbed parameters ensemble, six model parameters representing different properties of the land, ocean and atmosphere were selected with the aim of maximising behavioural diversity. In each case, a literature-based scaling factor or value range was chosen and values were sampled within these ranges. They are listed below:

1. Land surface albedo strongly affects the absorption of solar radiation by the land surface and therefore land surface temperature. This parameter was varied by Fischer et al. (2010). We used scaling factors with a range of 0.5–1.5.
2. The aerodynamic roughness length affects surface wind speed, as well as the efficiency of turbulent energy fluxes between the surface and the atmosphere. This parameter was perturbed by Fischer et al. (2010) and by Murphy et al. (2004). We used scaling factors with a range of 0.5–1.5.
3. Soil field capacity limits soil moisture, which affects evaporation and can have a significant impact on global temperatures (Ducharne and Laval 2000). We used scaling factors with a range of 0.9–1.5.
4. Ocean diffusivity controls the rate of heat diffusion within the ocean and was previously examined by Washington and Meehl (1989). We perturbed this parameter within a range of 400–800 m^2s^{-1}.
5. The critical relative humidity threshold for cloud formation (RH_{crit}) defines the humidity level above which clouds appear in the model. It has been used in large perturbed parameter ensembles, such as Murphy et al. (2004). We used a range of 0.65–0.75 for RH_{crit} over land, and 0.75–0.95 for RH_{crit} over the ocean, varied together.
6. The cloud albedo reduction factor accounts for the texture of clouds at the sub-grid level. We used values in the range 0.495–0.695 for convective cloud, and 0.765–0.965 for non-convective cloud, varied together.

Ideally, the generation of the perturbed structure ensemble should involve the comparison of single model simulations from multiple climate models, each using comparable initial conditions and perturbed parameters. Given that this would require coordination and commitment on the scale of a CMIP experiment, we instead constructed a structural perturbation ensemble using CSIRO Mk3L by selectively enabling or disabling individual model components, as described by Phipps (2010).

1. We varied the sea ice model, which has four states: no ice thermodynamics or dynamics, ice thermodynamics only, ice thermodynamics with leads, and full ice thermodynamics and dynamics.
2. We also varied between two alternative schemes for each of the following: atmospheric boundary layer, gravity wave drag, cumuliform cloud formation, stratiform cloud formation, the land surface, and the oceanic equation of state.

Although the relationship between model simulations produced in this way and a true multi-model ensemble is not entirely clear, we should still expect to see more behavioural diversity than in the perturbed parameters ensemble. The experiment used several combinations of each of these options, including the default (all on), default with each option changed individually (9 simulations), and a further 15 simulations with pseudo-random sampling of model structure. We note that 5 simulations became numerically unstable and failed to complete. These are omitted from further analysis.

To generate members of the perturbed parameters and perturbed structure ensembles, a low-discrepancy sequence, the Sobol' sequence (Reichert et al. 2002), was used to sample values from a uniform distribution over the intervals described above. This involved calculating an m-dimensional (here $m = 6$—the number of parameters) Sobol' sequence, of length n (the number of samples—simulations—we wanted to generate). The elements in the Sobol' sequence are m-dimensional vectors, $\mathbf{s}_n$, with each component a quasi-random value between 0 and 1. We can then take each of these samples, and map each component to the interval required for each model parameter—i.e. we take the first element to correspond to the ocean diffusivity values, which we want to map to [400, 800], and then we use $\textit{ocean diffusivity} = (800 - 400) \times \mathbf{s}_{n,k} + 400$. For the perturbed structure ensemble, a discretised version of this process is used: real-valued samples were generated using the Sobol' sequence, in the space $(0,1)^m$, and then each interval was split into p_i intervals, where $p_i =$ number of states in dimension i. The 7-dimensional Sobol' sequences

is then mapped from $(0, 1)^7 \rightarrow \{0, 1, 2, 3\} \times \{0, 1\}^6$ (4 sea-ice model switch states, 2 for each other structural switch). For example:

$$(0.83, 0.23, 0.35, 0.58, 0.92, 0.29, 0.61) \rightarrow (3, 0, 0, 1, 1, 0, 1)$$

is equivalent to: sea ice with full ice thermodynamics and dynamics; alternative version of atmospheric boundary layer scheme; alternative version of land surface scheme; default version of stratiform cloud scheme; default version of cumuliform cloud scheme; alternative version of oceanic equation of state; default version of gravity wave drag scheme.

Ensembles were analysed first using raw output and then using output corrected for any bias in the global mean. Cost functions of surface air temperature were calculated over all grid cells for which HadCRUT3 had $\geq$80 % data coverage for the period 1971–2010. In addition to the three ensembles generated here, analyses were also performed on 122 simulations from the CMIP5 Historical ensemble (WCRP 2013; Pirani 2008) for comparison. For this collection of CMIP5 models, the ensemble only extends to 2004, and so we only use the data from 1971 to 2004.

2.3 Analysis metrics

The analysis metrics we investigate include rank histograms (Hamill 2001; Annan and Hargreaves 2010), and a new approach—observation error curves—motivated and explained in more detail after initial results are described below. Both are tools to assess whether observations are drawn from the same distribution as ensemble members. Rank histograms compare the observations to the ensembles by calculating the rank of the observations relative to the models at each point in time and space. For example, if the observed value for a particular grid point and time-step is higher than that for all the models, it will be ranked $(n + 1)$th, where n is the number of models. If the observations and ensemble data are drawn from the same distribution, the distribution of ranks should be approximately uniform, as the observations have an equally likely chance of lying at any given rank, for any given data point. If the ensemble is over-dispersive (i.e. the spread in the model ensemble is greater than that in the observational data set) then the distribution will be higher in the middle of the range, as the observations are less likely to fall in the extreme high or low ranks. On the other hand, if the ensemble is under-dispersive, the distribution will be U-shaped, as the observations fall outside the model range more often. Note that this would not be true with strongly biased data: if the ensemble contained a significant bias, we would expect to see the observations rank consistently high or low. As we are using bias corrected data, we should expect that the ranks are relatively symmetrically distributed over the interval—that is, the integral of the density function over the lower half of the ranks should be approximately equal to that over the upper half.

