Clim Dyn (2015) 45:3169–3181
DOI 10.1007/s00382-015-2531-3

Weighting climate model ensembles for mean and variance estimates

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

Received: 17 July 2014 / Accepted: 13 February 2015 / Published online: 20 February 2015

Abstract Projections based on climate model ensembles commonly assume that each individual model simulation is of equal value. When combining simulations to estimate the mean and variance of quantities of interest, they are typically unweighted. Exceptions to this approach usually fall into two categories. First, ensembles may be pared down by removing either poorly performing model simulations or model simulations that are perceived to add little additional information, typically where multiple simulations have come from the same model. Second, weighting methodologies, usually based on model performance differences, may be applied, and lead to some improvement in the projected mean. Here we compare the effect of three different weighting techniques—simple averaging, performance based weighting, and weighting that accounts for model dependence—on three ensembles generated by different approaches to model perturbation. We examine the effect of each weighting technique on both the ensemble mean and variance. For comparison, we also consider the effect on the CMIP5 ensemble. While performance weighting is shown to improve the estimate of the mean, it does not appear to improve estimates of ensemble variance, and may in fact degrade them. In contrast, the model independence weighting approach appears to improve both the ensemble mean and the variance in all ensembles.

1 Introduction

It is common to evaluate climate models against twentieth century data Randall et al. (2007) and Flato et al. (2013). Evaluations of this kind are based on the view that skilful twentieth century hindcasts by climate models are an important basis for judging the value of these models in projecting the climate of the twenty-first century. However, there are many reasons why no single model simulation is an adequate representation of the true Earth's climate. These include: a limited understanding of the physics of the climate system; imperfect numerical schemes to simulate known components of the climate system; limitations in computational resources, which restrict both the completeness and resolution of climate models; a lack of precision in observationally-based initial conditions for simulations; and a lack of understanding of the amplitude and frequency of internal climate system variability, which limits our ability to assess model performance. In collating an ensemble of imperfect model simulations, the community typically assumes that these simulations are drawn from a sufficiently broad pool to constitute an independent sampling strategy (Knutti et al. 2010b). It then uses the distribution across the ensemble to make probabilistic estimates of changes in the climate system (e.g. Tebaldi and Knutti 2007).

In most cases, we actually have more information about the differences between ensemble members. For example, we know that some models appear to perform significantly better than others for some applications (Box 9.1, Flato et al. 2013). To account for this, performance-based weighting methodologies can be used, in which poorly-performing models are assigned lower weights than others (e.g. Gleckler et al. 2008; Reifen and Toumi 2009; Weigel et al. 2010; Tebaldi and Knutti 2007). This is an intuitive

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.

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

solution, and has the advantage over model selection (e.g. only using the best performing models) of taking some information from all models, while still reducing the susceptibility of the mean to apparent outlier-introduced bias.

There are of course many reasons why one model may perform better than another: one may have more accurate parameterisations, higher spatial resolution, be more tightly calibrated to relevant data sets, or include more physical components. Some model simulations may also perform better than others because of more accurate initialisation, or because of the imposition of more complete or more accurate external forcings. In all these cases there is an expectation that the advantages afforded by weighting such a model above others will persist throughout the twenty-first century, justifying the construction of weights using twentieth century observed data. A model might also perform well, however, because its particular realisation of stochastic internal variability happens to coincide with observations (or worse, observational errors better coincide with the model errors). In these cases, the efficacy of using weights derived on a historical reference period for future projection are clearly questionable. These and other risks associated with model weighting are well recognised (Reifen and Toumi 2009; Macadam et al. 2010; Weigel et al. 2010).

One issue that performance weighting does not address is the potential dependence in sampling strategy that the ensemble could represent. How independent are the future climate estimates that different model simulations provide? Shared parameterisations, initial conditions, or land surface data sets, for example, might mean that simulations from different research groups behave similarly. By reducing the impact of unusual model simulations, performance weighting may actually *increase* inter-model dependence in an ensemble. Consider a case where 5 models in a 6-member ensemble are essentially identical and the sixth does not perform as well as the others. Removing this outlier would falsely give a high degree of confidence in the result where inter-model spread is interpreted as probability. This could result in climate estimates or projections that are biased, and yet appear far more certain than they reasonably should, potentially leading to ill-informed decision making.

We already know that the structures of climate models are somewhat dependent (Masson and Knutti 2011). In our analysis below, we examine how dependence is manifested in model output. That is, how error covariances between different simulations affect the ensemble mean and ensemble variance. To do this, we use the independence-based weighting methodology proposed by Bishop and Abramowitz (2013). To understand how this weighting approach defines model dependence and how we apply it in the context of this work, we first clarify our assumptions about the relationship between an ensemble of model simulations and observations of the real world. Below we contrast three existing paradigms of interpretation.

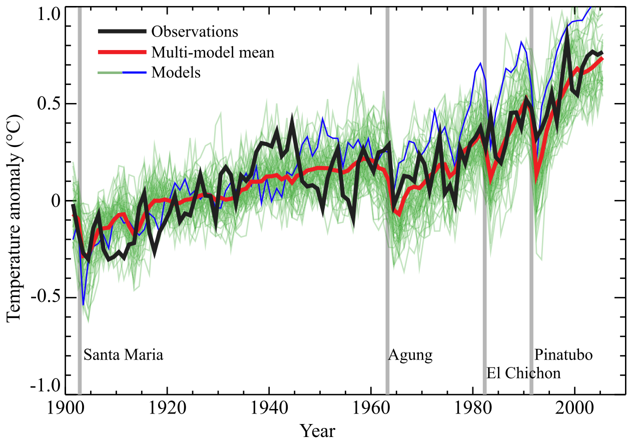

Fig. 1 The CMIP3 model ensemble, after Hegerl et al. (2007). Models are shown in *green*, the multi-model mean in *red*, and the observations in *black*. An arbitrary single model is shown in *blue* to highlight the differences between variability seen in the model simulations and observations and in the multi model mean

The *truth-plus-error* paradigm (Knutti et al. 2010b) views imperfect model simulations as centred around the observations, with model-observation discrepancy essentially a pseudo-random noise that represents flaws in the model, computational inadequacy, or initial condition uncertainty. It is the prevailing approach to interpreting model ensembles (Annan and Hargreaves 2010; Knutti et al. 2010a).

The assumption of random distribution of error in the truth-plus-error paradigm implies that, as an ensemble increases in size, the ensemble mean should converge toward the observations. That is, the error of the multi-model mean should converge to zero as model errors are averaged out, purely as a function of ensemble size. The expected error correlation of independent models in this context is the same as for random variables: zero. This raises the most problematic (yet least stated) implication of the truth-plus-error paradigm—that the climate system is *entirely* deterministic, with no stochastic internal variability. That is, the only barrier to an arbitrarily close match between the ensemble mean and observations is the number of independent models in the ensemble. While some may argue that the climate system is deterministic, for example, on 30-year averages and beyond, we know that it is not on shorter timescales and in fact have evidence of variability on longer timescales (e.g. PAGES 2k Consortium 2013). In reality of course, we only have one climate system—a sample size of one—and so definitively defining a timescale of internal variability is impossible.

