# Climate model dependence and the replicate Earth paradigm

**Author:**
Bishop, C; Abramowitz, Gabriel

**Publication details:**
Climate Dynamics
v. 41
Chapter No. 3-4
pp. 885-900
0930-7575 (ISSN)

**Publication Date:**
2013

**Publisher DOI:**
http://dx.doi.org/10.1007/s00382-012-1610-y

# Climate model dependence and the replicate Earth paradigm

Craig H. Bishop
Naval Research Laboratory,
Marine Meteorology Division
7 Grace Hopper Avenue, Monterey, California 93943-5502
craig.bishop@nrlmry.navy.mil

Gab Abramowitz
Climate Change Research Centre
University of New South Wales
Kensington, Sydney
gabriel@unsw.edu.au

19th September, 2012

**Keywords**

Climate model ensembles, model independence, climate uncertainty quantification, climate model bias correction.

*Corresponding Author:*

Gab Abramowitz
gabriel@unsw.edu.au
Climate Change Research Centre
University of New South Wales, NSW 2052
Australia
Ph: +61 2 9385 8958
Fax: +61 2 9385 8969

**Abstract**

Multi-model ensembles are commonly used in climate prediction to create a set of independent estimates, and so better gauge the likelihood of particular outcomes and better quantify prediction uncertainty. Yet researchers share literature, datasets and model code – to what extent do different simulations constitute independent estimates? What is the relationship between model performance and independence? We show that error correlation provides a natural empirical basis for defining model dependence and derive a weighting strategy that accounts for dependence in experiments where the multi-model mean would otherwise be used. We introduce the "replicate Earth" ensemble interpretation framework, based on theoretically derived statistical relationships between ensembles of perfect models (replicate Earths) and observations. We transform an ensemble of (imperfect) climate projections into an ensemble whose mean and variance have the same statistical relationship to observations as an ensemble of replicate Earths. The approach can be used with multi-model ensembles that have varying numbers of simulations from different models, accounting for model dependence. We use HadCRUT3 data and the CMIP3 models to show that in out of sample tests, the transformed ensemble has an ensemble mean with significantly lower error and much flatter rank frequency histograms than the original ensemble.

## 1. Introduction

Multi-model ensemble prediction aims to quantify climate prediction uncertainty by considering a number of simulations from a range of different models or modelling approaches. Synthesizing the information from this collection of simulations remains a somewhat subjective process. Some research suggests weighting (e.g. Krishnamurti et al, 2000; Giorgi and Mearns, 2002; Tebaldi et al, 2005) or sub-selecting (e.g. Perkins et al, 2007; Gleckler et al, 2008) different simulations or models, based on their performance. By far the most common and widely accepted approach is to simply use the multi-model mean (e.g. Lambert and Boer, 2001; Gleckler et al, 2008; although this is clearly not appropriate for all types of evaluation, for example variability estimates). Its acceptance is evident in the widespread use of the multi-model mean to represent "best guess" scenarios in the Intergovernmental Panel on Climate Change's Fourth Assessment Report (henceforth IPCC AR4; Meehl et al, 2007b), and was reinforced in the recent Expert Meeting on Assessing and Combining Multi-model Climate Projections in 2010 (Knutti et al, 2010a).

There are a number of perspectives on why the multi model mean performs so well. Imagine, for example, that each model's error time series (modelled minus observed) were a random number time series with variance equal to 1 and zero mean. We know that if we examine the mean of n>1 independent random number time series, its variance will be much lower than 1 (in fact it will approximate 1/n). While we will spend some time explaining why model errors do not behave like random number time series, it is nevertheless true that the multi-model mean tends to cancel out the eccentricities of individual models (both random variability and structural errors). This is clearly seen in Figure 1, taken from the IPCC AR4, where the red multi-model mean is a smoothed representation of yellow individual models. It is important to note, however, that the multi-model mean has very different properties to any particular model simulation. At least anecdotally from Figure 1, we can see that the multi-model mean has significantly less variance than other model time series. Perhaps more importantly, the observational time series appears to have variance more like an individual model than the multi-model mean, yet the mean consistently provides a better a-priori estimate than any individual model. In Section 4 we argue that this suggests a framework for interpreting multi-model ensembles that sees model simulations and observational time series as indistinguishable realizations of the Earths climate (e.g. Annan and Hargreaves, 2010; Annan and Hargreaves, 2011), in that the best estimate to any particular realization will be the multi-model mean, without any expectation that the mean will match that realization.

Whether or not one subscribes to this framework or one that sees the multi-model mean of perfect models converging to observations as an ensemble's size grows, the use of the multi-model mean as the best estimate works best when each model provides an independent estimate. Yet modelling groups share data sets, parametrisations and even sections of model code, so there are reasons to suspect that the assumption of statistical independence of every climate prediction may not be appropriate (Tebaldi and Knutti, 2007; Jewson and Hawkins, 2009; Knutti et al, 2010a). This issue is even more pressing when we consider coincident prediction (e.g. Giorgi and Mearns, 2002) – to what degree should model agreement be a sign of robustness?

Before discussing how best to define model independence in ensemble prediction, we use an analogy to highlight the critical distinction between model performance and model independence. Suppose we wish to estimate the horizontal coordinates of the peak of a hill by averaging the position of several walkers climbing the hill. While we want the walkers (the models) to be close to the top of the hill (the observation or truth), to achieve the best estimate we also want them to be spread evenly around the peak. This estimation technique is analogous to ensemble averaging, distance from the peak analogous to performance, and spread around the peak analogous to independence. This highlights the possibility that the mean of an ensemble of relatively poor performing but independent models could outperform the mean of an ensemble of relatively dependent but well performing models.

Intuitively, one may want to define model independence in terms of shared model structure or parametrisations (as though we were using evolutionary cladistics for species classification; in which case shared genetic history implies dependence, e.g. Masson and Knutti, 2011). Here, however, we take a more pragmatic approach and focus on the independence of models' simulations (perhaps more analogous with Linnaean taxonomy). In fact we suggest there is an obvious choice for empirically defining model dependence – correlation in model errors. Section 2 focuses on deriving weights that explicitly account for model dependence defined using correlation of model errors. Section 3a applies these weights to a "toy" example to examine their behaviour before Section 3b examines their application to a collection of climate models and observed surface temperature data. Section 4 discusses how different interpretations of the relationship between observations and a model ensemble lead to very different formalisations of the definition of dependence outlined in Section 2. Section 5 introduces an ensemble transformation process that both improves the predictive power of the multi-model mean and constrains ensemble variance so

that instantaneous ensemble variance reflects the variance of the climatic distribution of weather states. Section 6 presents discussion and conclusions.

## 2. Defining and weighting for model dependence

We aim to demonstrate that correlation in model errors is a good basis for a definition of model dependence. The association of dependence and error correlation *per se* is not new (e.g. Jun et al, 2008; Collins et al, 2010) but the approach we outline below offers both a compelling reason for it as well as an optimal weighting solution. To begin, we allow the (seemingly inappropriate) use of mean square error for time-series evaluation of climate model simulations. We will spend some time in Section 4 justifying this decision, but for now we note that this decision need not (and does not) imply an expectation that a perfect model should match observations. To emphasise this point, we will refer to mean square difference (MSD), rather than mean square error. Next, suppose we wish to find the linear combination of an ensemble of model simulations that minimizes MSD with respect to an observational data set. That is, for time steps $(1,\ldots,j,\ldots,J)$ and bias-corrected models $(1,\ldots,k,\ldots,K)$, we want to find

$$\mu_e^j = \mathbf{w}^T \mathbf{x}^j = \sum_{k=1}^{K} w_k x_k^j \qquad \text{so that} \qquad \sum_{j=1}^{J} \left(\mu_e^j - y^j\right)^2 \tag{1}$$

is minimized, where $x_k^j$ is the $j^{\text{th}}$ time step of the $k^{\text{th}}$ bias-corrected model, $y^j$ is the $j^{\text{th}}$ time step observation, $w_k$ is the $k^{\text{th}}$ model coefficient in the linear combination, $\mathbf{w}^T = \left[w_1, w_2, ..., w_K\right]$ and $\left(\mathbf{x}^j\right)^T = \left[x_1^j, x_2^j, ..., x_K^j\right]$. Bias-correction in this case simply refers to subtracting the mean error from a model's time series for the in-sample period. Additionally, we want to constrain the coefficients $w_k$ to sum to 1, so that this constrained least squares minimisation problem is solved using a Lagrange multiplier, $\lambda$:

$$F\left(\mathbf{w},\lambda\right) = \frac{1}{2}\left[\frac{1}{\left(J-1\right)}\sum_{j=1}^{J}\left(\mu_e^j - y^j\right)^2\right] - \lambda\left(\left(\sum_{k=1}^{K} w_k\right) - 1\right). \tag{2}$$

The solution to the minimisation of (2) (fully detailed in Electronic Supplementary Material A) can be expressed as

$$\mathbf{w} = \frac{\mathbf{A}^{-1}\mathbf{1}}{\mathbf{1}^T\mathbf{A}^{-1}\mathbf{1}} \tag{3}$$

where $\mathbf{1}^T = \overbrace{\left[1,1,...,1\right]}^{K \text{ elements}}$ and $\mathbf{A}$ is the $K$x$K$ difference covariance matrix

158 $$\mathbf{A} = \begin{pmatrix} \mathbf{c}_{1,1} & \cdots & \mathbf{c}_{1,K} \\ \vdots & \ddots & \vdots \\ \mathbf{c}_{K,1} & \cdots & \mathbf{c}_{K,K} \end{pmatrix}. \quad (4)$$

159 That is, $\mathbf{c}_{i,j}$ is the covariance of the *i*th and *j*th bias-corrected model minus observed time series (this
160 is effectively an error covariance matrix, without any expectation that errors should be zero for a
161 perfect model). Note that **A** is symmetric and that each diagonal term $\mathbf{c}_{k,k}$, the error covariance of
162 model *k* and model *k*, is just the error variance of model *k*, or $\sigma_k^2$. Note also that the denominator in
163 (3) (which is the sum of all of the elements of $\mathbf{A}^{-1}$) is constant for all *k*, and so is effectively just a
164 scaling factor. Each $w_k$ is then proportional to the sum of the elements in the $k^{\text{th}}$ row of $\mathbf{A}^{-1}$.
165
166 Now if we assume that error correlations between these *K* models are zero (setting all non-diagonal
167 terms in (4) to zero) we have

168 $$\mathbf{A}' = \begin{pmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_K^2 \end{pmatrix} \quad (5)$$