As well as applying rank histograms to actual values of temperatures, we also apply them to the temperature trends throughout 1971–2010. While a flat rank histogram of actual temperatures gives an indication that ensemble spread is a reasonable representation of system uncertainty, an over-dispersive ensemble might indicate at least two quite different possibilities. First, that the ensemble trend broadly matches observations, but its spread is too large. Second, that the ensemble trend is in the wrong direction, regardless of ensemble width. It is this second possibility that we aim to detect by constructing a histogram of trends. More specifically, we first estimate the gradient of the least squares fit to the temperature time series at each grid cell for observations and all models within a given ensemble, and then construct a rank histogram based on the values of the gradient. Note that temperature time series are corrected for any bias in the global mean, but are not corrected for any bias in the trend. We would not therefore necessarily expect classic U-shaped or humped histograms of gradient values.

For reference, we also include more common measures of performance, such as the probability density function (PDF) overlap of Perkins et al. (2007), and compare this with temporally sensitive measures such as root mean square error (RMSE) and correlation, calculated over all monthly time steps and $5° \times 5°$ grid cells. The bin width used for the PDF overlap metric was 1 °C.

3 Results

The global annual average temperature for each model ensemble is shown without bias correction in the top row of Fig. 1. Each ensemble performs reasonably well and the ensemble means follow a similar path, as expected given that they share the same time evolution of external forcings. Increasing concentrations of anthropogenic greenhouse gases cause the overall warming trend, while volcanic eruptions cause short-term cooling in the early 1980s (El Chichón, 1982) and early 1990s (Pinatubo, 1991).

There are, however, major differences between the model ensembles. Figure 1 highlights the ensemble biases and variance differences. The mean surface air temperature bias (taken over all grid cells and time steps) for each ensemble is also shown in the first column of Table 1. The initial conditions ensemble mean shows a small positive bias relative to the observations, and all individual members are warmer than the observations. The perturbed

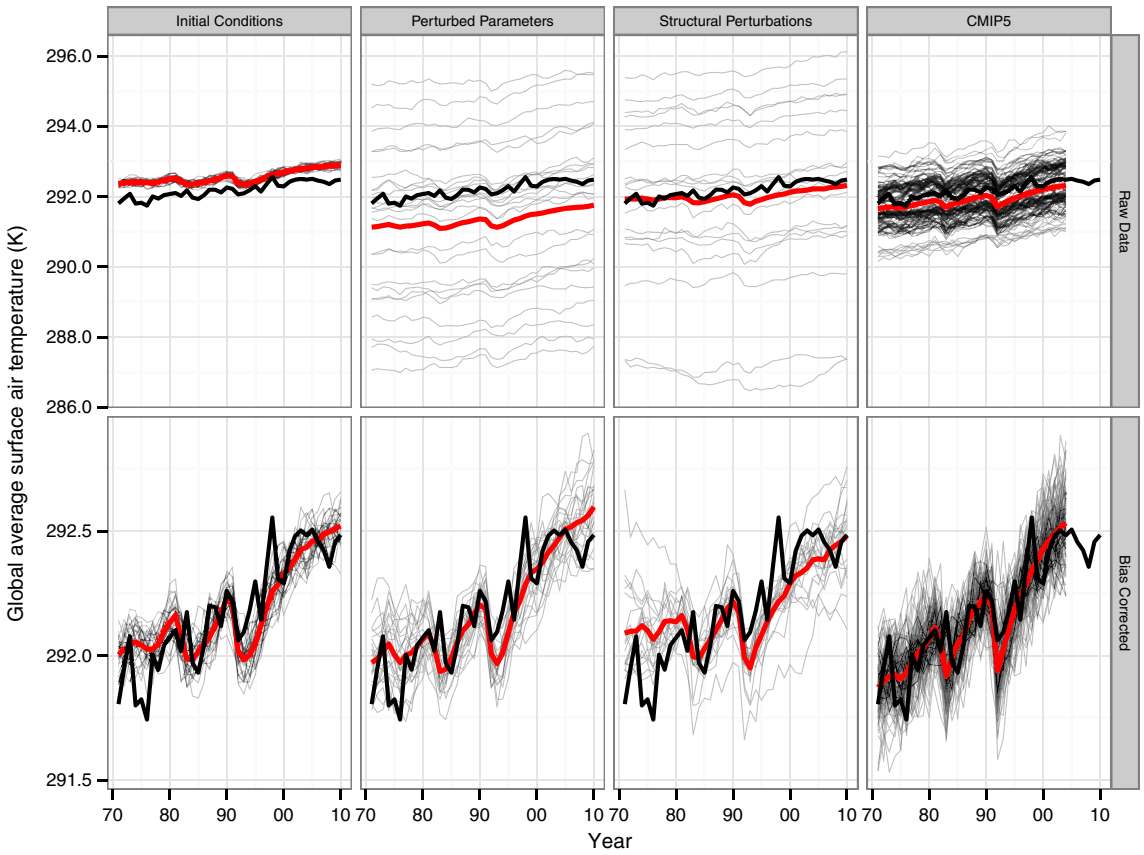

Fig. 1 Global annual mean surface air temperatures (K) for each model simulation. The *columns* represent the three generation techniques, plus the CMIP5 ensemble. The *first row* shows the raw model output, the second row shows bias corrected output (note different y-axis scales). *Thin black lines* represent individual model simulations, the mean is in *red*, and the observations are shown as a *thick black line*

Table 1 Ensemble statistics for surface air temperature (SAT)