It should be evident that the truth-plus-error paradigm is problematic: Fig. 1 shows significant differences between the multi-model mean (red) and the observations (black). In particular, there are long periods where the multi-model mean and the observations are quite distant (e.g. 1935–1945). Except where there are strong volcanic eruptions, the multi-model mean exhibits significantly less variability than the observations. The mean clearly does not converge toward the observations. Similar results were noted by Knutti et al. (2010b). It might be argued that this is evidence of a systematic bias, however that would require us, in this instance, to believe in the existence of a systematic bias that just happened to remove a lot of the variability over time.

Annan and Hargreaves (2010) were the first to explicitly refute and provide an alternative to the truth plus error paradigm. They assert that models and observations should be treated as indistinguishable random draws from the same underlying distribution—the "*statistically indistinguishable paradigm*". This clearly removes the expectation that the ensemble mean and observations should converge as the ensemble size increases. Since the observations have the same characteristics as a model, and contain a certain amount of error, as the model ensemble size increases the ensemble mean should *not* converge to the observations, but to the statistical centre of the distribution. The multi-model mean does not have the same attributes as a true Earth-like climate because the averaging process reduces the amplitude of unforced internal variability (Knutti et al. 2010a). Supporting this, Gleckler et al. (2008) show that the multi-model mean has smaller errors, and that the variance of the mean is lower, than individual models. Critically, the multi-model mean does *not* represent a potentially real climate. The meaning of ensemble spread is not explicitly described in Annan and Hargreaves (2010), but is assigned in a later blog post to "collective uncertainties about how best to represent the climate system" (Annan 2010).

The *replicate earth* paradigm (Bishop and Abramowitz 2013) expands on this idea by asserting that the distribution defined by internal climate system variability provides a basis for the distribution described in the indistinguishable paradigm. That is, an ideal, independent estimate of the climate system's state would be a random draw from a distribution that defined true internal climate system variability. Bishop and Abramowitz (2013) argue, however, that it is not safe to assume, as the indistinguishable paradigm does, that models represent independent draws from this underlying distribution. They argue that if we assume that the Earth's climate is partially chaotic, we can conceptually estimate the spread of possible climate system outcomes by imagining a vast number of replicate earths. All replicate earths would have observationally consistent climate forcing, so that by sampling the different outcomes across replicate earths we could estimate a *Climate Probability Distribution Function* (CPDF) for a variable of interest. In this case, our observational record would represent just one sample from the CPDF, and the instantaneous CPDF would represent the distribution of variable values drawn from the climate system's internal variability. The properties of the CPDF would clearly vary with time. The CPDF mean, a smoothed quantity with lower variability than any individual replicate earth, would represent the forced response of the natural system. Should the climate system be truly deterministic, on timescales longer than 30 years for example, then we would simply expect that the CPDF variance would collapse to zero when analysing 30-year average time series.

Bishop and Abramowitz (2013) argue that climate models are best viewed as *flawed* attempts to create replicate earths. A *perfect*, independent model simulation would be a random draw from the time-evolving CPDF. The mean of an ensemble, therefore, is an approximation to the mean of the CPDF, and the quality of that approximation relies on the independence of the models in the ensemble. An *imperfect* replicate earth ensemble would be an ensemble whose members only represent dependent subsets of the possibilities defined by the CPDF, biasing the ensemble as a whole.

Under the indistinguishable and replicate earth paradigms, model errors should actually be considered as a linear combination of two time-series (i.e. model minus observations), and so the expected correlation will be positive (Bishop and Abramowitz 2013). If the ensemble members are independent, and the time series is long enough, replicate earth-like model simulations (as is assumed in the indistinguishable paradigm) should have an error correlation of 0.5, because each error is effectively a linear combination of two replicate earths: the simulation minus observations.

We note that, regardless of ensemble interpretation paradigm, all weighting approaches are likely to suffer from issues such as sensitivity to the variable being weighted (a simulation in an ensemble might receive quite different weights depending on which variable is examined) and the metric used to construct the weights (some metrics target variability, means or entire distributions; some are sensitive to scale, while others are not). Their efficacy will also always be entirely dependent on the representativeness of the in-sample period used to train the weights with respect to the out-of-sample prediction period. These are clearly not insurmountable difficulties, but define clear prerequisites for the experimental setup of any meaningful application of weights.

In this paper we explore the impact of three different weighting techniques on ensemble performance, using four different ensembles (Sect. 2). We then discuss the results in Sects. 3 and 4 and conclude in Sect. 5.

2 Methodology

The first three ensembles we use, each covering the period 1971–2010, were generated using the CSIRO Mk3L climate system model version 1.2 (Phipps et al. 2011, 2012). All simulations were based on the default fully coupled ocean-atmosphere mode of Mk3L with each simulation perturbed from this baseline configuration. The model simulation period (1971–2010, with a spin-up from 1851) allowed a long period over which the simulations can be compared with reliable observational data of surface air temperature. All ensemble members were integrated over the period 1851–2010, following the protocol for the CMIP5 Historical experiment (Taylor et al. 2012). Accordingly, the model was driven with changes in orbital parameters, atmospheric greenhouse gas concentrations, solar irradiance and stratospheric sulphate aerosols due to volcanic eruptions. The three ensembles were generated by using perturbations from the baseline configuration of the model, using three approaches:

1. Perturb initial conditions. To generate the initial conditions ensemble, restart files with 100 year spacings (sourced from the control simulation used by Phipps et al. 2013), were used as initial conditions. This ensemble has 25 members.
2. Perturb model parameters. To generate this ensemble, six uncertain model parameters representing aspects of the land, ocean and atmosphere were selected with the aim of maximising behavioural diversity, and perturbed simultaneously within literature-based ranges. This ensemble has 25 members.
3. Perturb model structure. In this case, alternative aspects of individual model components were selectively disabled or enabled, including the land surface scheme, atmospheric boundary layer scheme, gravity wave drag scheme, cumuliform and stratiform cloud schemes, the oceanic equation of state, and the dynamical and thermodynamical components of the sea ice model. Five of the simulations failed to complete—this ensemble therefore has 20 members.

To generate the parameter values for the perturbed parameter ensemble members, we used literature-based ranges and perturbed all parameters simultaneously, using the low-discrepancy Sobol' sequence (Reichert et al. 2002) to sample parameter values uniformly within the pre-defined ranges. We used a similar method for the perturbed structure ensemble, first using one simulation with the default settings (all selected model components on); 9 simulations with default settings and one model component modified; and 15 simulations with a quasi-random selection of model component states using the Sobol' sequence to select the states. This method has the benefit of allowing further sampling if required, while maintaining the relative uniformity of coverage of the sample space. A complete description of the ensemble generation process, including parameter ranges and the sampling approach, is given in Haughton et al. (2014). The perturbed structure ensemble is intended to emulate an ensemble of single simulations from several climate models, each using comparable initial conditions and perturbed parameters. It was infeasible to actually generate a structural ensemble with multiple climate models, and so we adopted the above approach using CSIRO Mk3L.