169 so that using (5) in (3), the weight for model *k* is proportional to $1/\sigma_k^2$, the inverse of model *k*'s
170 error variance. That is, assuming zero error correlation leads to optimal weights based entirely on
171 relative differences in model performance.
172
173 While we will spend some time in Section 4 explaining why independent models in an ensemble
174 should *not* have zero error correlation, (5) illustrates that this minimization of error (or 'difference')
175 problem may be viewed as having has a solution in two parts: that related to the 'performance'
176 differences of each model (the diagonal terms of **A**) and that related to the level of covariance
177 between the errors of the models (the non-diagonal terms of **A**). This, we suggest, provides a natural
178 choice for an empirical definition of model dependence. The weights given in (3), therefore,
179 optimally weight models for dependence and differences in performance with respect to MSD for
180 the in-sample period.
181
182 A simple idealized example illustrates how important model dependence defined in this way can be
183 to the performance of the multi-model mean. First, in Electronic Supplementary Material (ESM) A
184 we show that the expected error variance of $\mu_e$ (from (1), the optimally weighted ensemble mean)
185 is given by

$$s_m^2 = \frac{\sum_{j=1}^{J}\left(\mu_e^j - y^j\right)^2}{J-1} = \frac{1}{\left(\mathbf{1}^T\mathbf{A}^{-1}\mathbf{1}\right)}. \quad (6)$$

Now consider an ensemble of $K$ models that all have error variance equal to 1, with error covariance defined by powers of a single correlation parameter, $-1 \le \rho \le 1$, so that

$$\mathbf{A}_\rho = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^K \\ \rho & 1 & \rho & \cdots & \rho^{K-1} \\ \rho^2 & \rho & 1 & \cdots & \rho^{K-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^K & \cdots & \rho^2 & \rho & 1 \end{bmatrix} \text{ and } \mathbf{A}_\rho^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho & 0 & \cdots & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 \\ 0 & -\rho & 1+\rho^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & -\rho & 1 \end{bmatrix} \quad (7)$$

(chosen so that $\mathbf{A}_\rho$ had a simple inverse). Since $\rho$ raised to some positive integer power is always less than $\rho$, it is clear that the inter-model error covariances are smaller for those elements far from the diagonal of $\mathbf{A}$ than those close to the diagonal. Since $\left(\mathbf{1}^T\mathbf{A}^{-1}\mathbf{1}\right)$ is simply equal to the sum of all the elements in $\mathbf{A}^{-1}$, it follows that using (7) in (6) gives

$$s^2 = \frac{1-\rho^2}{(K-2)\rho^2 - 2(K-1)\rho + K} \quad (8)$$

Figure 2 shows the values of $s^2$ for a $K$=5 member ensemble as a function of $\rho$. For $\rho = 1$ the error covariance between all models is 1, so that the error variance of the optimal combination of models is the same as that of a single simulation – the models are identical. For perfectly uncorrelated errors, $\rho = 0$, the error variance of the optimally weighted mean is $1/K = 0.2$ of the error variance of an individual simulation. This result has serious implications for ensemble interpretation.

If, for example, one subscribed to the 'truth-plus-error' paradigm (e.g. Tebaldi et al., 2005; Greene et al., 2006; Furrer et al., 2007; Smith et al., 2009), and believed that a 'perfect' model should match observations plus a noise term, then model independence is naturally defined as pair-wise zero error correlation between models (as we would define independence of random variables, or noise). This is the $\rho = 0$ case of Equation 8, which would mean our estimate of error variance for the mean of $K$ independent models would then be 1/$K$. This in turn would imply that the only barrier to perfect prediction of climate at *any* timescale is the number of independent models

available to us – that error vanishes as the ensemble grows very large. In Section 4 we explain in more detail why we believe the truth-plus-error is inappropriate for climate prediction and suggest an alternative approach.

While it also appears that the perfectly anti-correlated example, $\rho = -1$, can result in a zero error variance weighted mean, there are likely stricter bounds on the range of possible error correlations than we have imposed in this idealized example, depending on the size of the ensemble (we will explore this more in Section 4). Remember all of these 'models' perform equally well, in the sense that they have equal error variances. Equation (8) and Figure 2 serve to demonstrate that, all other things being equal, lowering the error correlation between ensemble members increases the utility of an ensemble prediction because it lowers the error variance of the ensemble mean.

**3a. A simple application of dependence weights**

To get a sense of how this weighting technique behaves, we now apply it to a very simple synthetic example. Suppose observations of a variable of interest were given by the function $y(t) = t + 15\sin(t/6)$ and two model simulations of this variable were given by

$$x_1(t) = t/1.98 + 15\sin(t/4) \quad \text{and} \quad x_2(t) = 1.5t + 18\sin(t/8),$$

applied to 170 discrete time steps (chosen so that trends and oscillations are visible). After bias correction, the 'observations' and these two models have the same mean. Their time series are shown in Figure 3a by the black, blue and red curves. These two models were chosen as they have almost identical MSD (839 units$^2$) – their 'performance' is the same. In this case, the two-member ensemble mean and weighted mean (using (3)) have identical MSD (around 161 units$^2$). If we now add three additional models by simply adding noise terms (Gaussian, standard deviation 6 units) to the first model, $x_1$, we have a five-member ensemble with four dependent members (shown in Figure 3a by the grey lines). In this case, the ensemble mean MSD is around 409 units$^2$ and the weighted ensemble mean MSD is around 159 units$^2$ (these are mean MSD values from 1000 independently generated 5-model ensembles of the type described above). An example is shown by the green and gold curves in Figure 3a, respectively.

The weighting technique preserves the performance of the mean of the original independent-member ensemble while the performance of the multi-model mean is clearly degraded. Across the 1000 different 5-member ensembles, the weight given by (3) to $x_2$, the second model, is almost

constant at 0.5 (variance of the weight value is around 0.0001). Consequently, the sum of the weights of the *dependent* models 1,3,4 and 5 is also 0.5.

Next, while there is some minor performance gain in the weighted mean from the addition of the noisy members of the ensemble (due to the small sample size), this example illustrates a key point when using this weighting technique where the relative dependence of the ensemble members is not known. The performance gain of the weighted mean over the mean in the case when there is dependence in the ensemble is *not* due to the weighting technique fitting noise (issues of sample size excepted, of course). Rather, the MSD of the weighted mean approximates the MSD of the mean of the effectively independent members.

In Section 5 we will modify this weighting approach to ensure that as well as all weights summing to unity, no model receives a negative weight. We note here that model weights obtained from particle filter approximations to Bayes' theorem (van Leeuwen, 2011; Snyder et al., 2008) share this feature. However, as will become clear, both our objectives and method differ from that of a particle filter.

**3b. Application of dependence weights to the CMIP3 ensemble**

We now apply the weighting technique to the 24 fully coupled global CMIP3 models (Meehl et al, 2007a) in the PCMDI online database (http://www-pcmdi.llnl.gov/ipcc/about_ipcc.php) that constitute the basis for the IPCC Fourth Assessment Report's climate projections (Meehl et al, 2007b). We use the HadCRUT3 monthly surface air temperature dataset (Brohan et al, 2006) and compare the two for the years 1970-1999. All model simulations are interpolated to the HadCRUT3 5° x 5° spatial resolution, and only one simulation from each model was used. This provides us with 24 residual time series of length 360 (30 years x 12 months) for each of the globe's 72x36 grid cells at this resolution. Any grid cells with more than 20% of observational data missing during this period are omitted. Figure 4 shows the RMSD at each grid cell of (a) the multi-model mean of the CMIP3 ensemble, and (b) the weighted mean of the CMIP3 ensemble using the weights given in (3), averaged across 30 out-of-sample tests (detailed below). The weighting clearly offers considerable reductions in RMSD globally, but several regions in particular show marked changes (e.g. China, South America, Northern Europe and Africa).

Figure 5 shows that the difference between Figures 4a and 4b is comprised of three incremental stages: bias-correction, performance weighting and ultimately dependence weighting. Each curve in

277 Figure 5 represents the density of global RMSD values for different model weighting techniques
278 across 30 out-of-sample tests. These tests use a bootstrapping approach – for any given weighting
279 technique, 29 years are used to define a set of model weights for and the remaining year from our
280 30 years of data is used to test them. The process is then repeated for all 30 possible testing years.
281 As there are only 30 tests for each weighting technique, the 30 RMSD values from these tests are
282 fitted to normal distributions in Figure 5. For each weighting technique, results are plotted for both
283 a 'global' configuration (involving a single weight for each model across all 72x36 grid cells,
284 considering each grid cell as an independent piece of information) and a 'per-cell' application
285 (allowing 72x36 separate weights for each model). For reference, the density of RMSD values for
286 the original CMIP3 ensemble mean is shown in black.
287
288 First we consider bias correction. Using a historical evaluation period to remove model biases is
289 common in weather forecasting (e.g. Glahn and Lowry, 1972; Wilson and Vallée, 2002) and
290 climate research (Meehl et al, 2007b; Reifen and Toumi, 2009). The dark blue and light blue curves
291 in Figure 5 show the global RMSD of the mean of the bias-corrected models, applied globally and
292 at each grid cell, respectively. As expected, this improves the multi-model mean RMSD, especially
293 in the 'per-cell' application.
294
295 Next, the green curves in Figure 5 show global RMSD density when bias-corrected ensemble
296 members are linearly combined using the performance weights from (3) when **A** is given by (5). As
297 we would expect, in both the global (dark green) and per-cell (light green) cases, performance
298 weighting gives additional skill above bias correction. Finally, we consider the optimal dependence
299 weights given by (3). These are represented in Figure 5 in the global and per-cell cases by the red
300 and orange curves, respectively, and clearly give the lowest global RMSD values.
301
302 The clearest feature of Figure 5 is the separation of 'global' and 'per-cell' weighting strategies.
303 Next, while there are many published examples of model performance weighting (e.g. Krishnamurti
304 et al, 2000; Giorgi and Mearns, 2002; Tebaldi et al, 2005) and apparently none of dependence
305 weighting, the improvements in performance provided by the dependence weights here are
306 significantly larger than those provided by performance weights alone. Lastly, we highlight the grey
307 curve in Figure 5, a persistence-like case, using the mean of observed values each month across the
308 29-year training period to predict that month in the testing year. We see that the dependence-
309 weighted ensemble performance is in fact superior to the average of 29 month-specific observed

temperatures at each grid cell - a remarkable achievement for an ensemble of 100-150 year predictions at coarse resolution.

We again wish to reassure the reader, despite the analysis above, that we do not expect a climate model to match monthly temperature time series. While it may seem that the weighting approach we outline is simply likely to favour those models whose internal variability happens to better match observations in the in-sample period, something of a noise fitting exercise, we now show that the result is stable across the 30 year period under consideration. Table 1 shows the global RMSD values for weights trained on one decade during this period and tested on another. While the in-sample periods (e.g. train weights on 1990s and test weighted mean on 1990s) clearly always show the best results, application to out-of-sample decades show that the weights are remarkably stable and remain effective. While it may be tempting to suggest that 30 years is too short a period to illustrate that the weights perform well out-of-sample, we note that 30 years is the World Meteorological Organization's reference length for defining climate as opposed to weather (http://www.wmo.int/pages/prog/wcp/ccl/faqs.html). While we fully acknowledge the potential for longer period climate cycles to cloud this result, the limited availability of observational data ensures this will always be a caveat. In Sections 4 and 5 we describe why, despite the lack of an expectation of a model's agreement with monthly time series observations, the dependence weighted time series is key to understanding ensemble spread.