Ensemble	Global time-mean SAT (K)	Mean SD of SAT (K)				Standard deviation of *a* values	
		Global		Per-cell		Global	Per-cell
		Raw	BC	Raw	BC		
Initial conditions	292.6	0.099	0.095	0.783	0.782	1.759	2.591
Perturbed parameters	291.3	2.207	0.156	2.615	1.393	1.183	1.592
Perturbed structure	292.0	2.362	0.412	3.575	2.828	0.657	0.945
CMIP5	291.9	0.650	0.183	1.73	1.602	0.873	0.968

The first column shows the average global mean for each ensemble for the entire period. Columns two to five show the mean value of the standard deviation taken across the ensemble at each data point, for both raw and bias corrected (BC) data, over the global and per-cell domains. The last two columns show the standard deviations of *a* values per ensemble, for bias corrected data over the global and per-cell domain (see explanation in text)

parameters ensemble mean has a larger, negative bias, but is less than 1K too cool, while the perturbed structure ensemble mean is closer to the observations. The CMIP5 ensemble performs similarly to our ensembles, with only a small bias in the ensemble mean.

The mean of the standard deviation of the models' global temperature errors at each time step is given for the raw data in columns 2 and 4 of Table 1. This is calculated by taking the standard deviation across the ensemble at each time-step and grid cell (for the per-cell calculation), and then taking the mean of these standard deviation values over time and space. In the global case, the standard deviation of global temperature values is taken across the ensemble for each time step, and then averaged over all

time steps. The variance between the models in the initial conditions ensemble is much lower than in both the perturbed parameters and perturbed structure ensembles. In the latter ensembles, the spread of the individual simulations is large, and each includes simulations that are both positively and negatively biased. The variance in the perturbed structure ensemble is slightly larger than in the perturbed parameters ensemble. The CMIP5 ensemble variance is much higher than the initial conditions ensemble, but much lower than our other ensembles.

We then bias correct individual model simulations by removing the difference between the global time mean of each simulation and the observations (shown in the bottom row of Fig. 1). This process is standard practice in climate change experiments (e.g. Solomon et al. 2007; Macadam et al 2010). After bias correction the inter-model variance (shown in columns 3 and 5 of Table 1) in the perturbed parameters and perturbed structure ensembles is greatly reduced. This suggests that much of the variance in the first three non-bias-corrected ensembles stems from divergence over the 120 years prior to the sample period. However, even after bias correction, there is still more ensemble variance in the perturbed parameter and perturbed structure ensembles. The perturbed structure ensemble has by far the greatest diversity of simulation behaviour, followed by the perturbed parameters ensemble, with the initial conditions ensemble having very little apparent behavioural diversity. This shows that there is an increase in diversity of simulation behaviour as the model is perturbed in increasingly complex ways—initial conditions provide little diversity, while structural changes provide the most. Bias correction has less impact on the CMIP5 ensemble in the per-cell case, perhaps due to model selection (whereby model developers self-select model variants with the best performance) during the development and submission process.

Table 2 shows the performance differences between each ensemble using a range of common metrics. Results for each metric are presented for both the ensemble mean and the average of the metric across all ensemble members. While we might anticipate that metrics that incorporate a temporal signal on a monthly time scale (such as RMSE and correlation) might randomly advantage those simulations that have coincident internal variability with the real world, Table 2 shows that in fact the performance rank of these ensembles is the same in these metrics as in the correlation-insensitive PDF overlap metric. These statistics also reinforce the ability of the CSIRO Mk3L model. The initial conditions ensemble members have on average better RMSE, correlation, and PDF overlap with observations over time and space than the CMIP5 ensemble members, both before and after bias-correction. The perturbed parameter and perturbed structure ensemble members, on the other hand, perform worse on average than the CMIP5 ensemble members under all metrics. In general, as anticipated, the ensemble mean performs better than an individual ensemble member, for each ensemble and each metric.

Spread across each ensemble clearly varies widely. We now focus on what we believe to be a key metric in assessing whether ensemble spread should be used to represent uncertainty. We compare ensemble spread relative to the spread in the observations, using rank histograms and observation error curves, both for the per-cell data set (718,560 data points for all time steps and grid cells), for the global data set (480 time steps of global average temperature), and for per-cell trends (one value for each of the 1,497 grid cells with greater than 80 % data coverage).

Figure 2 shows the ranks of global and per-cell monthly observed values, as well as linear trend gradients over time within each cell, compared to the models in each ensemble. It is clear that the initial conditions ensemble is under-dispersive, underestimating the variance in the observations, as indicated by the U-shaped histogram for both the global and per-cell cases. The perturbed parameter and perturbed structure ensembles clearly overestimate the variance in the observations in the per-cell case, as indicated by the strong bell-shapes in the histograms, although it seems regional diversity compensates somewhat in the case of the perturbed parameters ensemble, as the rank histogram for global mean temperatures is much flatter than for per-cell temperatures. The CMIP5 ensemble appears largely over dispersive in both cases; however, there are distinct up-ticks at both ends of the per-cell histogram, perhaps indicating extreme hot and cold seasons in some regions that are not captured by the spread of models, or potentially indicating errors in the observational data.

The large asymmetry in the perturbed parameters ensemble per-cell rank histogram indicates that the observations are ranking higher than the models more often than should be expected, given bias corrected data. This may be because the model means are based on a distribution that is highly skewed relative to the observations—e.g. the models are too hot in the tropics, where variability is lower, or too cold in more variable grid cells, such as polar regions. Since the ranks are not dealing with actual values, the bias correction does not guarantee a balance of high and low ranks.