Our fourth ensemble is the collection of CMIP5 historical simulations, a so-called "ensemble of opportunity", in that its makeup is determined by research groups' ability to contribute. Its structure is inevitably part multi-model ensemble, part initial conditions ensemble and part perturbed parameter ensemble. Contributing institutions, model names and the number of simulations from each model are shown in Table 1.

We use the HadCRUT3 surface air temperature data set as our observational reference (Brohan et al. 2006). Model simulations were re-gridded to the HadCRUT3 5° × 5° grid. All simulations were bias corrected (that is, each simulation's time and space mean is made to equal observations), in order to remove the effect of climatic drift over the spin-up period (see Haughton et al. 2014).

2.1 Model weighting and ensemble transformation

Our three approaches to weighting these three ensembles are (a) a simple ensemble mean and unweighted ensemble variance, as is standard practice (Solomon et al. 2007); (b) weights that account for performance differences between ensemble members, applied both to the mean and ensemble variance estimates; and (c) weights that account for *both* performance and dependence between ensemble members, again applied to both mean and variance estimates.

The performance weights we use are inversely proportional to each simulation's error variance—an optimal performance weighting approach for mean square error (MSE) based cost functions (a more explicit definition will be given in the description of independence weighting below). Application of these weights to the ensemble mean estimate is straightforward. To apply them to the ensemble variance estimate, we use a sample weighted variance formula:

$$s^2 = \frac{V_1}{V_1^2 - V_2} \sum_{k=1}^{K} v_k (x_k - \bar{x})^2 \tag{1}$$

where x_k are the ensemble members, $\bar{x}$ is the ensemble mean, v_k are performance weights, $V_1 = \sum_{k=1}^{K} v_k$

and $V_2 = \sum_{k=1}^{K} v_k^2$. If we use normalised weights, we get $V_1 = 1$, and so Eq. 1 becomes:

$$s^2 = \frac{1}{1 - \sum v_k^2} \sum_{k=1}^{K} v_k (x_k - \bar{x})^2$$

For very homogeneous weights, with a large sample size (K), the denominator approaches $1 - 1/K$, and the weighting has very little effect, but when weights are highly heterogeneous, and the sample size is small, the difference between unweighted and weighted variance becomes larger. We use the weighted variance formula for all variance calculations on weighted ensembles.

Bishop and Abramowitz (2013) use error-covariance-based weights to account for model dependence within an ensemble, following the definition of dependence within the replicate earth paradigm described above. While this is just one way of defining independence, it is the only approach that we are aware of that explicitly and quantitatively accounts for model dependence affecting ensemble performance.

Bishop and Abramowitz (2013) argue that since CPDF spread describes essentially unpredictable internal climate system variability, the best possible estimate to any particular random draw from the CPDF, or replicate earth, is the CPDF mean (at least in a MSE sense). They then construct an optimal linear combination of existing ensemble members, μ^j, and argue that this linear combination is the best estimate we can get to the CPDF mean—the true climate response to external forcing. That is,

$$\mu^j = \sum_{k=1}^{K} w_k x_k^j \text{ such that } \sum_{j=1}^{J} (\mu^j - y^j)^2$$

$$\text{is minimised, and } \sum_{k=1}^{K} w_k = 1$$

where $j \in \{1 \ldots J\}$ are time steps, y^j are observations and x_k^j is the jth time step of the kth model simulation. The analytical solution to this problem is expressed in terms of the matrix of pair-wise error covariances between each simulation:

$$A = \begin{pmatrix} \mathrm{cov}(\mathbf{x}_1^{err}, \mathbf{x}_1^{err}) & \mathrm{cov}(\mathbf{x}_1^{err}, \mathbf{x}_2^{err}) & \cdots & \mathrm{cov}(\mathbf{x}_1^{err}, \mathbf{x}_K^{err}) \\ \mathrm{cov}(\mathbf{x}_2^{err}, \mathbf{x}_1^{err}) & \mathrm{cov}(\mathbf{x}_2^{err}, \mathbf{x}_2^{err}) & \cdots & \mathrm{cov}(\mathbf{x}_2^{err}, \mathbf{x}_K^{err}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(\mathbf{x}_K^{err}, \mathbf{x}_1^{err}) & \mathrm{cov}(\mathbf{x}_K^{err}, \mathbf{x}_2^{err}) & \cdots & \mathrm{cov}(\mathbf{x}_K^{err}, \mathbf{x}_K^{err}) \end{pmatrix} \quad (2)$$

where $\mathrm{cov}(\mathbf{x}_n^{err}, \mathbf{x}_m^{err})$ is the error covariance between the nth and mth bias corrected model simulations. The matrix A is then inverted, and the column corresponding to model simulation $\mathbf{x}_k$ is summed, and normalised by dividing through by the sum of the components of the inverted matrix, to give a value w_k for each model:

$$\mathbf{w} = \frac{A^{-1}\mathbf{1}}{\mathbf{1}^T A^{-1}\mathbf{1}} \quad (3)$$

We note that the result in Eq. 3 is reported by both Potempski and Galmarini (2009) and Bishop and Abramowitz (2013). As this is an analytic solution for the minimum error variance estimate in-sample, Bishop and Abramowitz (2013) use μ as an estimate of the CPDF mean.

Note that these coefficients weight for both performance and independence: if we set the off-diagonal terms in A to zero, the solution for the kth simulation, w_k', is proportional to the error variance of the simulation—precisely the performance weights we discussed above. This zero pairwise error covariance scenario is equivalent to the assumption of independence in the truth-plus-error paradigm.

The values $\mathbf{w}$ are defined such that $\sum_{k=1}^{K} w_k = 1$, but they can individually be either larger than 1 or negative. Unfortunately this means that these w_k cannot be considered weights for calculating weighted variance. To overcome this problem Bishop and Abramowitz (2013) use a transformation that modifies both the w_k and the ensemble members themselves. This transformation ensures that (a) the linear combination of modified weights and modified ensemble members is equal to μ^j, and (b) the time average of the instantaneous ensemble variance around the CPDF mean estimate (as estimated by μ) is the same as the variance of observations around the CPDF mean estimate over time. The transformation is a two-step process that first normalises the independence coefficients w_k to weights $\tilde{w}_k$ that are positive and sum to 1, and then inflates and deflates the variance of each ensemble member about the raw ensemble mean and CPDF mean estimate, respectively. The resultant elements are linear combinations of the original model simulations. The new ensemble members retain high correlation with their original corresponding simulations (~0.95), but have different variability structures, and cannot be considered as physically consistent model simulations. With positive weights now available, we can calculate projection variance using these transformed ensemble members. We reiterate that the transformation process does not modify the CPDF mean estimate: the independence coefficient-based linear combination of the original models is the same as the independence-weighted mean of the transformed ensemble members.

We reiterate that the independence weights outlined above account for *both* performance and independence. To aid distinguishing the two weighting approaches however, we will refer to these simply as independence weights below.