## 4. Model independence, climate system uncertainty and the replicate Earth paradigm

In Section 2 we noted in a simple example that if we assumed pair-wise model error correlation to be zero, optimal model weights (with respect to MSD) were entirely proportional to differences in model performance (error variance in this case). Intuitively, we might therefore want to insist upon independent model simulations as having pair-wise zero error correlation – as the standard statistical definition, $f(x_1,x_2)=f(x_1)f(x_2)$, would suggest for two independent random variables. But is this appropriate for error in climate models? Below we discuss consequences of this approach and argue for an alternative conception of a multi-model ensemble that naturally yields a positive value for the expected error correlation of independent simulations.

Firstly, a zero-error-correlation definition of independence implies that a 'perfect', independent model should reproduce observed data plus an independent noise term. Equivalently, the error of the mean of an ensemble of perfectly independent models should converge to zero as the ensemble size grows large. Observational data in this case would always be the mean of a distribution of an

ensemble of independent models. This view of a 'perfect ensemble prediction' amounts to a "truth-plus-error" paradigm of interpretation (e.g. Tebaldi et al, 2005). While it could be argued that very long time averages of observations that smooth out, for example, decadal oscillations, are predictable, high impact weather events likely to be affected by climate change such as droughts, flooding and Tropical Cyclones clearly have limited predictability. Annan and Hargreaves (2010) noted the inappropriateness and prevalence of this truth-plus-error paradigm in climate model evaluation and proposed instead using an 'indistinguishable' paradigm of interpretation which assumes *a priori* that the observations and all models belong to the same distribution. Here, we present a third paradigm of interpretation – the 'replicate Earth' paradigm. We introduce the idea by noting a critical and hitherto unmentioned assumption of the truth-plus-error paradigm – that the climate system and all the processes that affect it are entirely (i.e. deterministically) predictable from climate forcing variables. That is, since the truth-plus-error paradigm assumes a zero-error-correlation definition of independence, the only barrier preventing a perfect prediction by the mean of an ensemble at *any* time scale is the number of independent members it contains. This is clearly inappropriate for sub-decadal time scale events and may be inappropriate for much longer time scale events as well. This is a consequence of defining the observed time series to be the centre of the distribution of an ensemble of independent models – the ensemble mean will converge to the observations as the ensemble grows large.

In contrast to the truth-plus-error viewpoint, the replicate Earth paradigm accommodates the possibility that there may be inherent limits to the predictability of the atmosphere and ocean at any time scale. Suppose there were a very large number of Earth replicates that experienced immeasurably similar climate forcing (e.g. orbital, solar, greenhouse gas forcing) to our own Earth. Each replicate Earth would have different instantaneous realisations of atmospheric and ocean states as a result of the climate system's chaotic processes. Imagine that the behaviour across this very large number of Earth replicates defined time-evolving Climatological Probability Density Functions (CPDFs) that defined the probability of the occurrence of particular ranges of climate variables or even particular types of events. If the climate were constant, we could approximate the CPDF using historical data, but this is clearly not the case, especially when $CO_2$ concentrations are rapidly increasing. It is therefore impossible to empirically determine the properties of the CPDF in the presence of changing climate. This, we suggest, is the role of climate models.

Climate models can be viewed as imperfect attempts to create replicate Earths. We suggest that an ideal ensemble prediction would be comprised of replicate Earths that were independent and

identically distributed (IID) draws from the CPDF defined by a very large ensemble of replicate Earths. In this ideal, the models/replicate Earths comprising the ensemble prediction are independent because they are independently drawn from the CPDF – not because of zero error correlation. These models are perfect because they behave like replicate Earths but their distance from the observations on our Earth (that is, their error) is not zero. This distance has a strict lower bound determined by the inherent variance of the range of states permitted by a particular set of climate forcing conditions. The chaotic nature of atmospheric and oceanic flow causes the trajectories of two replicate Earths in almost identical states to diverge with time and ultimately to be statistically indistinguishable from independent random draws from the CPDF. In this sense, 'perfect' and 'independent' models are essentially synonymous. The mean of an ensemble of perfect models is therefore simply an approximation of the mean of the CPDF, and since the real Earth itself is also a random draw from the CPDF, we should *not* expect observations of it to match this mean, but rather be equivalent to a different perfect model. Unlike the truth-plus-error paradigm, we should not expect the error of the mean of an ensemble of replicate Earths (with respect to our Earth's observations) to tend to zero as the ensemble size increases.

Figure 1, taken from the IPCC AR4, seems to broadly support the concept behind the replicate Earth paradigm. It shows global mean surface temperature, expressed as an anomaly, for an ensemble of climate models (shown in yellow), their mean (shown in red) and the observational record (shown in black). The observational record seems much more like an individual model than the multi-model mean: it is the most extreme value on a few occasions and has model-like variability. Knutti et al (2010b) note that the CMIP3 ensemble mean error converges to a large non-zero value, also supporting the replicate Earth concept.

We can in fact show that the anticipated level of error covariance between two (perfectly independent) replicate Earths (that is, the off-diagonal $\mathbf{c}_{i,j}$ in (4)) is $\overline{\sigma_r^2}$, the time average of the instantaneous CPDF variance (see ESM B for derivation). To visualise this quantity, imagine in Figure 1 determining the variance of the yellow lines at a single point in time (about the red line), and averaging this for all time steps. We can also show that the MSD of the mean of a $K$-member replicate Earth ensemble to "our Earth" is $\overline{\sigma_r^2} + \overline{\sigma_r^2} / K$ (see Equation B8), so that as the ensemble size becomes infinite the MSD of its mean would converge to $\overline{\sigma_r^2}$.

Annan and Hargreaves (2010, 2011) suggest that the appropriate paradigm for ensemble interpretation is one that assumes that climate models and observations are drawn from the same distribution (the indistinguishable paradigm). If each climate model were perfect (i.e. a replicate Earth) then we would agree with this approach. The fact that their indistinguishable paradigm *assumes* that models are replicate Earth-like is reinforced by a result we derive in ESM B – the anticipated error correlation of replicate Earths is 0.5, precisely the estimated correlation presented in the appendix of Annan and Hargreaves (2011). In the next section we demonstrate that the CMIP3 ensemble is in reality *not* replicate Earth-like, and derive a transformation process to bring it closer to being so.

## 5. Transformation to a replicate Earth-like ensemble

We now explore the extent to which the current generation of climate models, that is the CMIP3 ensemble, behaves like a replicate Earth ensemble. To do this, we identify two key properties of an ensemble of replicate Earths. When trying to visualize these properties, it may be helpful to refer to Figure 1.

1. **The equally weighted mean of an ensemble of replicate Earths is *the* linear combination of replicate Earths that minimizes the distance from our Earth's observations over an extended time period.**

That is, the best estimate (in terms of mean square distance) of any particular replicate Earth (a random draw from the CPDF) will be the mean of the CPDF. (Of course, the random draw from the CPDF that is of particular to us is the real Earth). Note that this property follows directly from the statistical fact that the entity that minimizes the expected squared distance from individual realizations of any probability distribution is the mean of the distribution.

2. **The time average of the instantaneous CPDF variance should be approximately equal to the variance of the real Earth about the CPDF mean over time.**

This is essentially saying that the variance of the real Earth about the CPDF mean should be the same as the variance of all the other replicate Earths about the CPDF mean. We could equivalently phrase this as '*the time average of the variance of an ensemble of replicate Earths should be approximately equal to the MSE of the replicate Earth ensemble mean (with respect to the real Earth's observations)*'. This property holds for all ensemble predictions that represent the distribution of truth (Leutbecher and Palmer, 2008). One could imagine in Figure 1 determining the variance of the yellow lines at a single point in time (about the red line), and averaging this for all time steps. Our assertion simply states that this should be roughly equivalent to the variance of the black line about the red line (as though the red line were the

444 CPDF mean). To understand this assertion, note that if the CPDF were not changing in time, the
445 ergodic assumption would be valid and the distribution of observations from a single Earth (e.g.
446 our Earth) would precisely define the CPDF: the variance of the observations about the mean of
447 the observations would be precisely equal to the variance of the CPDF. However, if the
448 trajectory of CPDF variance were a strongly non-linear function of time then the time average
449 of CPDF variance would only be approximately equal to the variance of the real Earth about the
450 CPDF mean over time. Since we do not know *a priori* how the CPDF variance will be affected
451 by increasing $CO_2$ concentrations and a changing climate, the possibility of fluctuations in
452 CPDF variance must be allowed for. (Indeed, the study of Schär et al. (2004) suggests that the
453 variance of the CPDF of European summertime temperatures will increase with increasing $CO_2$
454 concentrations.)
455
456 In ESM B, we provide a mathematical proof that properties 1 and 2 would be satisfied by an
457 ensemble of replicate-Earths (or equivalently by an ensemble of long simulations from models that
458 perfectly represented physical processes from the nanoscale to the global scale but which had
459 differing, equally plausible initial conditions).
460
461 It should be immediately clear that the CMIP3 ensemble does not satisfy property 1. We showed in
462 Section 2 that an optimized linear combination of models performs significantly better than the
463 multi model mean in out of sample tests. We can also show that property 2 is not satisfied. That is,
464 the instantaneous variance of monthly surface temperature in the CMIP3 ensemble, averaged over
465 all months 1970-1999, is quite different from the variance of the observations about the multi-
466 model mean (equivalent to the error variance of the multi-model mean). We will see clear evidence
467 of this when discussing Figure 6 shortly.
468
469 While any inference of CPDF properties in the presence of changing climate can only be based on
470 model predictions (since we only have one real sample, our Earth), we can extract better estimates
471 of them than the original CMIP3 ensemble provides, by using the two properties above. We know,
472 for example, that our linear combination of models from Section 2, $\mu_e$, is a better candidate for the
473 CPDF mean than the multi-model mean, since it is the minimum error variance estimate we can
474 have for our set of model simulations (i.e. best estimate to the replicate Earth that is the real Earth).
475 Also, if we could interpret each model's weight from (3) as the probability that that particular
476 model were a replicate Earth, we could estimate instantaneous CPDF variance using

$$\sigma_e^{2j} = \sum_{k=1}^{K} w_k \left( x_k^j - \mu_e^j \right)^2 \tag{9}$$

as though our variable of interest (in this case monthly surface temperature) were a discrete random variable. Using these two pieces of information, we now present a technique to transform the raw CMIP3 ensemble into a more replicate Earth-like ensemble.