The histograms of trend gradients indicate even greater differences in behaviour between the ensemble types. The initial conditions ensemble is again highly under-dispersive, indicating that the temporal trends are too homogeneous across grid cells, and that there are not enough extreme trends. The structural ensemble exhibits a large bias toward low or negative trends (as indicated by the large number of high-ranking observation trends), which is likely largely due to the handful of simulations that have a

Table 2 Simulation statistics per ensemble, calculated per-cell, for raw and bias corrected data

	Ensemble	RMSE		Correlation		PDF overlap	
		mmmean	Mean	mmmean	Mean	mmmean	Mean
Raw	Initial conditions	2.086	2.323	0.986	0.983	0.943	0.945
	Perturbed parameters	2.403	3.569	0.985	0.976	0.947	0.847
	Structural perturbations	2.590	4.878	0.980	0.956	0.936	0.858
	CMIP5	1.687	2.636	0.991	0.979	0.960	0.935
BC	Initial conditions	2.050	2.293	0.986	0.983	0.972	0.973
	Perturbed parameters	2.205	2.819	0.985	0.976	0.927	0.926
	Structural perturbations	2.577	4.297	0.980	0.956	0.919	0.882
	CMIP5	1.672	2.552	0.991	0.979	0.971	0.947

Each statistic is calculated first for the multi-model mean, and then for the average of the statistic for each simulation in the ensemble. The three statistics calculated are root mean squared error, correlation, and the PDF overlap score from Perkins et al. (2007)

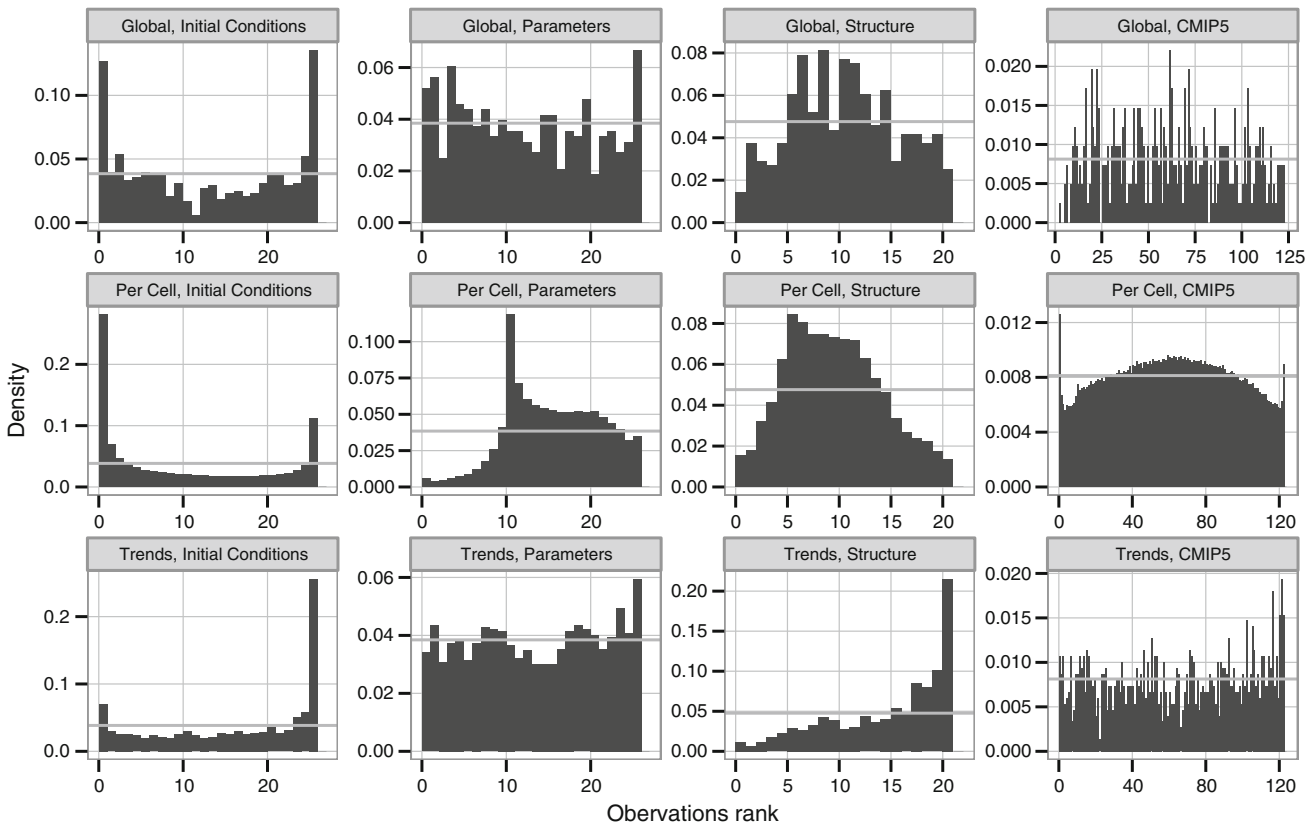

Fig. 2 Rank of observations relative to bias-corrected model simulations in each ensemble. The *columns* represent the three generation techniques, plus the CMIP5 ensemble. The histograms in the *top row*are based on global monthly mean surface air temperatures, while those in the *middle row* are based on per-cell data. The histograms in the *third row* are based on the linear trend gradients in surface air temperature at each grid cell. The *grey lines* indicate the expected values for a perfectly uniform distribution. High observation ranks for a particular metric indicate that the model ensemble more often exhibits lower values for that metric. Likewise, a narrow distribution for the observation ranks indicates that the model ensemble has a wider distribution

low or negative global trend over the period. The histograms for both the perturbed parameters ensemble and the CMIP5 ensemble are quite flat, indicating reasonably good performance of trends across the spatial domain. This is notable for the perturbed parameter ensemble especially, as there are clear major trends in the spin-up period 1851–1970 (not shown in Fig. 1, but all simulations in that ensemble used the same initial conditions).