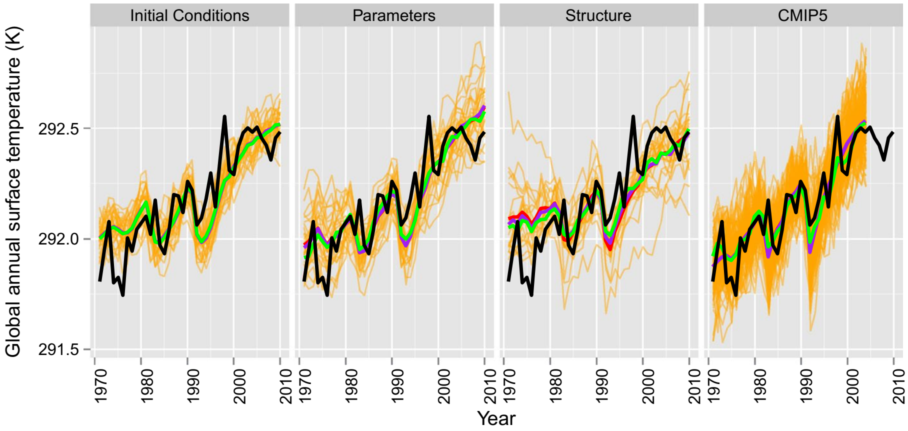

Fig. 2 Global annual mean surface air temperatures (K) of bias corrected model simulations, grouped by generation method. The unweighted mean is in *red*, the performance weighted mean is in purple, and the independence weighted mean is in *green*. Bias-corrected models are in *orange*, and observations are in *black*

The weighting of Bishop and Abramowitz (2013) can be applied in many ways, using different cost functions or temporal and spatial resolutions. In particular, it can be applied on a per-cell basis— that is, each model simulation has one weight per grid cell, and the time series for each grid cell from different models are combined using independent weights—or it can be applied globally, using all data to calculate a single weight per simulation. Bishop and Abramowitz (2013) use both global and per-cell weighting. For this study we use only global weighting, based on per-cell data and monthly time steps aggregated into a single vector. There are therefore K weights for K model simulations.

As illustrated by Bishop and Abramowitz (2013), weighting model simulations for independence is somewhat analogous to removing model simulations from the ensemble—if two models are dependent, their simulations contain very similar information. Bishop and Abramowitz (2013) investigated out-of-sample performance of the approach by testing the transformation on one decade of the late twentieth century and testing on others, with stable positive results. We also note that Abramowitz and Bishop (2014) further investigate this by using a perfect model approach with CMIP5 historical and RCP projections, and again find the correction to be stable out-of-sample.

3 Results

Figure 2 shows time series of annual averages of global mean surface air temperature for all members of the four bias corrected ensembles. Note that CMIP5 Historical simulations stop after 2004. Three means are displayed for each ensemble: the unweighted multi-model mean (red), the performance weighted mean (purple), and the independence weighted mean (green). In much of Fig. 2 they are indistinguishable. As noted above, these use global weights, calculated using all grid cells and monthly time steps of available data. They are *not* calculated using global average temperatures. We therefore need not necessarily expect that weighting would have a major effect on global averages, since the discrepancies between models may be dominated by regional differences. The weights *could* be calculated over globally averaged annual data, however the resulting mean would likely be over-fitted, as there are 20–25 free variables (weights per ensemble), and only 40 data points.

RMSE values of each ensemble mean across time and space are given in Table 2. Under standard unweighted averaging, the initial conditions ensemble mean performs notably better than either the perturbed parameter or structural ensembles, and the structural ensemble has the worst performing mean. For the initial conditions ensemble, all three means are almost identical, and there is very little improvement due to either the performance or independence-weighting. For the perturbed parameter ensemble, the performance weighted mean over the training period shows a small improvement over the unweighted mean at the global scale (1.6 %), while the independence weighted mean shows a slightly larger improvement (3.5 %). For the perturbed structure ensemble, the performance improvement of the two weighted means over the unweighted mean is larger, although the improvement due to the performance weighting (15 %) relative to that due to the independence weighting (20 %) is larger than for the perturbed parameters ensemble. Under the independence weighting, the perturbed structure ensemble out-performs the perturbed parameter ensemble, and compares favourably with the performance of the initial conditions ensemble. CMIP5 shows a similar pattern of improvement to the latter two ensembles under both weighting methodologies (0.4 and 3.1 % respectively). The CMIP5 ensemble outperforms all of the Mk3L ensembles under this metric.

We now examine the effect of each weighting methodology on the variance of the ensembles. The top rows of Figs. 3 and 4, as with Fig. 2, show time series of global

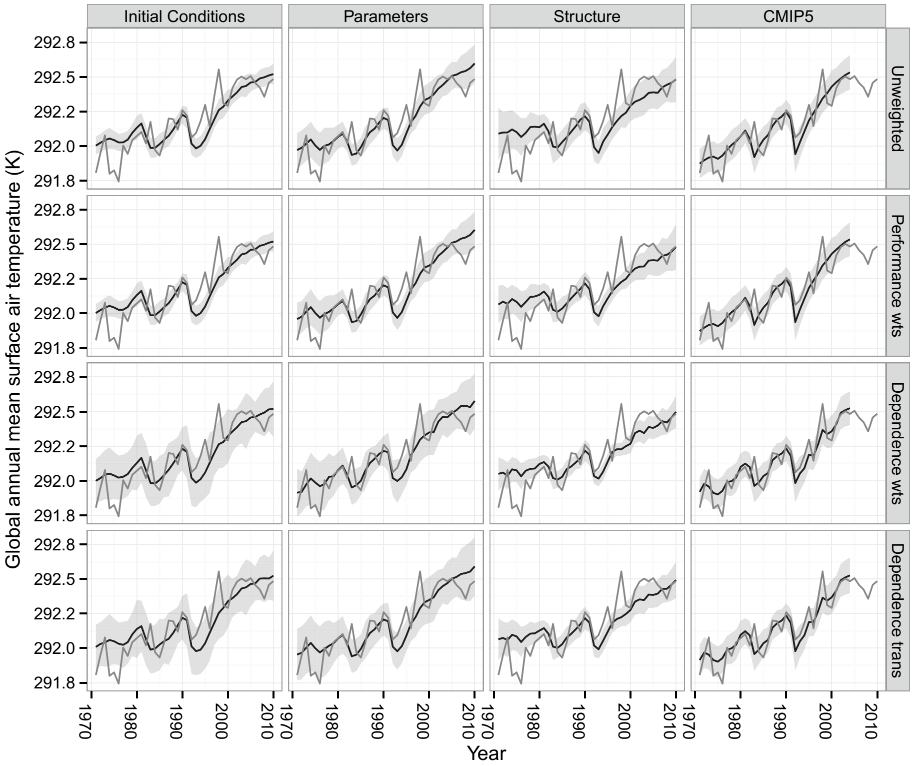

Fig. 3 Global average annual surface air temperature from HadCRUT3 (*grey solid line*) and each bias corrected ensemble for the period 1971–2000 (1971–2004 for CMIP5). Columns represent different ensembles and rows different approaches to weighting. Each weighted ensemble mean is shown in *black*, with the shaded region showing the standard deviation of simulations in the ensemble, at each time step. Weighted standard deviations are calculated using weights applied to global annual averages