To begin, we note that model weights obtained from (3) are not necessarily positive and may therefore also be greater than 1 (since they sum to 1). Ensuring their positivity, however, is not possible without modifying the original models' time series. But we can modify them in such a way that their weighted combination still provides our minimum error variance estimate, and CPDF mean estimate, $\mu_e$. Recalling that the value of the $k^{\text{th}}$ ensemble member at the $j^{\text{th}}$ observation time is given by $x_k^j$ and that the vector listing all $K$ of these ensemble values is denoted by $\mathbf{x}^j$, one can proceed by expressing $x_k^j$ in terms of a perturbation from the multi-model mean $x_k^j = \overline{x}^j + x'^j_k$ where $\overline{x}^j = \frac{1}{K}\sum_{k=1}^{K} x_k^j$, then we can express $\mu_e$ as

$$\mu_e^j = \mathbf{w}^T \mathbf{x}^j = \tilde{\mathbf{w}}^T \mathbf{z}^j \tag{10}$$

where the $k^{\text{th}}$ element of the $k$-vector $\mathbf{z}^j$ is given by $z_k^j = \overline{x}^j + \alpha x'^j_k$

$$\tilde{\mathbf{w}} = \frac{\left( \mathbf{w}^T + (\alpha - 1)\frac{\mathbf{1}^T}{K} \right)}{\alpha} \tag{11}$$

and $\alpha = 1 - K\min(w_k)$ where $\min(w_k)$ is the lowest (most negative) of the preliminary weights obtained from (3). A complete derivation of (10) and (11) is provided in ESM C, where we also show that the $\tilde{\mathbf{w}}$ weights still sum to 1 and are now all positive.

With this transformation of model time series and weights, we are now in a position to interpret the $\tilde{w}_k$ as probabilities in (9) with $z_k^j$ in place of the $x_k^j$. This gives us an estimate of the CPDF variance. With the weights interpreted as relative probabilities of occurrence, the formal definition of variance becomes $\sigma_e^{2j} = \sum_{k=1}^{K} \tilde{w}_k \left( \tilde{x}_k^j - \mu_e^j \right)^2$ where $\tilde{x}_k^j$ is a soon to be defined modification of $z_k^j$.

From property 2 above, for the ensemble to be replicate Earth-like, the time average of this variance must be equal to the variance of our observations about the CPDF mean estimate, $\mu_e$. That is, property 2 requires that

$$\frac{1}{J}\sum_{j=1}^{J}\sigma_e^{2j} \;=\; s_e^2 \;=\; \frac{\sum_{j=1}^{J}\left(\mu_e^j - y^j\right)^2}{J-1}. \qquad (12)$$

We can ensure (12) by further modifying our model time series, this time to change the ensemble variance, by letting

$$\tilde{x}_k = \mu_e + \beta\left(\overline{x} + \alpha x'_k - \mu_e\right) \qquad (13)$$

where

$$\beta = \sqrt{\frac{s_e^2}{\frac{1}{J}\sum_{j=1}^{J}\sum_{k=1}^{K}\tilde{w}_k\left[\left(z_k^j - \mu_e^j\right)\right]^2}}. \qquad (14)$$

A more detailed mathematical justification for (13) and (14) is provided in ESM C. Critically, it also contains a proof that this new transformation still preserves the CPDF mean estimate

$$\mu_e^j = \mathbf{w}^T\mathbf{x}^j = \tilde{\mathbf{w}}^T\mathbf{z}^j = \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}^j \qquad (15)$$

and that (12) holds with

$$\sigma_e^{2j} = \sum_{k=1}^{K}\tilde{w}_k\left(\tilde{x}_k^j - \mu_e^j\right)^2. \qquad (16)$$

In summary, (11) and (13) give us a weighted ensemble that satisfies both properties 1 and 2 provided that $\tilde{w}_k$ is interpreted as the *relative* probability that the transformed model $\tilde{x}_k^j$ is a replicate Earth. We say "relative probability" rather than "actual probability" because these weights only depend on the relative performance of the models rather than their absolute performance.

Figure 3b gives a visual indication of how this transformation process affects the simple ensemble we examined earlier in Figure 3a, and shows the value of $\alpha$ and $\beta$ used for the transformation. Recall from above that single values of $\alpha$ and $\beta$ scale all models' deviation from the multi-model mean and weighted mean, consecutively. The result is a transformed ensemble whose variance about the weighted mean, $\mu_e$, approximates the variance of the observations about $\mu_e$. Note also that the strong anomalous trends shown in model 1 and model 2 are reduced as a result. Perhaps most importantly, the mean of this transformed ensemble (green line in Figure 3b) is now much closer to the weighted mean of Equation 15. That is, the equally weighted mean of these models is much closer to being the best estimate of the observations (Property 1 above) – the models have become more replicate Earth-like.

To show that the transformed CMIP3 ensemble is more replicate Earth like, we examine whether the CPDF that the observations (our Earth) are drawn from is the same as the CPDF implied by the weighted transformed models. The CPDF implied by the weighted transformed ensemble can be sampled by randomly sampling the transformed ensemble members with frequencies that are proportional to the weights $\tilde{\mathbf{w}}$. However, before considering the CPDF implied by the weighted transformed ensemble, it is of interest to assess the CPDF that would be implied if the 24 transformed ensemble members had been evenly weighted. To do this, we consider a histogram of the rank of the observed value of monthly surface temperature amongst all modelled values on the real line. For example, if for a particular month at a particular grid cell, the observed temperature is hotter than all 24 models, we increment the histogram bin associated with the 25$^{th}$ rank by one. After repeating this process for all grid cells and all months of data in each of the 30 out-of-sample tests described in Section 3, we have a 25-bin rank frequency histogram (RFH), as shown in Figure 6. It shows histograms for the original CMIP3 ensemble (black), the ensemble of bias-corrected models (blue) and also the ensemble of transformed models (red) given by (13). If the observations and models are from the same distribution, the histogram should be flat. Both the original and bias corrected ensembles populate the central ranks much more than the extreme ranks – they are 'over-dispersive' (a sign that the ensemble is unlikely to be over-confident in its *range* of predictions). The histogram for the evenly weighted transformed ensemble members (red) is clearly much flatter thus indicating that the CPDF implied by the assumption that all of the transformed ensemble members are equally likely is closer to the true CPDF than that implied by assuming that all of the original bias corrected ensemble members are equally likely (the blue curve). This can also be seen in the example shown in Figure 3b – the mean of the transformed ensemble is much closer to the CPDF mean estimate, $\mu_e$, than the mean of the original models.

We note that our interpretation of ensemble spread being over-dispersive (supporting the findings of Annan and Hargreaves, 2010) is somewhat at odds with Jewson and Hawkins (2009), who argue that the ensemble spread is too narrow. The ensemble transformation they present is therefore intended to inflate ensemble spread, and this is done as a function of correlation in model projections (rather than focusing on correlation of model errors, as we do here).

The assumption that the transformed ensemble members are equally likely is obviously inconsistent with properties 1 and 2 of replicate Earth ensembles. These properties are only satisfied when the weights/probabilities $\tilde{\mathbf{w}}$ are applied to the transformed ensemble members. We now present a method to combine the transformed time series and weight information that borrows heavily from

565 the resampling methods used in particle filters (van Leewen, 2009). In doing so, we obtain
566 ensembles whose evenly weighted sample mean is identical to (15) and whose evenly weighted
567 variance is identical to that given by (16). The procedure also suggests an "effective" number of
568 independent models in the CMIP3 ensemble.
569
570 Since we know that our 24-member ensemble shows more dependence than a replicate Earth
571 ensemble, we cannot expect it to make probabilistic predictions with the same level of accuracy as a
572 24-member ensemble of replicate Earths. We therefore create ensembles of varying sizes by
573 randomly sampling the 24 transformed members $\tilde{x}_k$ with frequencies related to their corresponding
574 weights $\tilde{w}_k$. In doing so, we want an ensemble that:

575 1) has sample mean $\mu_e^j$ - the linear combination of ensemble members minimizing error
576 variance;
577 2) has time averaged variance equal to the time averaged error variance of the ensemble mean
578 $s_e^2$; and
579 3) has a flat rank histogram, suggesting that the ensemble members are drawn from the same
580 CPDF as the observed data set.

581 To create an ensemble of size $M<K=24$, we:

582 I. Divide the unit interval into $K$ sections using

583 $$\left[0,\ \tilde{w}_1,\ \ldots,\ \sum_{i=1}^{k}\tilde{w}_i,\ \ldots,\ \sum_{i=1}^{K-1}\tilde{w}_i,\ 1\right] \qquad (17)$$

584 Then, for each required prediction (in this case at each grid cell for each month), randomly
585 select $M$ uniformly distributed random numbers in the unit interval [0:1] and select the k$^{th}$
586 ensemble perturbation $\left(\tilde{x}_k^j - \mu_e^j\right)$ when a random number falls in the k$^{th}$ interval of (17).

587 II. Remove the mean of the $M$ selected perturbations (so that at any point in time, the
588 perturbations sum to zero).

589 III. Multiply these perturbations by $\sqrt{M/\left(M-1\right)}$ to ensure that the removal of the ensemble
590 mean does not change the variance.

591 IV. Add the resulting perturbations to the minimum error variance estimate $\mu_e^j$.

592 This procedure gives us an $M$ member ensemble whose mean is precisely equal to $\mu_e^j$ and whose
593 time averaged mean square deviation about the mean is be precisely equal to $s_e^2$ - the time averaged

error variance of $\mu_e^j$. While this ensemble is neither over-dispersive nor under-dispersive under this second moment measure, we have not yet constrained higher order moments of the ensemble distribution and histograms are only guaranteed to be flat if all moments of the ensemble distribution are identical to those of the distribution from which the observation is drawn. These higher moments are likely to be affected by the number $M$ of randomly selected members. When $M$ is close to the original ensemble size $K$, it is extremely likely that the random selection procedure will select the same member more than once – particularly if one of the ensemble weights is much larger than the others. Such repeated selection of the same member will clearly affect the 3rd and 4th moments of the ensemble distribution.

Letting $M$ be much smaller than $K$ reduces the chances of selecting the same member more than once and hence lessens its possibly deleterious effects on the higher moments of the ensemble distribution. A disadvantage of decreasing $M$ is that our requirement that the $M$ perturbations sum to zero increases its impact on the ensemble distribution's 3rd and 4th moments as $M$ is decreased.

With these facts in mind, it is unreasonable to expect the histogram to be flat for all sub-selected ensemble sizes $M$. However, since a flat histogram enables impact assessments that could not otherwise be performed, it is of great interest to find an $M$ value that has an approximately flat histogram. We considered both the per-cell weights and the global weights, and examined histograms for values of $M$ between 3 and 24, with 4 to 9 shown in Figure 7.