Another method of examining the over- or under-dispersion of the ensemble is to assume that the ensemble

accurately represents the distribution of uncertainty, as per the indistinguishable paradigm (Annan and Hargreaves 2010), and examine whether the observations are drawn from this distribution. We let the observations be a linear combination of the ensemble mean and the ensemble standard deviation, scaled by a random variable, a, at each grid-cell and time step:

$$obs_i = \bar{x_i} + a_i \sigma_{x_i}$$

where i indexes both the grid-cell and time step. Inverting this equation, we obtain

$$a_i = \frac{(obs_i - \bar{x_i})}{\sigma_{x_i}}$$

If the observations are drawn from the same distribution as the ensemble, the distribution of the random variable a should approximate the standard normal distribution ($\mathcal{N}(0,1)$). Note however that the distribution is not necessarily zero-centred in the per-cell case, because bias correction is performed globally. The variance of a is inversely related to the over- or under-dispersion of the ensemble, relative to the observations.

The standard deviation of a for each ensemble for global and per-cell data is shown in the last two columns of Table 1. This gives us a useful quantitative measure of over- or under-dispersion based on continuous values rather than rank, and shows that the initial conditions and perturbed parameter ensembles are both under-dispersive ($\sigma_a > 1$), while the perturbed structure ensemble is over-dispersive ($\sigma_a < 1$), in both the global and per-cell cases. The distribution of a for each ensemble is shown in Fig. 3. We can see that both the perturbed structure ensemble and the CMIP5 ensemble are over-dispersive, whereas the initial conditions and perturbed parameter ensemble both appear to be under-dispersive, regardless of whether the analysis is performed globally or per-cell. This result appears to contradict the apparent over-dispersion displayed in the per-cell rank histogram of the perturbed parameters ensemble (see Fig. 2). This discrepancy may be due, for example, to a long tail of cold temperatures over the globe being eliminated in the conversion from continuous temperature data to ordinal ranks, or to ensembles not being uniformly over- or under-dispersive everywhere. The distributions of the observational trends confirm the findings in the trend rank histograms—the initial conditions ensemble is under-dispersive, and both the initial conditions and structural ensembles have a large bias toward lower trends, relative to the observations. This low bias in

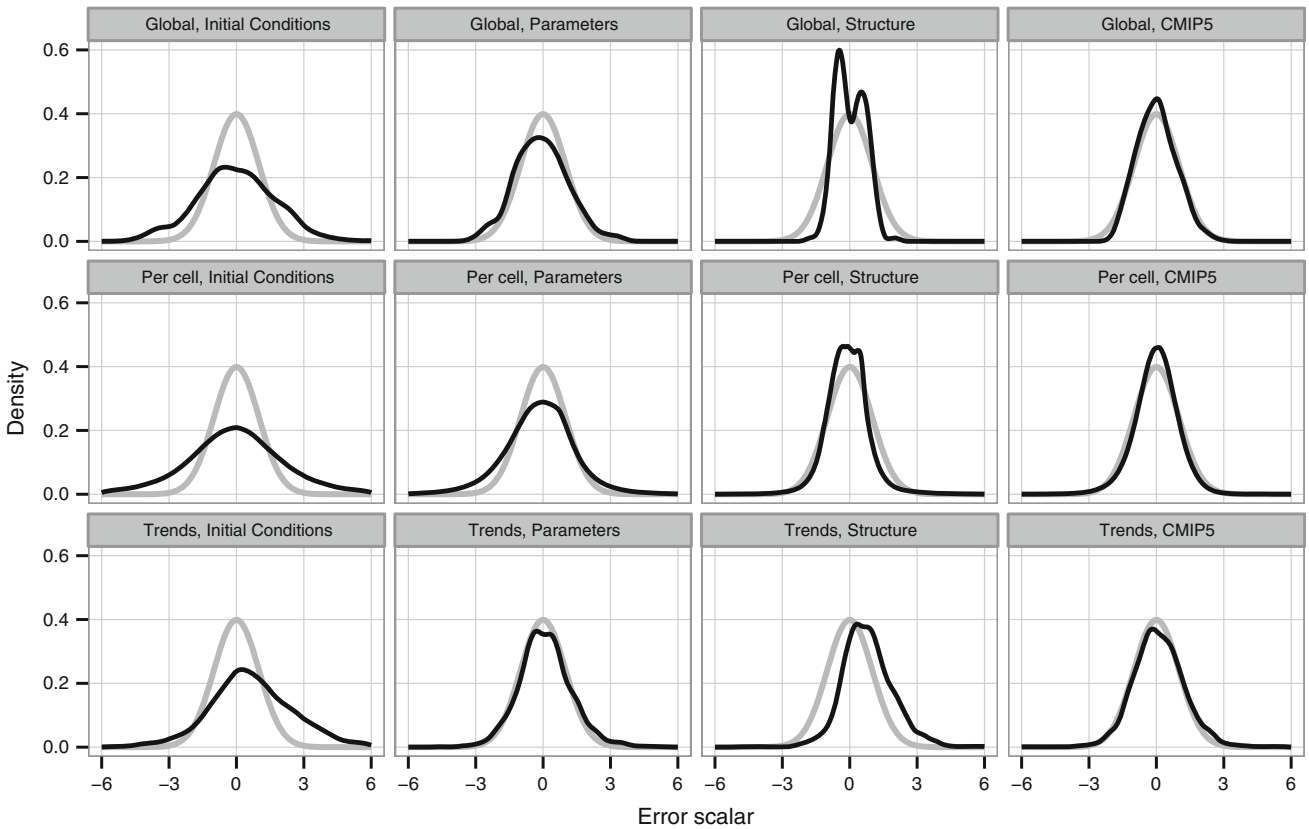

Fig. 3 Distributions of observations relative to the ensemble means on a per-cell basis, normalised by the standard deviation of the ensemble at each grid point and time-step. The *rows* and *columns* correspond to the same domains as in Fig. 2. A standard normal distribution ($\mathcal{N}(0,1)$) is shown in grey, for reference

trends is also visible in the rank histogram equivalents for these two ensembles.

