

Understanding the CMIP3 multi-model ensemble

J. D. ANNAN * AND J. C. HARGREAVES

RIGC/JAMSTEC, Yokohama, Japan

* *Corresponding author address:* J. D. Annan, Research Institute for Global Change, 3173-25 Showamachi,
Yokohama, Japan

E-mail: jdannan@jamstec.go.jp

ABSTRACT

The CMIP3 multi-model ensemble has been widely utilised for climate research and prediction, but the properties and behavior of the ensemble are not yet fully understood. Here we present some investigations into various aspects of the ensemble's behaviour. In particular, we explain why the multi-model mean is always better than the ensemble members on average, and we also identify the properties of the distribution which control how likely it is to out-perform a single model. Our analyses further support the paradigm of a statistically indistinguishable ensemble, and indicate that the current ensemble size is too small to adequately sample the space from which the models are drawn.

1. Introduction

Global climate models are the primary means by which projections of climate change are made. The World Climate Research Programme's Coupled Model Intercomparison Project phase 3 (CMIP3) multi-model dataset contains results from more than 20 of the major global climate models developed around the world. While this resource has proved very valuable, it has also highlighted the significant differences between model projections, and major questions remain as to the most appropriate treatment of this set of diverse results. Each model provides a different projection, due to differences in their parameterisations and numerical methods. There has been extensive discussion on how best to interpret the ensemble and analyse its outputs, in order to make credible predictions of future climate change. Various statistical weighting schemes have been proposed, (eg Giorgi and Mearns 2002; Smith et al. 2009; Knutti et al. 2009), but with little consensus as yet. Competing

paradigms for the interpretation of the ensemble of models may lead to substantially different methods for processing their outputs (Annan and Hargreaves 2010) and thus it is important that we develop a fuller understanding of the multi-model ensemble.

The purpose of this paper is to investigate some of the properties of the multi-model ensemble, in order to better understand various aspects of its behaviour. Our analysis runs primarily along lines motivated by geometrical considerations. In Section 2, we will, for the first time, present a simple explanation for the oft-noted good performance of the multi-model mean. In Section 3, we consider the performance of the multi-model mean in more detail, specifically the phenomenon of it frequently, but not inevitably, outperforming all models in the ensemble. We show that this depends on various parameters of the sampling distribution. We analyse the CMIP3 ensemble in Section 4, reconciling the behaviour of the ensemble with its effective dimension, which raises some questions about the adequacy of the sample size. We summarise our results in Section 5.

2. Why is the multi-model mean so good?

Right from the earliest days of the analysis of the CMIP1 ensemble of climate models, it has been noted during comparisons with observational data that the multi-model mean tends to have a lower RMSE than most, if not all, individual models (Lambert and Boer 2001). This phenomenon has been repeatedly replicated, and indeed highlighted, in subsequent research (eg Gleckler et al. 2008). One possible explanation that has sometimes been proposed is the paradigm of models being independent samples from some distribution centred on the truth, in which case the multi-model mean could be expected to converge to the

truth as more models are added to the ensemble (eg Tebaldi and Knutti 2007, and references therein). However, this hypothesis has no credible theoretical or philosophical foundation that we know of, and is convincingly refuted by analysis of the models themselves (Knutti et al. 2009). In fact the ensemble of models can be much more plausibly considered as statistically indistinguishable from the truth (Annan and Hargreaves 2010). Therefore, a claim of truth-centredness can hardly be invoked as a basis for the good performance of the multi-model mean. Thus, this question appears to have remained unanswered until now. Here we present a simple explanation of this effect, which is purely algebraic in nature, and is entirely independent of climate science, climate models or ensemble generation methods.

If we write O for an arbitrary vector of observations of climatic variables, m_i for the equivalent outputs from the i th member of an ensemble of n models, and $M = \frac{1}{n} \sum m_i$ for the multi-model mean (all sums are over i unless otherwise stated), then we can perform the standard manipulation ($\|\cdot\|$ is the Euclidean distance norm):

$$\begin{aligned} \frac{1}{n} \sum \|m_i - O\|^2 &= \frac{1}{n} \sum \|(m_i - M) - (O - M)\|^2 \\ &= \frac{1}{n} \sum \|m_i - M\|^2 - \frac{2}{n} \sum (m_i - M) \cdot (O - M) + \|O - M\|^2 \end{aligned}$$

By the definition of M , the cross product term is zero, so we find:

$$\frac{1}{n} \sum \|m_i - O\|^2 = \frac{1}{n} \sum \|m_i - M\|^2 + \|O - M\|^2 \quad (1)$$

and thus the mean of the squared distances between the individual models and the observations is greater than the square of the distance from the multi-model mean to the observations, by an amount which depends solely on the spread of the models around their

mean. We note that this simple algebraic identity makes no appeal to any properties of the errors in the underlying physics of the models, nor how the ensemble was generated. Furthermore, it does not even depend on where the observations happen to lie, and therefore does not require that the errors of different models have a tendency to cancel. On the contrary, the result holds even where the sign of errors is the same across all ensemble members.