For the per-cell weights – the most accurate configuration – it was found that the exterior bins were over populated for $M>5$ and underpopulated for $M<5$ (see orange lines in Figure 7). Note that such departures from flatness *do not* indicate under or over dispersion under a 2nd moment measure. This is because, by construction, the mean square deviation of the ensemble members about $\mu_e^j$ is precisely equal to the mean square error $s_e^2$ of $\mu_e^j$. Hence, in this case, the overpopulation of the extreme ranks must be associated with a mismatch between the 3rd, 4th and higher moments of the ensemble distribution and the true distribution. A similar investigation is also shown in Figure 7 for the *global* perturbed model case (red lines in Figure 7). They show that the extreme ranks are overpopulated for $M$=10 (not shown) and underpopulated for $M$=8. Although the population of the extreme ranks for $M$=9 is similar to that of the interior ranks, ranks 2 and 9 are a little underpopulated. Nevertheless, as is indicated by the blue and black dashed curves, the RFHs associated with the raw and bias corrected ensembles are less flat than that delivered by our method

627 for the $M$=9 case. We note that both of these estimates of the effective number of independent
628 climate models in the CMIP3 ensemble (5 and 9) are within the range of existing estimates (Pennel
629 and Reichler, 2008; Annan and Hargreaves, 2011).
630
631 To assess whether the improvement to the RFHs is because of our ensemble transformation
632 procedure and not just due to randomly selecting a smaller ensemble size, we randomly selected
633 just $M$ of the raw, globally bias corrected and per-cell bias corrected CMIP3 ensemble members
634 using a set of 24 weights all equal to 1/*24* and computed the resulting RFHs. These are shown by
635 the black, blue and green dashed curves respectively in Figure 7. As expected, these curves show
636 that the extreme ranks of the global and per-cell cases are under and over populated, respectively.
637
638 The flatter RFHs associated with the replicate-Earth like ensembles obtained by resampling the
639 weighted transformed ensemble shows that the relative frequency of events in this replicate-Earth
640 like ensemble is more likely to be related to the probability of their occurrence in the real world,
641 making the ensemble better suited for use in quantitative societal/economic/ecological impact
642 models. These pseudo-replicate Earths provide an estimate of the CPDF that accounts for
643 performance and dependence differences between models, as well the rescaling the ensemble
644 variance to be closer to the variance of the observational record about the CPDF mean.
645
646 **6. Discussion and conclusion**
647 We have introduced a new way to interpret multi-model climate ensembles, the replicate Earth
648 paradigm, that offers a justifiable approach to including inherent climate system uncertainty at all
649 timescales when evaluating ensemble performance. It is significantly different conceptually from
650 the two prevailing ensemble interpretation paradigms – the so called "truth plus error" and
651 "indistinguishable" paradigms. We outlined two key properties of a replicate Earth ensemble and
652 showed that the current generation of climate models are not replicate Earth-like. We then derived a
653 transformation process to make a given ensemble more replicate Earth-like, maximizing its
654 predictive ability in that the resulting ensemble mean provides the best estimate to observations and
655 the resulting ensemble variance becomes a reliable predictor of the error variance of the ensemble
656 mean.
657
658 The technique yields a positive weight for each transformed ensemble member that can be
659 interpreted as the relative probability that the transformed ensemble member is a replicate Earth. By
660 randomly resampling the transformed ensemble members with frequencies given by their weights,

evenly weighted ensembles were produced with flat rank frequency histograms and low error variance means. An ensemble with a flat rank frequency histogram and small ensemble mean error variance has a much better chance of accurately predicting changes in frequencies of weather events than one with a larger ensemble mean error variance and a non-flat rank frequency histogram. Hence, the resampled transformed ensemble is better suited to quantitative assessments of climate change impacts than the original ensemble.

One question that we have not answered is whether *climate change itself* will lead to model error covariances that are significantly different to those associated with the HadCRUT3 data set (that is, the extent to which historical data is representative of the future system). We note, however, that this issue equally applies to bias-corrections derived from historical data, which appear well accepted by the community and are prevalent in IPCC representations of climate projections. Note that for the purposes of this discussion, we have also assumed that the sample size provided by historical data is large enough to rule out spurious fluctuations in the weights associated with too small a sample size (see Weigel et al, 2010 for examples of issues this may cause). We also reiterate that the inferences we've made about the CPDF are entirely model based. This is of course unavoidable. They may well change markedly as models improve.

In deciding on how best to apply this approach to future projections, data availability for the variables of interest would likely determine whether the per-cell or global application is more appropriate, noting the issue of sample size discussed above. There also may be utility in simply using the rescaled ensemble described in Equation (13) without the resampling process described above, although this requires further investigation. Further work should also consider how the method we have presented might be best extended to multiple climate model variables (Gleckler et al., 2008).

We also demonstrated that even in a simple optimization problem, accounting for *both* model performance differences *and* model dependence is critical to extracting the most predictive ability from an ensemble, regardless of whether one subscribes to the replicate Earth or truth-plus-error paradigm. The weighting technique outlined in Section 2 provides a justifiable way to weight multi-model ensembles where some models may be represented by many simulations and others by only a few – an issue that will face those interpreting the CMIP5 ensemble. The weights reflect each simulation's contribution to the overall predictability of the entire ensemble. This suggests that to

694 the extent that adequate resources are available, a *diversity* of skillful climate model types must be
695 encouraged to improve ensemble predictive ability.
696
697 While our example showed that error correlation provides a natural definition for model
698 dependence, the level of correlation associated with "statistical independence" depends on the
699 assumptions inherent in the ensemble interpretation paradigm. We showed that the 'truth-plus-
700 error' paradigm leads to the rather counterintuitive conclusion that the mean of an infinite ensemble
701 of models with zero inter-model error correlation would be equal to the truth at all time and length
702 scales. We also discussed the 'indistinguishable' paradigm of Annan and Hargreaves (2010), noting
703 that ensembles of error prone models have statistical properties that strongly distinguish them from
704 replicate-Earths or 'perfect' models. We suggested that while the indistinguishable paradigm is
705 justifiable with replicate Earths, it is not with today's climate models.
706
707 The replicate Earth paradigm is not an entirely new idea. We already accept that climate models
708 should not be able to reproduce weather – that weather is partially chaotic. There has, however,
709 been an implicitly assumed time constant in most climate research (perhaps 30 years or so) beyond
710 which we have by and large assumed that climate is *entirely* predictable. If this were true then the
711 CPDF of 30-year averages of variables from replicate Earths would have zero variance across an
712 ensemble. If it were not true because of longer timescale modes of climate variability that were not
713 directly forced then 30 year variable averages from replicate Earths would not have zero variance.
714 In this way, the replicate Earth paradigm naturally accommodates both predictable and
715 unpredictable time averages.
716
717 Given this somewhat fluid interpretation of the distinction between weather and climate, we now
718 contrast the CPDF that we wish to extract from an ensemble of climate predictions with the
719 Weather PDF (WPDF) that forecasters attempt to infer from ensembles of 1-15 day weather
720 forecasts (Gneiting and Raftery, 2005, and references therein). The WPDF is the distribution of
721 possible weather trajectories over, for example, the next 15 days given the last 3-5 weeks of
722 atmospheric/ocean observations. In terms of replicate Earths, this distribution is defined by the
723 distribution of 1-15 day trajectories of all replicate Earths having the exact same set of (error prone)
724 observations over something like the preceding 3 weeks. In contrast, the CPDF is the distribution of
725 possible 1-1500 year trajectories of all replicate Earths having the same anthropogenic greenhouse
726 gas and aerosol forcing and approximately the same ocean heat content in the late 19$^{th}$ century. If
727 one were to extend the trajectories of the replicate Earths comprising the WPDF out to climate time

728 scales they would converge to the CPDF as chaotic processes caused the predictability associated
729 with knowledge of recent observations to be lost.
730
731 To test how well some ensemble of models approximates the CPDF, one needs a long time series of
732 observations such as the HadCRUT3 data set. In contrast, the ability of an ensemble of model
733 forecasts to approximate the WPDF can be tested with repeated realizations of relatively short time
734 sequences of observations (1-15 days). The transformation process described in this paper could be
735 applied equally well to multi-model ensembles of 1-15 day forecasts as it could to multi-model
736 climate predictions. However, since the time-averaged observations used for the climate application
737 are different to the instantaneous observations used in the weather forecasting application it is
738 possible that some models would receive small weights for the weather forecasting application and
739 large weights for the climate prediction application.
740
741 This raises the fact that one not only needs to consider the length of the observational record used to
742 derive the weights (e.g., 15 days for WPDF versus 30 years for CPDF), but also the degree of time
743 averaging applied to the observations. In our study, we chose to use monthly mean data over a 30
744 year period. We could equally well have used shorter observation averaging periods such as a week,
745 a day or even 1-hour averages. Alternatively, we could have used longer observation averaging
746 periods such as 3 monthly, annual or decadal averages. Further experimentation will be required to
747 determine the sensitivity of weights derived from our method to the observation averaging period.
748 Recalling that the model weights determined by our method are designed to provide a linear
749 combination of models that minimize the distance from observations and noting that high frequency
750 variations in the atmosphere are inherently unpredictable, we speculate that weights from our
751 method will be more sensitive to changes in the length of the observation period (15 day segments
752 versus 30 yr segments) than they would be to the observation time averaging period.
753
754 This speculation is based on the fact that by using a relatively small number of ensemble weights to
755 minimize the distance to a very large set of independent observations, the replicate Earth paradigm
756 not only anticipates model-observation mismatch in perfect models at *any* time and/or space scale
757 but also ensures that models that accurately capture long time scale trends receive more weight than
758 those that do not. It makes no assumption about a particular time constant that separates weather
759 and climate, only that the partially chaotic natural system has a spread of 'true' outcomes that
760 define the CPDF. Critically though, this interpretation does not imply that models provide little
761 information about the real world. On the contrary, by ameliorating the deleterious effect of