4 Discussion

There are two key results here. Firstly, and most obviously, the different ensemble generation techniques lead to very different ensembles. Perturbed structure ensembles are over dispersive and perturbed initial conditions ensembles are highly under dispersive, at least for the modelling system used in this study. If one were to naïvely use spread from these ensembles to estimate uncertainty in the climate system they would likely over and underestimate it, respectively. The results for the perturbed parameters ensemble are less certain, with seemingly contradictory results for the per-cell rank histogram and observation error distributions.

Our analysis of CMIP5 shows it to be over dispersive for surface temperature, and this is supported by the analysis conducted by Annan and Hargreaves (2010), which shows that the CMIP3 ensemble is over-dispersive, both for sea level pressure and surface air temperature. The reasons why CMIP5 is not *more* over dispersive than our structural ensemble, despite it being a multi-model ensemble, are not immediately clear. The fact that CMIP5 is an ensemble that is part perturbed structure, part perturbed parameter and part perturbed initial condition should not change our expectation of greater dispersiveness. The answer may at least in part be due to model selection in the CMIP5 submission process, discussed below.

Second, while we acknowledge that different perturbation approaches (initial conditions, parameter, structural perturbation) target very different types of uncertainty in the modelling system, these do not directly translate to the climate system in a meaningful way. For example, initial conditions ensembles aim to sample internal variability within the modelling system (Phipps et al. 2013). Different members aim to represent the spread of possible states of the climate system, given that the initial conditions are only imprecisely known, if at all. This is essentially the same as attempting to generate replicate Earths, since if we had a perfect model, replicate Earths might be generated by sampling initial condition uncertainty. However, the under-dispersion in the perturbed initial conditions ensemble shown here clearly illustrates that it does not represent a replicate Earth ensemble. Conversely, perturbed physical parameter and perturbed structure ensembles are designed to sample the uncertainty in our knowledge of real physical systems. These also aim to sample uncertainty in the approach taken to approximate those processes on spatial scales larger than those on which the processes themselves operate.

It could be argued that the separation of uncertainties in initial conditions, model parameters and model structure, while conceptually useful, might be hindering accurate prediction. For example, consider the case where an under-dispersive perturbed initial conditions ensemble is used to determine the magnitude of future climate variability or uncertainty in predictions. Our results provide good reason to believe that this approach will lead to underestimates of these sources of uncertainty. If, however, we can create an ensemble that includes perturbed parameter and perturbed structure members and that generates a flat rank histogram, it is more likely to give reliable results, despite its apparent mixing of uncertainty domains. We did not investigate whether different parameter values or structures of the model would lead to significantly different variances resulting from perturbations to the initial conditions. However, we have no reason to believe that this would necessarily be the case. We suggest that, rather than perturbing ensembles using initial conditions, parameter variations or structural variation in the hope that these lead to appropriate estimation of uncertainty, researchers should instead prove that variability generates flat rank histograms before using that ensemble to estimate variability or uncertainty in climate projections.

Our results apply only to surface air temperature and only provide a basic estimation of ensemble spread. We note that this study does not aim to explore ensemble spread alone, but rather that our goal was to explore the differences between ensembles generated by different perturbation techniques. We have only examined surface temperature here, because the surface temperature record is the most reliable observational record for the last four decades. While other variables may exhibit different results, Yokohata et al. (2012; Fig. 1) indicate that at least to a first approximation, ensembles that are over- or under-dispersive for surface temperature are likely to be correspondingly over- or under-dispersive for other variables.

The implications of our results for the CMIP "ensemble of opportunity" approach, the most commonly-used approach to assess uncertainty in future climate projections, are considerable. We have shown that even a limited implementation of a perturbed structure approach within a single model is sufficient to produce an over-dispersive ensemble. It might therefore be expected that a true multi-model ensemble approach, which would include a greater diversity of model structures, would also generate over-dispersive ensembles. However, this only applies strictly to a multi-model ensemble with one simulation from each model, and with no model selection. Model selection is inherent in CMIP3 and CMIP5. While the CMIP ensembles are constructed as an "ensemble of opportunity", the ensemble members themselves are not entirely the product of opportunity: modelling groups choose which simulations

to submit, inevitably resulting in an ensemble of *good* model simulations, each of which perform well individually. This may reduce the over-dispersion exhibited in the ensembles. Nevertheless, Annan and Hargreaves (2010) showed that the CMIP3 ensemble is over-dispersive, and our results indicate the CMIP5 ensemble is over-dispersive in surface air temperature, although this may not always be the case (Oldenborgh et al 2013). Care should therefore be taken when using variance as an uncertainty estimate from these ensembles.

If our results are generalisable, initial conditions perturbation alone is likely not an adequate approach for ensemble generation, if ensemble spread is to be used as an estimate of uncertainty. This is only true under the replicate Earth (Bishop and Abramowitz 2013) or indistinguishable paradigms (Annan and Hargreaves 2010), where we expect some variance about the mean, and where such variance in the ensemble members should be statistically indistinguishable from that in the observations. Under the truth-plus-error paradigm (Knutti et al 2010), the statement makes no sense, as model-observation errors are assumed to be independent, and as such there is no "internal variability" within the observations to speak of: if enough models are included, errors will average out, and we will be left with an accurate representation of the "truth", i.e. the observations. We believe that this makes the truth-plus-error paradigm unsupportable, and inconsistent with any concept of "internal variability".

While we cannot definitively generalise this result to other models, we see no reason why it should not apply widely. The lack of analyses of ensemble variance within climate model evaluation literature, for example using rank histograms or something akin to observation error curves, gives us no cause to discount this as a reasonable assumption. We suggest that greater use of rank histograms or observation error curves (e.g. Annan and Hargreaves (2010; Bishop and Abramowitz 2013; Oldenborgh et al. 2013) would be valuable. We also note that there may be reasons why CSIRO Mk3L's internal variability might be lower than other models—its relatively coarse spatial resolution for example. However, evaluation of the internal climate variability simulated by CSIRO Mk3L on inter-annual to interdecadal time scales indicates that its characteristics are comparable both with reconstructions of past climate and with that simulated by other models (Fernández-Donado et al. 2013; Phipps et al. 2013). We are not aware of any *a priori* reason why a model that has a higher resolution should necessarily exhibit greater behavioural diversity under any particular physics parameterization, but this issue does need to be explored in greater detail in the future.