3. When is the mean better than all of the models?

Not only is the multi-model mean better than most models, it is not infrequently observed to be better than *all* of the models in the sample. This, however, depends on the particular comparison that is being made. For example, Figure 1 of Lambert and Boer (2001) indicates that of the 15 models which participated in the CMIP1 experiment, the multi-model mean has the best representation of surface air temperature and precipitation for both summer and winter seasons, but that a few individual models are better for sea level pressure. Here we will call such a model, that is closer to a set of observations (in terms of having a lower RMS error) than the multi-model mean, a ‘nearer neighbour’ for those data. Gleckler et al. (2008) analysed the CMIP3 ensemble, considering a wide range of climatic variables, and reported that in most but not all cases, the multi-model mean scores better than all of the individual models. We now consider this phenomenon in more detail.

If we consider a single model m_i and perform similar algebraic manipulation to before,

we obtain:

$$\begin{aligned} \|m_i - O\|^2 &= \|(m_i - M) - (O - M)\|^2 \\ &= \|m_i - M\|^2 - 2(m_i - M) \cdot (O - M) + \|O - M\|^2 \end{aligned} \quad (2)$$

This time the cross product term does not vanish, and we can see that the model will lie closer to the observations than the multi-model mean does, precisely when $\|m_i - M\|^2 - 2(m_i - M) \cdot (O - M) < 0$ or equivalently when $2 \cos \theta > \|m_i - M\|/\|O - M\|$, where θ is the angle between the two vectors $O - M$ and $m_i - M$. If we keep the lengths of these two vectors fixed while allowing their angle to vary, this condition requires that the angle has to lie below some threshold which depends on the ratio of the vector lengths but which is always less than $\pi/2$. The probability of this condition holding will depend on the sampling distributions of the models and data.

a. Isotropic case

We present some numerical calculations to illustrate the consequences of the above analysis. We wish to estimate the probability of a model being a nearer neighbour to the data, and how this may be affected by various factors relating to the sampling distributions of models and data. We adopt the paradigm of the statistically indistinguishable ensemble (Annan and Hargreaves 2010) and therefore start by generating synthetic models and observations from the same distributions. Initially we use the multivariate d -dimensional Normal $N(0, 1)^d$, and consider how the results vary with d . The metric we adopt is the Euclidean distance (equivalent to the widely used root mean square difference, up to a scaling factor), and thus in this example the sampling distribution is isotropic in the metric space. In order to eliminate

dependence of the results on sample size, we measure the distance of the observations from the known mean of the sampling distribution, rather than an empirically estimated mean of a finite sample. For realistic ensemble sizes this approach has very little influence on our results.

The results are plotted as the solid dark blue line in Figure 1. It is immediately apparent that there is a strong dependence on the dimension of the sampling distribution. This is basically due to the well-known phenomenon of random vectors being increasingly close to perpendicular (with high probability) in high dimensional spaces. When the vectors $m_i - M$ and $O - M$ are sufficiently close to perpendicular, Equation 2 shows that the distance $\|m_i - O\|$ will exceed $\|O - M\|$, so a model is increasingly unlikely to be a nearer neighbour to the data as the dimension of the sampling distribution increases.

We can relax the assumption of a statistically indistinguishable ensemble by changing the sampling distribution of the models by a scaling factor while keeping the observational sampling distribution unchanged. The base case of the statistically indistinguishable ensemble is contrasted in Figure 1 with experiments in which the models are sampled from a distribution which is either half, or twice, the width of that from which the observations were picked. It may seem counterintuitive at first, but the narrower the ensemble distribution is (and therefore the more likely that the observations are well outside the ensemble range), the higher is the probability that a randomly sampled model will be a nearer neighbour to the data, when compared to the multi-model mean. This is actually easy to explain in geometrical terms. For the statistically indistinguishable case, the two vectors $m_i - M$ and $O - M$ will typically be of similar length, and when this is the case, then in order for O to be closer to m_i than it is to M , the angle between these two vectors must be rather acute, satisfying $\cos \theta > 0.5$.