762 correlated model errors on ensemble predictions, it provides a transformed ensemble whose spread
763 better represents our uncertainty in prediction.
764
765 To avoid over-fitting the data, one must ensure that the number of observations far exceeds the
766 number of ensemble members for which weights are sought. In our examples, we sought weights
767 for 24 ensemble members. The number of observations used to estimate these weights was 902,016
768 and 348 in the global and per-cell examples, respectively. Delsole (2007) found that 22
769 observations of seasonally averaged temperatures were not sufficient to accurately constrain the
770 weights for a 6 member multi-model ensemble. Doblas-Reyes et al. (2005) found that a weighting
771 method for seven models improved over the simple multi-model mean with 40 years of data, but not
772 with 20 years. A difference between our method and that used by Delsole (2007) and Doblas-Reyes
773 et al. (2005) is that their method is based on the inversion of the outer product of model predictions
774 whereas our method is based on the inversion of the outer-product of model-observation
775 mismatches. The condition number of these matrices is different. If one method gives a very ill-
776 conditioned matrix (a concern of DelSole, 2007) and the other does not, it is possible that the
777 numerical accuracy of the inversions of the matrices might differ significantly. Our finding of a
778 significant decrease in mean square error in out of sample tests suggests that we did have a
779 sufficiently large number of observations for our method to usefully constrain the weights – even in
780 the per-cell example.
781
782 As noted by Palmer et al. (2008), an ideal (or more replicate-Earth like) set of models for estimating
783 the CPDF would be a set that could not only be shown to provide a good approximation to the
784 CPDF when compared to long data sets like HadCRUT3 but also be shown to accurately
785 approximate the WPDF when used for short term forecasts. If such a set of *quasi-independent*
786 models could be obtained then ensemble forecasts like those currently used to define the WPDF
787 could be seamlessly extended forward in time to provide seasonal, annual, decadal and centennial
788 projections. In practice, such ensembles of forecasts will inevitably suffer from *dependent model*
789 *errors* and the nature of these dependent model errors will depend on the time scale of the quantities
790 being forecast. In principle, the replicate-Earth transformation presented here could be used to
791 ameliorate the deleterious effects of correlated model errors on ensemble prediction performance at
792 a range of time scales.
793
794 We have shown that while error correlation provides a natural choice for a definition of model
795 dependence, the level of error correlation we should expect from independent estimates in an

ensemble depends strongly on the ensemble interpretation paradigm to which we subscribe. We introduced the replicate Earth paradigm of interpretation, which assumes the possibility of inherent uncertainty in the climate system at any spatial or temporal scale, and so does not anticipate perfect model-observation matching at any scale. An ensemble of perfect model (or replicate Earths) in this paradigm were shown to have well defined statistical properties that were not present in the CMIP3 ensemble. We then outlined an ensemble transformation process to transform the CMIP3 ensemble to one that did have these properties and showed that this transformed ensemble provided an improved prediction of the distribution of out-of sample observations.

**Acknowledgements:** The CMIP3 modelling groups, PCMDI and the WCRP's Working Group on Coupled Modelling (WGCM) for making the WCRP CMIP3 multi-model dataset available – support is provided by the Office of Science, U.S. Department of Energy. CHB was supported by the U.S. Office of Naval Research Grant# 4304-D-0-5. We also thank an anonymous reviewer for providing extensive and constructive feedback.

812 **References**
813
814 Annan JD, Hargreaves JC (2010) Reliability of the CMIP3 ensemble. Geophys. Res. Lett., 37, doi:
815 10.1029/2009gl041994.
816
817 Annan JD, Hargreaves JC (2011) Understanding the CMIP3 ensemble. J. Clim. 24, 4529–4538. doi:
818 10.1175/2011JCLI3873.1
819
820 Brohan P, Kennedy JJ, Harris I, Tett SFB, Jones PD (2006) Uncertainty estimates in regional and
821 global observed temperature changes: a new dataset from 1850. J. Geophys. Res. 111, D12106.
822
823 Collins M, Booth BB, Bhaskaran B, Harris GR, Murphy JM, Sexton DMH, Webb MJ (2010)
824 Climate model errors, feedbacks and forcings: A comparison of perturbed physics and multi-model
825 ensembles. Climate Dyn., 36, 1737–1766, doi:10.1007/s00382-010-0808-0.
826
827 DelSole T (2007) A Bayesian Framework for Multimodel Regression. J. Clim., 20, 2810–2826.
828
829 Doblas-Reyes FJ, Hagedorn R, Palmer TN (2005) The rationale behind the success of multi-model
830 ensembles in seasonal forecasting. Part II: Calibration and combination. Tellus, 57A, 234–252.
831
832 Furrer R, Knutti R, Sain SR, Nychka DW, Meehl GA (2007) Spatial patterns of probabilistic
833 temperature change projections from a multi- variate Bayesian analysis. Geophys. Res. Lett., 34,
834 L06711, doi:10.1029/2006GL027754.
835
836 Giorgi F, Mearns L (2002) Calculation of average, uncertainty range and reliability of regional
837 climate changes from AOGCM simulations via the 'reliability ensemble averaging' (REA) method.
838 J. Clim., 15, 1141–1158.
839
840 Glahn HR, Lowry DA (1972) The Use of Model Output Statistics (MOS) in Objective Weather
841 Forecasting. J. Appl. Meteor., 11, 1203–1211.
842
843 Gleckler PJ, Taylor KE, Doutriaux C (2008) Performance metrics for climate models. J. Geophys.
844 Res., 113, D06104.
845
846 Gneiting T, Raftery AE (2005) Weather Forecasting with Ensemble Methods. Science, 310, 248-
847 249
848
849 Greene AM, Goddard L, Lall U (2006) Probabilistic multimodel regional temperature change
850 projections. J. Clim., 19, 4326-4346.
851
852 Jewson S, Hawkins E (2009) CMIP3 ensemble spread, model similarity, and climate prediction
853 uncertainty, http://arxiv.org/abs/0909.1890.
854
855 Jun M, Knutti R, Nychka D (2008) Spatial analysis to quantify numerical model bias and
856 dependence: How many climate models are there? J. Amer. Stat. Assoc., 103, 934–947.
857
858 Knutti R, Abramowitz G, Collins M, Eyring V, Gleckler PJ, Hewitson B, Mearns L, *Good Practice*
859 *Guidance Paper on Assessing and Combining Multi Model Climate Projections*, in: Meeting Report
860 of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining

Multi Model Climate Projections Stocker TF, Qin D, Plattner G-K, Tignor M, Midgley PM (eds.) IPCC Working Group I Technical Support Unit, University of Bern, Bern, Switzerland, 2010.

Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA, (2010) Challenges in combining projections from multiple models. J. Clim., 23, 2739-2758.

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, 4196–4216.

Lambert SJ, Boer GJ (2001) CMIP1 evaluation and intercomparison of coupled climate models. Climate Dyn., 17, 83–106.

Leutbecher M, Palmer TN (2008) Ensemble Forecasting. J. Comput. Phys., 227, 3515-3539

Masson D, Knutti R (2011) Climate model genealogy, Geophys. Res. Lett., 38, L08703, doi:10.1029/2011GL046864.

Meehl, G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer, and K. E. Taylor, 2007: The WCRP CMIP3 multi-model dataset: A new era in climate change research. *Bul. Am. Meteorol. Soc.,* **88**, 1383-1394.

Meehl GA *et al.* 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* (ed. Solomon, S. *et al.*) (Cambridge Univ Press, Cambridge, UK and New York, 2007).

Palmer TN, Doblas-Reyes FJ, Weisheimer A, Rodwell MJ (2008) Toward Seamless Prediction: Calibration of Climate Change Projections Using Seasonal Forecasts. Bull. Amer. Meteor. Soc., 89, 459–470.

Pennell C, Reichler T (2011) On the Effective Number of Climate Models. J. Clim., 24, 2358–2367, doi: http://dx.doi.org/10.1175/2010JCLI3814.1

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, 4356-4376.

Pirtle Z, Meyer R, Hamilton A (2010) What does it mean when climate models agree? A case for assessing independence among general circulation models, Environmental Science & Policy, 13, 351–361.

Reifen C, Toumi R (2009) Climate projections: Past performance no guarantee of future skill? Geophys. Res. Lett., 36, L13704.

Schär C, Vidale PL, Lüthi D, Frei C, Häberli C, Liniger MA, Appenzeller C (2004) The role of increasing temperature variability in European summer heatwaves. Nature 427, 332-336.

Smith RL, Tebaldi C, Nychka DW, Mearns LO (2009) Bayesian modeling of uncertainty in ensembles of climate models. J. Am. Stat. Assoc. 104, 97-116.

909 Snyder C, Bengtsson T, Bickel P, Anderson JL (2008) Obstacles to High-Dimensional Particle
910 Filtering. Mon. Wea. Rev., 136, 4629–4640.
911
912 Tebaldi C, Smith RW, Nychka D, Mearns LO (2005) Quantifying uncertainty in projections of
913 regional climate change: A Bayesian approach to the analysis of multi-model ensembles. J. Clim.,
914 18, 1524-1540.
915
916 Tebaldi C, Knutti R (2007) The use of the multimodel ensemble in probabilistic climate projections.
917 Phil. Trans. R. Soc. A, 365, 2053-2075.
918
919 van Leeuwen PJ (2009) Particle Filtering in Geophysical Systems. Mon. Wea. Rev., 137, 4089–
920 4114.
921
922 Weigel AP, Knutti R, Liniger MA, Appenzeller C (2010) Risks of Model Weighting in Multimodel
923 Climate Projections. J. Climate, 23, 4175–4191.
924
925 Wilson LJ, Vallée M (2002) The Canadian Updateable Model Output Statistics (UMOS) System:
926 Design and Development Tests. Wea. Forecasting, 17, 206–222.

## Figure Captions

**Figure 1:** Global mean surface temperature over the last century, expressed as an anomaly. An ensemble of climate models is represented by the yellow lines, their multi-model mean in red, and the observational record in black (originally Figure TS.23.a in IPCC AR4 Working Group 1 Technical Summary).

**Figure 2:** Mean square error of the optimal linear combination of ensemble members as a function of the error correlation parameter $\rho$ for an idealized 5 member ensemble having an error covariance matrix given by Equation (8).

**Figure 3:** (a) Synthetic model ensemble example illustrating the dependence weighting approach, specifically how the estimate it creates is invariant with the addition of dependent models to an ensemble, unlike the multi-model mean. (b) Transformation of the ensemble in (a) to be more replicate Earth-like in that the multi-model mean produces a better estimate and the ensemble variance is constrained, as described in Section 5.

**Figure 4:** The root mean square difference (RMSD) between (a) CMIP3 multi-model ensemble mean, (b) independence-weighted ensemble at each grid cell and the HadCRUT3 data set for years 1970-1999. White regions indicate greater than 20% missing data. Results are averaged over the 30 out-of sample tests described in Section 3b.

**Figure 5:** The relative global root mean square difference of the multi-model ensemble mean with several possible weighting strategies including simple bias-correction, weighting for performance differences and weighting for model dependence. Results from thirty out-of-sample tests are fitted to normal distributions. The persistence-equivalent prediction uses the mean of each month's observed value for all years except the testing year in each out-of-sample experiment.

**Figure 6:** Rank frequency histograms or Talagrand diagrams of the observation amongst variants of the 24-member CMIP3 ensemble. All monthly temperature predictions for 30 years at all grid cells are used. The black histogram shows the original ensemble; blue the globally bias corrected ensemble; red the globally transformed models, as described in Section 5.