It is difficult to know how the results for our perturbed parameters ensemble would compare with other perturbed parameter ensemble experiments, as the combination of parameters and their ranges are clearly specific to the model in question. It is obviously possible to create a perturbed parameter ensemble that is under-dispersive: simply choose a set of parameters that are known to have only small impacts, or only perturb the parameters over small ranges. Conversely, it should be possible to create an over-dispersive ensemble by choosing non-physical parameter ranges. Yokohata et al. (2012) lists a number of perturbed parameter experiments that show over- or under-dispersive climate sensitivity relative to the CMIP3 ensemble. We did not seek to generate a particular type of result, but rather to generate a perturbed parameter ensemble with realistic, literature-driven perturbations. The types of analyses we present could be valuable for examining the outcomes of the large perturbed parameters ensemble experiments typically used for making projections—especially for defining reasonable default parameter ranges, as changes in parameter ranges can be related to changes in ensemble variance.

At the other end of the spectrum it is unlikely that the perturbed structure ensemble presented here is representative of a true multi-model perturbed structure ensemble—we have perturbed only a small sub-set of the CSIRO Mk3L model structural components. While it is true that many modern GCMs share components (at least the theoretical background, if not the numerical implementation), these models have many more components than the seven that we have perturbed here. None of the CSIRO Mk3L model switches were indefensible—the alternative schemes are all plausible representations of aspects of the climate system. It would therefore seem intuitive that a true multi-model perturbed structure ensemble is likely to have broader spread than our perturbed structure ensemble does.

5 Conclusions

We have used a climate model, perturbed by initial conditions, physical parameters and structural choices, to generate ensembles of simulations. We have shown that there are significant differences between ensembles generated in these different ways. In particular, perturbed structure ensembles are likely to overestimate internal climate variability. This is therefore likely to be the case as well in true multi-model ensembles such as CMIP5, where there is a larger diversity of model components within the ensemble. In contrast, we have shown that ensembles generated by perturbing initial conditions tend to have excessively narrow spread, underestimating variability. There appears to be no obvious reason to believe that this conclusion would be significantly different for other

models, despite known differences in models' internal variability.

If, therefore, we wish to create ensembles that are unlikely to over or underestimate variability or uncertainty, our results highlight that we may need to create ensembles that are part initial conditions perturbation and part parameter or structural perturbations. While these are traditionally very separate sources of uncertainty within a modelling system, it appears that a careful blending of perturbation approaches to achieve optimal dispersion may be the best approach to generate reliable results. In shorter term experiments, such as seasonal prediction, where internal variability can be dominant, the nature of the blending process may be different to longer term climate prediction, where model structure and parameter values might play a different role in ensemble dispersiveness.

While an analysis over an in-sample period is never necessarily representative of how the system would perform in the future, we suggest that this approach applies equally to ensembles of opportunity (such as the CMIP ensembles) and grand ensembles (mixed initial conditions and perturbed parameter ensembles, such as the climate*Prediction*.net experiments, Stainforth et al. 2005) and warrants further investigation. Perhaps the best that we can hope for in the near term is the development of a subsampling technique that might create smaller optimal ensembles from the collection submitted to CMIP experiments. A key step in this process, however, is an understanding of what is offered by each approach to ensemble generation. Our results, by isolating the effects of different ensemble generation techniques, contribute to the goal of understanding how best to generate unbiased, well-distributed climate model ensembles.

We firstly acknowledge the insightful and constructive comments of a number of anonymous reviewers. This work was supported in part through the ARC Centre of Excellence in Climate System Science, which is supported by the Australian Commonwealth Government (CE110001028) and by ARC Discovery Grant DP110102618. 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 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. All CMIP5 data was accessed via the NCI ESGF node. This research was undertaken with the assistance of resources provided at the Australian National University through the National Computational Merit Allocation Scheme supported by the Australian Government.

References

Annan JD, Hargreaves JC (2010) Reliability of the CMIP3 ensemble. Geophys Res Lett 37:5. doi:201010.1029/2009GL041994. http://www.agu.org/pubs/crossref/2010/2009GL041994.shtml

Bishop CH, Abramowitz G (2013) Climate model dependence and the replicate earth paradigm. Clim Dyn 41(3-4):885–900. doi:10.1007/s00382-012-1610-y

Brohan P, Kennedy JJ, Harris I, Tett SFB, Jones PD (2006) Uncertainty estimates in regional and global observed temperature changes: a new data set from 1850. J Geophys Res 111(D12). doi:10.1029/2005JD006548. http://www.agu.org/pubs/crossref/2006/2005JD006548.shtml

Ducharne A, Laval K (2000) Influence of the realistic description of soil water-holding capacity on the global water cycle in a GCM. J Clim 13(24):4393–4413. doi:10.1175/1520-0442(2000)013<4393:IOTRDO>2.0.CO;2. http://journals.ametsoc.org/doi/abs/10.1175/1520-0442(2000)013<4393%3AIOTRDO>2.0.CO%3B2

Fernández-Donado L, González-Rouco JF, Raible CC, Ammann CM, Barriopedro D, García-Bustamante E, Jungclaus JH, Lorenz SJ, Luterbacher J, Phipps SJ, Servonnat J, Swingedouw D, Tett SFB, Wagner S, Yiou P, Zorita E (2013) Large-scale temperature response to external forcing in simulations and reconstructions of the last millennium. Clim Past 9(1):393–421. doi:10.5194/cp-9-393-2013. http://www.clim-past.net/9/393/2013/