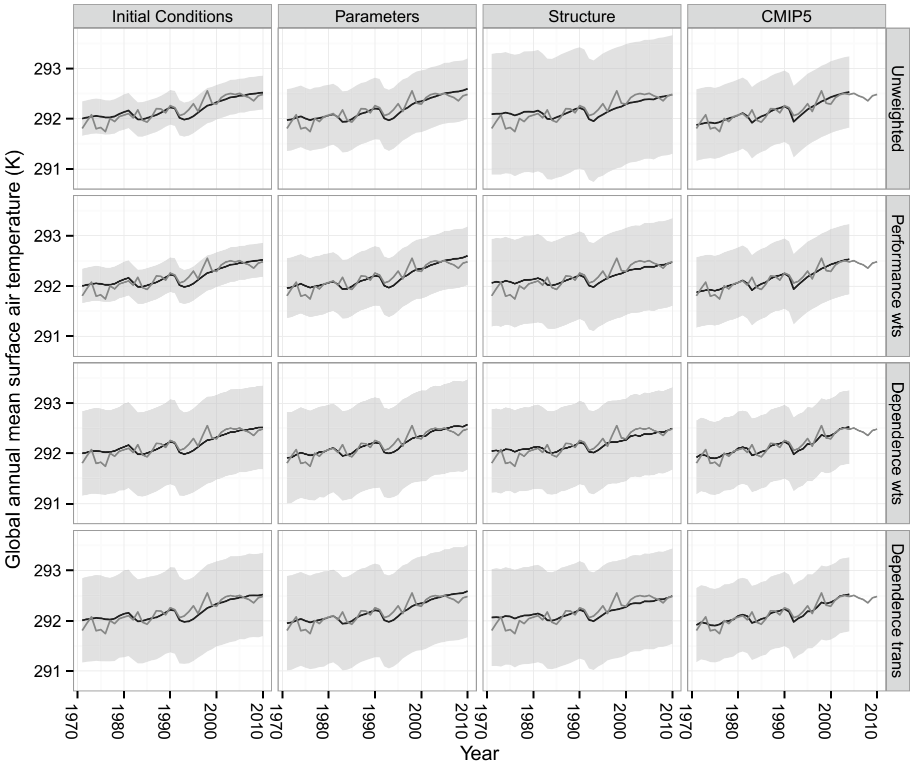

Fig. 4 As for Fig. 3, but using weighted standard deviation calculated on the monthly per-cell data, then averaging standard deviation to global annual scale for plotting. The *shaded region* represents one-half standard deviation

annual average surface air temperature. Observations are in dark grey, the unweighted ensemble mean is in black, and the light grey shading shows one standard deviation of ensemble spread for each year. The second and third rows of Fig. 3 show the application of the performance and independence weights, respectively. The ensemble means for the second and third rows are the same as the purple and green lines in Fig. 2. The shaded ensemble standard deviation for each of these two rows is calculated using the weighted variance in Eq. 1 and the appropriate set of weights.

The weights we use here were derived on all 5° × 5° grid cells and at all monthly time steps. As such, they account for performance differences and error covariance in different regions and in different seasons. The considerable regional biases that different models exhibit likely dominate error covariances between simulations, and so exhibit a significant influence on the weights constructed in this way. The application of these weights to the calculation of ensemble variance in globally averaged annual temperature, as is done in row three of Fig. 3, seems somewhat inappropriate. With this in mind, we now present two alternative methods of calculating the variance of the ensembles at this scale.

In Fig. 3, the standard deviation of the ensemble is calculated using globally averaged annual surface temperature. In Fig. 4, the standard deviation of the ensemble is calculated at each grid cell and monthly time step, and then averaged over the globe and each year. As we would expect, the averaged standard deviation in the latter figure is considerably larger than that shown in Fig. 3, since it captures the regional and seasonal differences between models. As a result, the shaded regions in Fig. 4 show only one-half of the standard deviation (so that variations in the ensemble means are still visible). Since the mean and standard deviations in this figure are calculated at different scales, it is the relative changes in the spread that is of importance in this figure. We reiterate that deriving weights using global annual averages would be inappropriate, as we have 20–25 model time series with only 40 constraining data. It should also be clear that this issue does not affect weighting for the ensemble mean. It is an open question whether the application of these weights to calculating ensemble variance *after* global annual averages are calculated is appropriate.

The fourth rows of Figs. 3 and 4 show a second approach to addressing this issue. They show the variance of the independence transformed ensemble calculated *without* using a weighted variance. The similarity of this result to the weighted transformed ensembles in the row above it in both figures shows that the transformation process itself provides most of the change to an ensemble's variance, rather than the weights (at least at the global scale).

We can also examine the spread of the ensembles by looking at the percentage of observations that fall within one standard deviation of the ensemble mean. Results are shown in Table 3. For a normal distribution, the value would be expected to approach 68.3 %, but the appropriate shapes of the distributions for these ensembles (representing CPDF distribution estimates) are not known so this assumption should be treated with caution. What is clear is that the improvement in ensemble variance under the performance weighting is far less consistent than the improvement to the mean: the initial conditions ensemble variance estimate does not change, and the perturbed parameters ensembles variance estimate actually degrades slightly under the performance weighting (assuming that we are expecting a value near 68 %). Since we *don't* know the variance of the observations around the true CPDF mean, it is difficult to say whether the variance in the structural ensemble or CMIP5 ensemble is improved or degraded under the performance weighting. It is possible that performance weighting improves variance estimates for ensembles that are very over-dispersive (such as our structural ensemble), however this data does not provide compelling evidence. We also note that we are only considering error covariance based performance weights. It is theoretically possible that, under other cost functions, performance-weighted variance does not perform so poorly.

Under the independence transformation, variance of the initial conditions and perturbed parameters ensembles improve dramatically. We also note that the percentage of observations that lie within one standard deviation of the ensemble mean *after* the independence transformation varies by only 9 % across all of these very different ensembles (as opposed to 40 % before transformation). This apparent convergence raises the intriguing question of whether the appropriate expected value of the percentage of observations that fall within one standard deviation of the true CPDF mean might be around these values. We also note that the range (65.41–73.98 %) includes 68.3 %, which is the value that would be expected if the CPDF followed a normal distribution.

To better understand dependence within each of these ensembles, we now also examine the pair-wise error correlations between simulations in each ensemble. Note that while we used error covariance when applying Bishop and Abramowitz's methodology, error correlation as a normalised measure is a little more intuitive and allows direct comparison between model ensembles.