**Figure 7:** Rank frequency histograms for different sizes of model ensembles incorporating both dependence weights and ensemble variance rescaling. Dashed plots show the original, globally bias corrected and per-cell bias corrected ensembles at each grid cell and for each monthly value (black, blue and green respectively). Resampled global and per-cell transformed model ensembles are shown in red and orange, with the frequency of each ensemble member's inclusion from the original 24 model ensemble determined by the transformed models' weights.

966 **Table 1:** Root mean square difference of the raw multi-model mean, bias-corrected mean,
967 performance weighted mean and independence weighted mean. Values are shown separately for the
968 1990s (top), 80s and 70s (bottom) with the training decade for corrections and weights shown on
969 the left of the first column and testing decade shown on the right of the first column. In-sample tests
970 are shaded.
971

| | MM mean | BC mean | Perf. weights | Indep. weights |
|---|---|---|---|---|
| 90s > 90s | 1.867 | 1.780 | 1.753 | 1.687 |
| 70s > 90s | 1.867 | 1.780 | 1.752 | 1.696 |
| 80s > 90s | 1.867 | 1.780 | 1.754 | 1.698 |
| 80s > 80s | 1.897 | 1.812 | 1.785 | 1.723 |
| 70s > 80s | 1.897 | 1.812 | 1.783 | 1.735 |
| 90s > 80s | 1.897 | 1.812 | 1.784 | 1.736 |
| 70s > 70s | 1.957 | 1.876 | 1.840 | 1.775 |
| 80s > 70s | 1.957 | 1.877 | 1.843 | 1.787 |
| 90s > 70s | 1.957 | 1.877 | 1.842 | 1.784 |

972

**Figure 1:**

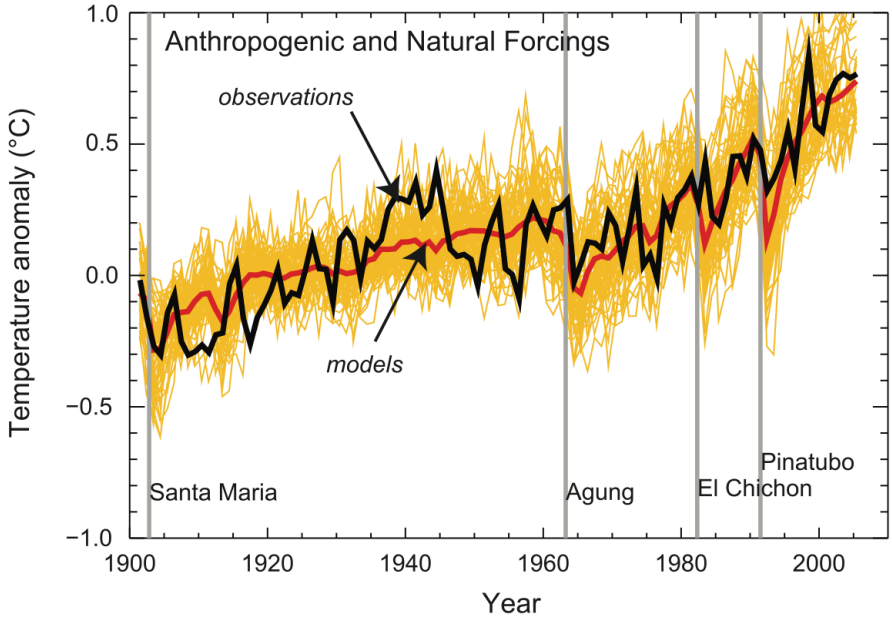
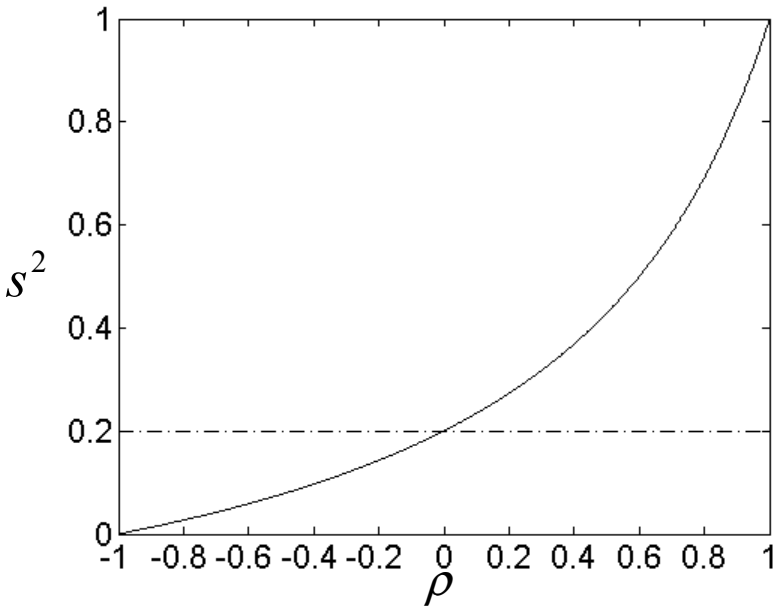
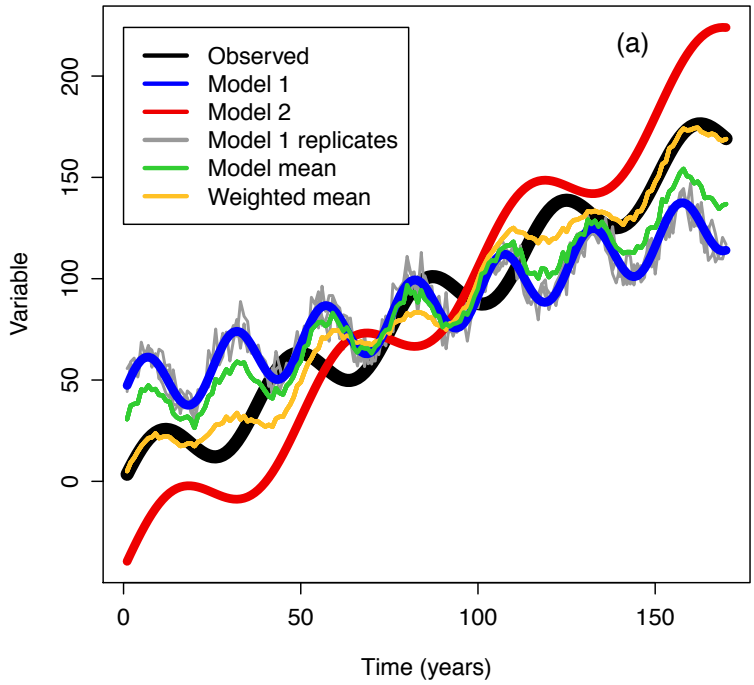
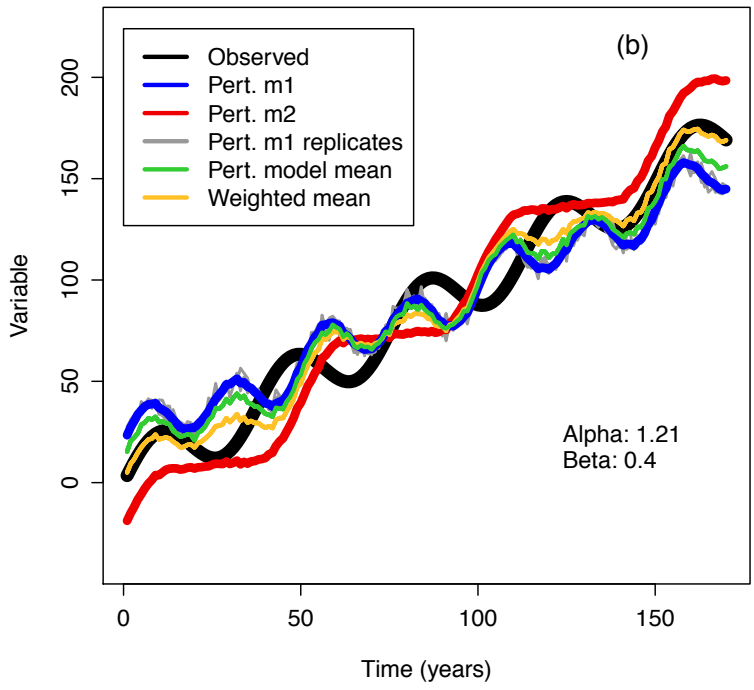

Anthropogenic and Natural Forcings
observations
models
Santa Maria
Agung
El Chichon
Pinatubo
Temperature anomaly (°C)
Year
1.0
0.5
0.0
−0.5
−1.0
1900
1920
1940
1960
1980
2000