Fischer EM, Lawrence DM, Sanderson BM (2010) Quantifying uncertainties in projections of extremesa perturbed land surface parameter experiment. Clim Dyn 37(7-8):1381–1398. doi:10.1007/s00382-010-0915-y,

Hamill TM (2001) Interpretation of rank histograms for verifying ensemble forecasts. Mon Weather Rev 129(3):550–560. doi:10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2

Knutti R, Abramowitz G, Collins M, Eyering V, Gleckler PJ, Hewitson B, Mearns LO (2010) Good practice guidance paper on assessing and combining multi model climate projections. In: Stocker TF, Qin D, Plattner GK, Tignor M, Midgley GF (eds) Meeting report of the intergovernmental panel on climate change expert meeting on assessing and combining multi model climate projections. IPCC Working Group I Technical Support Unit. University of Bern, Bern, Switzerland

Macadam I, Pitman AJ, Whetton PH, Abramowitz G (2010) Ranking climate models by performance using actual values and anomalies: implications for climate change impact assessments. Geophys Res Lett 37(16). doi:10.1029/2010GL043877. http://www.agu.org/pubs/crossref/2010/2010GL043877.shtml

Meehl GA, Stocker TF, Collins WD, Friedlingstein P, Gaye AT, Gregory JM, Kitoh A, Knutti R, Murphy JM, Noda A, Raper SC, Watterson IG, Weaver AJ, Zhao ZC (2007) Global climate projections. In: Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) Climate change 2007: the physical science basis. Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge, pp 747–845

Murphy J, Sexton D, Barnett D, Jones G, Webb M, Collins M, Stainforth D (2004) Quantification of modelling uncertainties in a large ensemble of climate change simulations. Nature 430(7001):768–772. http://www.nature.com/nature/journal/v430/n7001/abs/nature02771.html

Oldenborgh GJv, Reyes FJD, Drijfhout SS, Hawkins E (2013) Reliability of regional climate model trends. Environ Res Lett 8:014-055. doi:10.1088/1748-9326/8/1/014055. http://iopscience.iop.org/1748-9326/8/1/014055

Perkins SE, Pitman AJ, Holbrook NJ, McAneney J (2007) Evaluation of the AR4 climate models simulated daily maximum temperature, minimum temperature, and precipitation over australia

using probability density functions. J Clim 20(17):4356–4376. doi:10.1175/JCLI4253.1

Phipps SJ (2010) The CSIRO Mk3L climate system model v1.2, Technical report no. 4. The Antarctic Climate & Ecosystems CRC, Hobart, Tasmania, Australia

Phipps SJ, Rotstayn LD, Gordon HB, Roberts JL, Hirst AC, Budd WF (2011) The CSIRO Mk3L climate system model version 1.0—part 1: description and evaluation. Geosci Model Dev 4(2):483–509. doi:10.5194/gmd-4-483-2011. http://www.geosci-model-dev.net/4/483/2011/

Phipps SJ, Rotstayn LD, Gordon HB, Roberts JL, Hirst AC, Budd WF (2012) The CSIRO Mk3L climate system model version 1.0—part 2: response to external forcings. Geosci Model Dev 5(3):649–682. doi:10.5194/gmd-5-649-2012. http://www.geosci-model-dev.net/5/649/2012/

Phipps SJ, McGregor HV, Gergis J, Gallant AJE, Neukom R, Stevenson S, Ackerley D, Brown JR, Fischer MJ, van Ommen TD (2013) Paleoclimate data-model comparison and the role of climate forcings over the past 1500 years. J Clim doi:10.1175/JCLI-D-12-00108.1

Pirani A (ed) (2008) WCRP Coupled Model Intercomparison Project—Phase 5, CLIVAR echanges, vol 16. National Oceanography Centre, Southampton, UK. http://cedadocs.badc.rl.ac.uk/310/1/feb2008.pdf

Randall D, Wood R, Bony S, Colman R, Fichefet T, Fyfe J, Kattsov V, Pitman A, Shukla J, Srinivasan J, Stouffer R, Sumi A, Taylor K (2007) Climate models and their evaluation. In: Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt K, Tignor M, Miller H (eds) Climate change 2007: the physical science basis. Contribution of working group I to the fourth assessment report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge

Reichert P, Schervish M, Small MJ (2002) An efficient sampling technique for bayesian inference with computationally demanding models. Technometrics 44(4):318–327. doi:10.1198/004017002188618518

Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt K, Tignor M (eds) (2007) Contribution of working group I to the fourth assessment report of the Intergovernmental Panel on Climate Change. IPCC Fourth Assessment Report: climate change 2007. Cambridge University Press, Cambridge. http://www.ipcc.ch/publications_and_data/publications_ipcc_fourth_assessment_report_wg1_report_the_physical_science_basis.htm

Stainforth D, Aina T, Christensen C, Collins M, Faull N, Frame D, Kettleborough J, Knight S, Martin A, Murphy J, et al (2005) Uncertainty in predictions of the climate response to rising levels of greenhouse gases. Nature 433(7024):403–406. doi:10.1038/nature03301. http://www.nature.com/nature/journal/v433/n7024/abs/nature03301.html

Tebaldi C, Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. Philos T Roy Soc A 365(1857):2053–2075

Washington WM, Meehl GA (1989) Climate sensitivity due to increased CO2 experiments with a coupled atmosphere and ocean general circulation model. Clim Dyn 4(1):1–38. doi:10.1007/BF00207397. http://www.springerlink.com/content/k13643612510589x/abstract/

WCRP (2013) CMIP5: overview. http://cmip.llnl.gov/cmip5/index.html

Yokohata T, Annan JD, Collins M, Jackson CS, Tobis M, Webb MJ, Hargreaves JC (2012) Reliability of multi-model and structurally different single-model ensembles. Clim Dyn pp 1–18. doi:10.1007/s00382-011-1203-1. http://www.springerlink.com/index/Y51442386248051N.pdf