Figure 5 shows histograms of pair-wise error correlations between the simulations in each ensemble. There is a clear change in homogeneity between the ensembles: initial conditions ensemble simulation pairs all have very similar correlations, while the perturbed parameters ensemble error correlations are much broader, and those of the structural ensemble broader again. The average pair-wise error correlation for the initial conditions ensemble is 0.79, for the perturbed parameters ensemble 0.60, and for the structural ensemble 0.35. The 95 % confidence interval for the initial conditions ensemble and perturbed parameters ensemble both exclude 0. The confidence interval of the initial conditions ensemble (0.782, 0.80) also excludes 0.5, the value

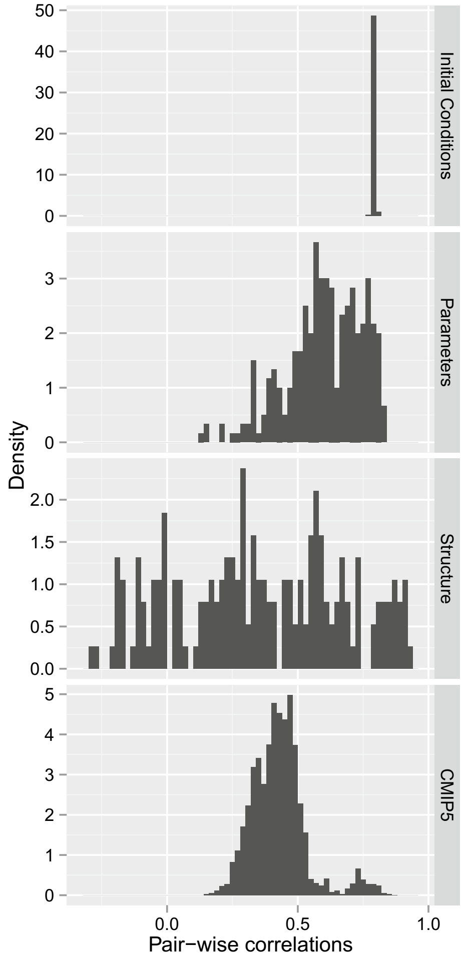

Fig. 5 Density of pair-wise error correlations between simulations in each ensemble. There are 300 pairs in the first two ensembles, and 190 in the structural ensemble. The bimodality seen in the CMIP5 ensemble is due to intra-model simulation dependence, as shown in Fig. 6

expected by the indistinguishable paradigm. The confidence interval for the perturbed structure ensemble is far broader, partly due to the sample size, with a confidence interval of (−0.27, 0.98), and includes both 0 and 0.5.

The narrowness of the initial conditions ensemble may be expected from Fig. 2, where we see that the variance between the simulations is sufficiently small that the range frequently does not span the observations. In particular, in the mid 1970s, the observations are lower than all the models, while in the mid-late 1990s, the opposite is true. This alone would add significantly to the correlations between simulation errors. There are likely similar patterns in seasonal and spatial trends that we do not see in Fig. 2 because of global averaging. This clearly points to strong dependence between the simulations in the initial conditions simulations.

Given those factors, the spread of the correlations between structural ensemble simulations is somewhat surprising—there are even simulation pairs with negative correlation. This means that the patterns differ between the simulations so much that the variance introduced by the observations is outweighed by the variance between the simulations.

The CMIP5 ensemble is distinctly bimodal, with a significant cluster in the range of 0.7–0.9. This cluster is almost entirely due to high correlations between pairs of simulations from the same modelling system within a single institution. All correlations between model pairs for CMIP5 are shown in Fig. 6, which clearly shows strong clustering from model simulations from nearly every modelling institution (see Table 1 for a complete list of simulations). This indicates that most model simulations from a single institution are highly dependent.

Most of the inter-institutional model simulation pairwise correlations lie in the range of 0.2–0.6. This indicates that the these model pairs are reasonably independent. However, Fig. 6 also shows some inter-institutional clustering—for example, the Australian ACCESS 1.0 and 1.3 models, which share the atmospheric component with the UKMO HadGEM2 model (shown by the @ in Fig. 6). As such, this model pair exhibits less independence than most model pairs in the ensemble.

Although the CMIP5 ensemble could be considered a structural ensemble, or at least a mixed structural/parameter/initital conditions ensemble, it is clear that even considering the error correlations among the inter-model or inter-institution simulations (left part of the bi-modal distribution seen in the fourth panel of Fig. 5) there is not as much diversity as is seen in our structural ensemble. This may be due to a number of factors, including deliberate model tuning. Our pseudo-random structural perturbations had no targeted performance goal behind them, while all of the CMIP5 simulations were produced with the intention of replicating historical obervations accurately. Deliberate or subconscious selection of the best simulations during the submission process for CMIP5 might also be a factor.

4 Discussion

Performance weighting is in widespread use, and is an active area in ensemble research, yet almost all weighting procedures consider only the weighted mean, and ignore the effect on variance (e.g. Giorgi 2005; Krishnamurti et al. 2000). Recent reviews of model combination methodologies and issues related to weighting also appear to have largely ignored the application of any kind of weighting to ensemble variance (Weigel et al. 2010; Knutti et al. 2010a). This may well be because it can be very problematic: as we have shown, performance weighting may reduce the quality of ensemble variance in some situations.

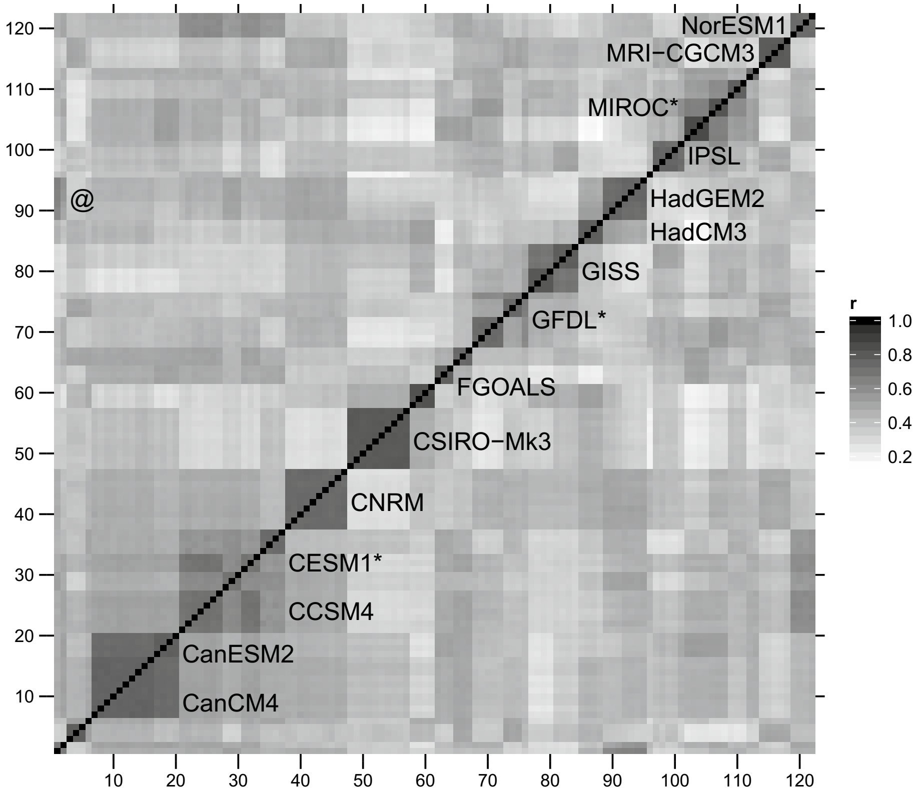

Fig. 6 Correlation between the error time series of each pair of model simulations. Model simulations are sorted alphabetically, as listed in Table 1, and major groups' simulations are annotated. The * indicates a model group with multiple model variations. *Darker cells* indicate higher correlation. Some models from different groups share high correlation, for example the Australian ACCESS model, and the UK HadGEM2 model, from which ACCESS was forked (indicated by the @)

As a consequence, performance-weighted ensembles may underestimate the variability we should expect to see in the future climate. For small ensembles, this would likely only be compounded by using simple variance calculations in the place of appropriate weighted variance calculations, as explained in Sect. 2. As we have only considered error variance based performance weighting, it is theoretically possible that other cost functions do not suffer as severely from this problem. However, the very nature of performance weighting reduces the impact of the more extreme samples, and in doing so likely reduces the variance of ensembles.