**Figure 2:**



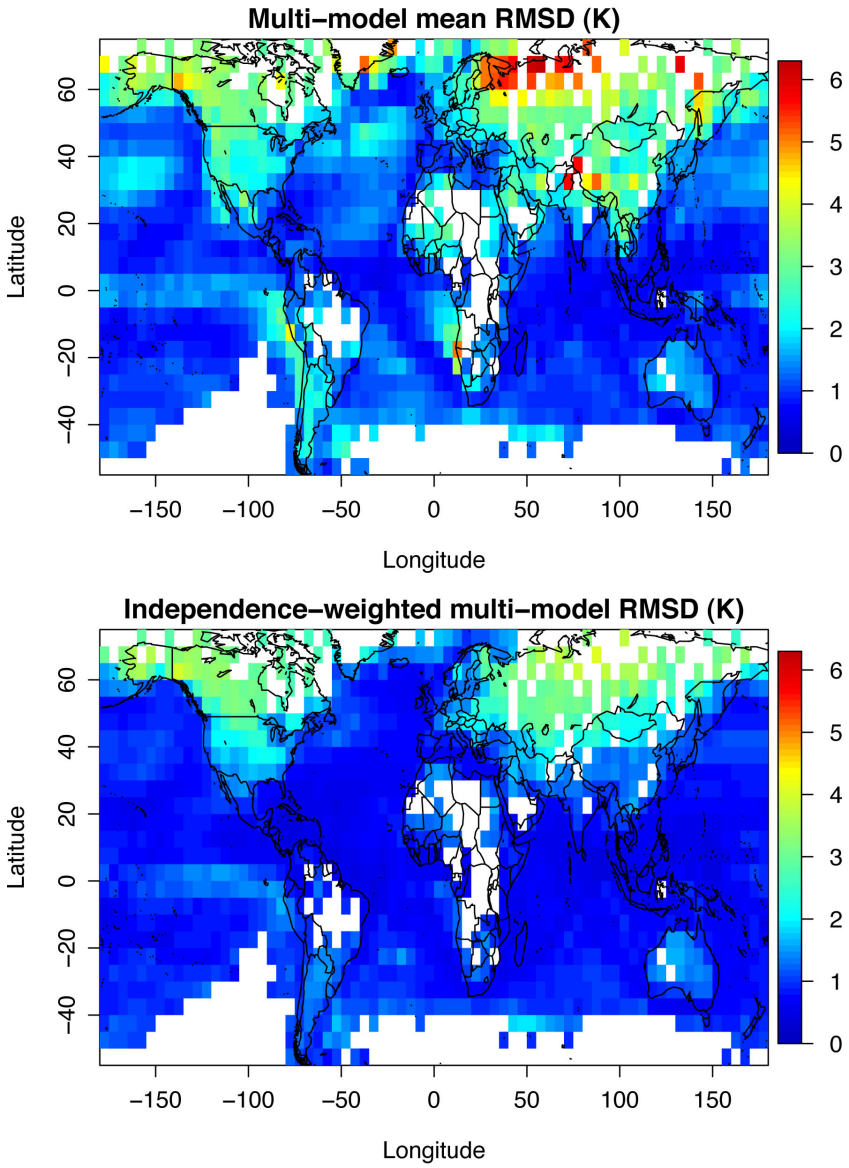
983 **Figure 3:**



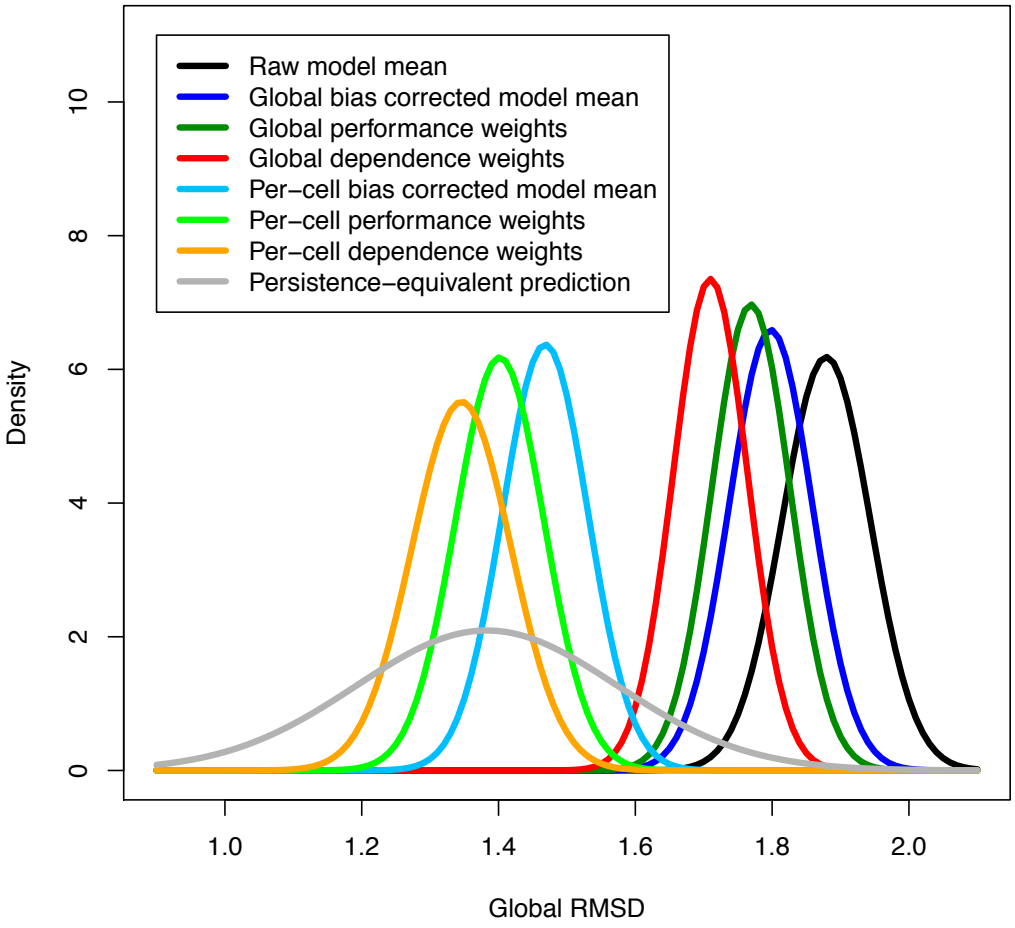

984

985 **Figure 4:**
986



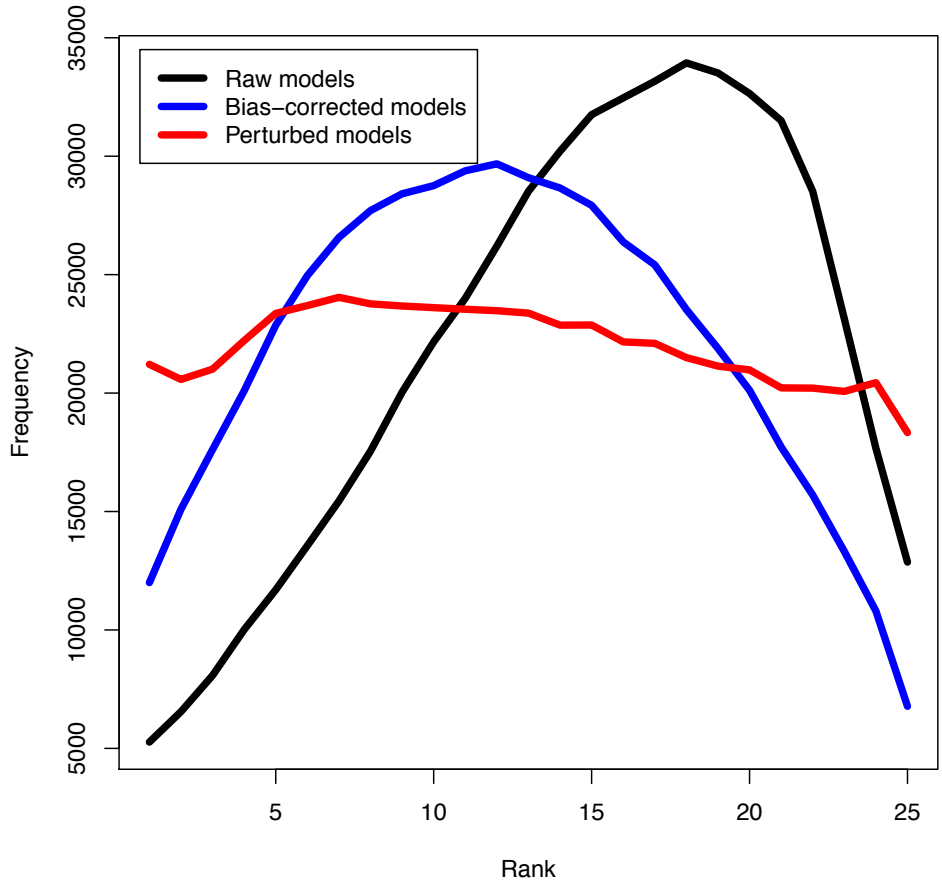

987

988 **Figure 5:**



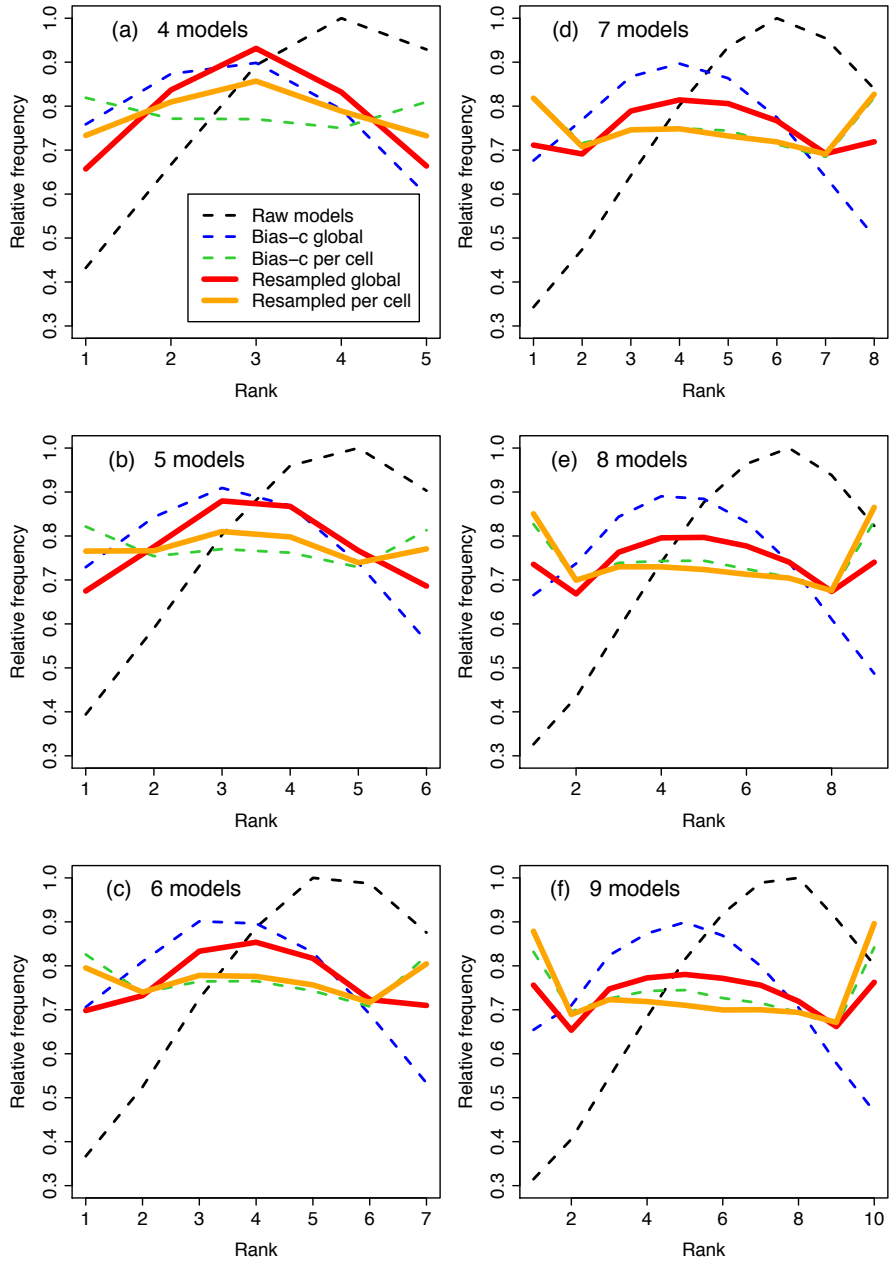

989
990

991
992 **Figure 6:**




993
994

**Figure 7:**


(a) 4 models
(b) 5 models
(c) 6 models
(d) 7 models
(e) 8 models
(f) 9 models
Raw models
Bias–c global
Bias–c per cell
Resampled global
Resampled per cell
Relative frequency
Rank