In contrast, the means and variance of the ensembles both improved under the independence transformation, and improvements were far more consistent from each ensemble. In the replicate earth transformation process, the distance of the CPDF mean estimate from the observations is critical in determining how broad the CPDF variance estimate should be. The CPDF variance estimate is just the time-averaged variance between the observations and the CPDF mean estimate. This means that if the CPDF mean tracks the observations very closely, the difference between the mean and the observations, and hence the CPDF variance estimate, will be very small. If, on the other hand, the CPDF mean estimate is very smooth, the internal variability in the observations will ensure that the CPDF variance estimate is larger.

The obvious implication here is that an ensemble of simulations from models which are in some way over-fitted or have coincident internal variability with the observed system will likely underestimate the variance. On the other hand, an ensemble of poorly performing models is likely to overestimate variance. It is also worth noting here that the CPDF variance estimate is, to some degree, a function of the number of models. The more models that are added to the ensemble, the more tightly the CPDF mean estimate can be fitted to the observations. However, this is not likely to be a problem except for very large ensembles with a very short period, or low spatial resolution. In our ensembles, we have at most 25 models with which to fit hundreds of thousands of data points, so over-fitting is unlikely to be a problem.

The CPDF mean, ultimately, represents the mean response to all large-scale forcings that all replicate earths would share. So the CPDF mean should respond to, for example, changes in CO_2, solar irradiance, and volcanic and anthropogenic aerosols, which are shared inputs to models. It should not respond directly to chaotic fluctuations in internal model processes, such as ENSO cycles; these processes are represented by the spread of the CPDF about the mean. However, it *should* capture changes in the patterns of those chaotic fluctuations: for example, a state-shift that shuts down the North Atlantic thermohaline circulation, or a shift to a permanent El Niño like state, if those

Table 1 List of CMIP model simulations used in this study

Simulation	Model	Institution
1	ACCESS1.0	CSIRO-BOM
2	ACCESS1.3	
3–5	bcc-csm1.1	BCC
6	BNU-ESM	GCESS
7–16	CanCM4	CCCma
17–20	CanESM2	
21–26	CCSM4	NCAR
27	CESM1-BGC	NSF-DOE-NCAR
28–30	CESM1-CAM5	
31–33	CESM1-FASTCHEM	
34–37	CESM1-WACCM	
38–47	CNRM-CM5	CNRM-CERFACS
48–57	CSIRO-Mk3-6.0	CSIRO-QCCCE
58–61	FGOALS-g2	LASG-CESS
62–64	FGOALS-s2	LASG-IAP
65–67	FIO-ESM	FIO
68–72	GFDL-CM3	NOAA GFDL
73–76	GFDL-ESM2G	
77–80	GISS-E2-H	NASA GISS
81–84	GISS-E2-R	
85–88	HadCM3	MOHC
89–91	HadGEM2-CC	
92–95	HadGEM2-ES	
96	inmcm4	INM
97–100	IPSL-CM5A-LR	IPSL
101	IPSL-CM5A-MR	
102–104	MIROC-ESM	MIROC
105	MIROC-ESM-CHEM	
106–8	MIROC4h	
109–111	MIROC5	
112	MPI-ESM-LR	MPI-M
113	MPI-ESM-P	
114–118	MRI-CGCM3	MRI
119–121	NorESM1-M	NCC
122	NorESM1-ME	

Simulation numbers correspond to the rows and columns of Fig. 6. More details of the modelling institutions are available at http://cmip.llnl.gov/cmip5/availability.html

changes represent a forced response to the given boundary conditions. The difficulty then lies in determining which changes are important, affecting every replicate earth, and which are replicate-specific. The possibility of bifurcations in the distribution of possible states would clearly complicate this viewpoint.

The results shown in Figs. 5 and 6 highlight a striking difference between the truth plus error and replicate earth paradigms. Under the truth plus error paradigm we expect the observations to be the centre of the model distribution. With this understanding, we should expect the model error to be randomly distributed around the observations, and hence expect a set of independent models to have a mean error correlation of zero.

In contrast, under the indistinguishable paradigm we expect that the observations are similar to model simulations, as both are drawn from the same distribution. Under the replicate earth paradigm, the same is true, but only if the model simulations adequately represent replicate earths. In both cases, “model errors” are actually a linear combination of two samples (the model, minus the observations), and some of the variance between a pair of model errors is contributed by the observations. Thus, if the models and observations are independently drawn from the same distribution, as in the indistinguishable paradigm, or are true replicate earths, then the expected correlation between error pairs is actually 0.5 (Annan and Hargreaves 2010; Bishop and Abramowitz 2013).

Under the replicate earth paradigm, if the models and observations are *not* drawn from the same distribution—for example, if the models’ distribution has less variance—then the observations contribute more variance to the linear combination, and we should expect higher error correlations. Likewise, if the models’ spread is higher than the observations, we should expect error correlations to be lower.

Figure 5 shows that the truth plus error paradigm would be hard to justify with any of these ensembles. The results for the initial conditions ensemble would be hard to support even under the indistinguishable paradigm, where we would *always* expect 0.5 error correlation, assuming the sample is long enough. The perturbed parameters ensemble error correlations also clearly indicate a problem with the truth-plus-error paradigm, although they do not indicate a clear distinction between the indistinguishable and the replicate earth paradigms. But under the replicate earth paradigm, if models are *not* replicate earth-like, we actually expect the error correlations to vary depending on the variance in the observations relative to the variance in

Table 2 RMSE values for the means of each ensemble, relative to the observations for the entire period (1971–2010, 1971–2004 for CMIP5)

	Initial conditions	Perturbed parameters	Perturbed structure	CMIP5
Unweighted	2.050	2.205	2.577	1.672
Performance weighted	2.050	2.170	2.183	1.666
Independence weighted	2.049	2.127	2.056	1.620

Table 3 Percentage of observed data that fall within one weighted standard deviation of the simulations about the ensemble mean (1971–2010, 1971–2004 for CMIP5)

	Initial conditions	Perturbed parameters	Perturbed structure	CMIP5
Unweighted	38.05	53.51	77.40	62.51
Performance weighted	38.05	52.88	73.43	61.95
Independence weighted	74.13	72.43	72.14	65.06
Independence transformed	73.98	72.94	76.35	65.41

the models, and due to the very small spread of the initial conditions ensemble, a mean pair-wise error correlation of 0.79 seems entirely understandable.

We also reiterate that we only have one true sample of the CPDF— the underlying distribution of probable earth states over time, given known boundary conditions—with which to estimate what the entire CPDF looks like. Hence it is difficult to categorically state the extent to which any of these ensembles truly represent the CPDF. We cannot know for certain that our observations are not a stark outlier—that most other replicate earths would not be quite different (effectively suggesting that our observational record is too short). If they are, then our CPDF estimates will clearly be biased. The results in Table 3 do, however, go some way to allaying this concern.

Finally, as the process has been described in Bishop and Abramowitz (2013), this weighting approach is limited to one variable at a time. While in one sense, separate applications to temperature and precipitation may lead to physically inconsistent best estimates, neither the CPDF mean or the unweighted ensemble mean are true climate realisations to begin with, so this is less of an issue than it may seem. We note that Abramowitz and Bishop (2014) show an application of the approach to precipitation, and that Potempski and Galmarini (2009) may offer an approach to allow weighting several variables at once.

5 Conclusions

We have shown that different weighting methodologies have significantly different effects on the four ensembles we considered. In particular, although means generally improve under performance-based weighting, we have shown that variance may not improve, and may in fact worsen. There remains the possibility that this result is particular to error variance-based performance weighting, and may be quite different for other cost functions, but without evidence it seems unlikely.

In contrast, independence-based weighting significantly improves both the ensemble mean and variance. The improvement to the ensemble mean is notably better than that under performance weighting, and the improvement to the variance is both better and far more consistent across different ensembles. This suggests that independence weighting could provide large gains in projection accuracy, including estimates of uncertainty, which, by reducing uncertainty around actions needed to avert the worst of climate change, could be of considerable benefit.

Finally, more work clearly needs to be done in testing the efficacy of this process using different variables, resolutions and observationally-based products. In particular, it would be useful to conduct similar experimentation using different cost functions as the basis for the dependence measure. Nevertheless, the significant gains apparent across the range of ensembles shown here suggest that post processing can yield considerably improved ensemble mean and variance estimates.

Acknowledgments 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 U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. 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

Abramowitz G, Bishop CH (2014) Climate model dependence and the ensemble dependence transformation of CMIP projections. J Clim. doi:10.1175/JCLI-D-14-00364.1. http://journals.ametsoc.org/doi/abs/10.1175/JCLI-D-14-00364.1

Annan JD (2010) IPCC experts new clothes. http://julesandjames.blogspot.com.au/2010/08/ipcc-experts-new-clothes.html

Annan JD, Hargreaves JC (2010) Reliability of the CMIP3 ensemble. Geophys Res Lett 37:5. doi:10.1029/2009GL041994. 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

Flato G, Marotzke J, Abiodun B, Braconnot P, Chou SC, Collins W, Cox P, Driouech F, Emori S, Eyring V, Forest C, Gleckler P, Guilyardi E, Jakob C, Kattsov V, Reason C, Rummukaine M (2013) Evaluation of climate models. In: Stocker T, Qin D, Plattner GK, Tignor M, Allen S, Boschung J, Nauels A, Xia Y, Bex V, Midgley P (eds) Climate change 2013: the physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change, Cambridge University Press, Cambridge. http://www.climatechange2013.org/images/report/WG1AR5_Chapter09_FINAL.pdf

Giorgi F (2005) Climate change prediction. Clim Change 73(3):239–265. doi:10.1007/s10584-005-6857-4

Gleckler P, Taylor K, Doutriaux C (2008) Performance metrics for climate models. J Geophys Res 113(D6):D06,104. doi:10.1029/2007JD008972. http://www.agu.org/pubs/crossref/2008/2007JD008972.shtml

Haughton N, Abramowitz G, Pitman AJ, Phipps SJ (2014) On the generation of climate model ensembles. Clim Dyn, pp 1–12. doi:10.1007/s00382-014-2054-3

Hegerl GC, Zwiers FW, Braconnot P, Gillett NP, Lou Y, Marengo Orsini JA, Nicholls N, Penner JE, Stott PA (2007) Understanding and attributing climate change. In: Understanding and attributing climate change, Cambridge University Press, Cambridge. http://www.ipcc.ch/publications_and_data/publications_ipcc_fourth_assessment_report_wg1_report_the_physical_science_basis.htm

Knutti R, Abramowitz G, Collins M, Eyering V, Gleckler PJ, Hewitson B, Mearns LO (2010a) Good practice guidance paper on assessing and combining multi model climate projections, IPCC working group I technical support unit. 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. University of Bern, Bern

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. doi:10.1175/2009JCLI3361.1. http://0-journals.ametsoc.org.library.newcastle.edu.au/doi/abs/10.1175/2009JCLI3361.1

Krishnamurti TN, Kishtawal CM, Zhang Z, LaRow T, Bachiochi D, Williford E, Gadgil S, Surendran S (2000) Multimodel ensemble forecasts for weather and seasonal climate. J Clim 13(23):4196–4216. doi:10.1175/1520-0442(2000)013. http://journals.ametsoc.org/doi/abs/10.1175/1520-0442(2000)013%3C4196:MEFFWA%3E2.0.CO%3B2

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

Masson D, Knutti R (2011) Climate model genealogy. Geophys Res Lett 38(8). doi:10.1029/2011GL046864

PAGES 2k Consortium (2013) Continental-scale temperature variability during the past two millennia. Nature Geosci 6(5):339–346. doi:10.1038/ngeo1797. http://www.nature.com/ngeo/journal/vaop/ncurrent/full/ngeo1797.html

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 26(18):6915–6936. doi:10.1175/JCLI-D-12-00108.1. http://journals.ametsoc.org/doi/abs/10.1175/JCLI-D-12-00108.1

Potempski S, Galmarini S (2009) Est modus in rebus: analytical properties of multi-model ensembles. Atmos Chem Phys 9(24):9471–9489. doi:10.5194/acp-9-9471-2009. http://www.atmos-chem-phys.net/9/9471/2009/

Randall D, Wood R, Bony S, Colman R, Fichefet T, Fyfe J, Kattsov V, Pitman AJ, 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) Cilmate models and their evaluation. In: 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. http://www.tandfonline.com/doi/abs/10.1198/004017002188618518

Reifen C, Toumi R (2009) Climate projections: past performance no guarantee of future skill? Geophys Res Lett 36(13). doi:10.1029/2009GL038082. http://www.agu.org/pubs/crossref/2009/2009GL038082.shtml

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

Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. Bull Am Meteorol Soc 93(4):485–498. doi:10.1175/BAMS-D-11-00094.1. http://journals.ametsoc.org/doi/abs/10.1175/BAMS-D-11-00094.1

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

Weigel AP, Knutti R, Liniger MA, Appenzeller C (2010) Risks of model weighting in multimodel climate projections. J Clim 23(15):4175–4191. http://journals.ametsoc.org/doi/abs/10.1175/2010JCLI3594.1