If $\|O - M\| = 2\|m_i - M\|$ then we only require the weaker condition $\cos \theta > 0.25$. In the case of a wide ensemble, the converse holds and the cosine of the angle has to be large. In fact, for any sufficiently distant ensemble member where $\|m_i - M\| > 2\|O - M\|$, m_i cannot be a nearer neighbour for the data as this would require $\cos \theta > 1$. Therefore, the closer the observations are to the multi-model mean (relative to the ensemble spread), the fewer nearer neighbours we would expect to find.

b. Anisotropic case

In many geophysical applications, the number of degrees of freedom of gridded data sets may be very large, but correlations across the grid are often significant and most of the variability can generally be explained by a relatively small number of empirical orthogonal functions (EOFs). Therefore, we now perform experiments for two families of sampling distributions which are anisotropic in the metric space, to investigate how this may affect the probabilities of nearer neighbours. In order to provide a compact and meaningful comparison across different families of sampling distribution, we use the formula for the effective number of degrees of freedom presented as Equation 4 of Bretherton et al. (1999). That is, if the i th EOF of the sampling distribution explains a fraction f_i^2 of the total variance, then the number of effective dimensions N_{ef} can be defined as $N_{ef} = 1 / \sum f_i^2$. For the isotropic case considered above, the number of effective dimensions equals the number of true dimensions d . If, instead, we use a distribution in which the eigenvalues λ_i drop off geometrically, $\lambda_i / \lambda_{i-1} = k$ for some ratio $k < 1$ then Bretherton et al. show that the effective dimension in this case is given by $(1 + k) / (1 - k)$ and that the first N_{ef} EOFs will explain roughly

$1 - e^{-2} = 86\%$ of the total variance. We construct such a distribution by using a high-dimensional multivariate Normal where the i th dimension is scaled so as to be sampled from $N(0, k^i)$. The dotted lines in Figure 1 show how the probability of a nearer neighbour for this distribution changes with the effective dimension, again for the three cases of wide, perfect, and narrow ensembles.

Finally, we also consider the case where the eigenvalues decrease as $\lambda_i \propto 1/\sqrt{i}$. For this slow decay, the effective dimension increases without bound with the number of eigenvalues, and thus it can (as in the isotropic case) be adjusted to choice by changing the number of dimensions of the sampling distribution. For this distribution of eigenvalues, the first N_{ef} EOFs explain around 80% of the total variance for $N_{ef} \simeq 5$, decreasing to around 50% at $N_{ef} \simeq 40$. The probability of nearer neighbours for this distribution are shown by the dashed lines in Figure 1. For a more rapid decay such as $\lambda_i \propto 1/i$, N_{ef} can easily be shown to be bounded above by 2.5 irrespective of the number of underlying dimensions. Such a low upper bound renders this distribution irrelevant to our investigations.

From these experiments, which sample a wide range of behaviours for the distributions of eigenvalues, we see that the probability of a nearer neighbour may be influenced to some extent by the distribution of variance among the EOFs, with the isotropic case having higher probabilities than the other two families of distributions. Nevertheless, the effective number of degrees of freedom, and relative widths of the distributions from which the models and observations are sampled, remain the dominant effects.

4. CMIP3 analysis

The theoretical insights developed in Section 3 can be used to analyse to output from the CMIP3 models, first exploring the nearer neighbour phenomenon and then continuing with a more detailed analysis of the effective dimension of the ensemble based on an EOF decomposition. We observe that the selection of a single model can be interpreted as a degenerate re-weighting in which one model is assigned full weight and the others all receive zero weight. Therefore, the existence (or otherwise) of models which outperform the multi-model mean may have some bearing on the debate over re-weighting of models according to their performance. We will not, however, pursue this major topic here but intend to consider it in a separate paper.

We use output from the set of 24 climate models analysed by Annan and Hargreaves (2010), for which data for the 20C3M scenario are available from the CMIP3 database. We analyse fields of three climatic variables: surface air temperature (SAT), with observational data obtained by Jones et al. (1999); precipitation (PPT), using the data of Adler et al. (2003); and sea level pressure (SLP) versus the data of Allan and Ansell (2006). All model and observational data sets are firstly regridded onto 5 degree global grids and averaged over the years 1961-1990 (temperature and sea level pressure) or 1979-1999 (precipitation), and we restrict our attention here to annual mean values. In principle, observational uncertainty should be accounted for by adding equivalent pseudo-errors onto the model outputs before any comparison. In this paper, as is common more widely in the evaluation of climate models, we ignore the issue of observational uncertainties, as we expect them to be small compared to inter-model differences. As an initial check of model behaviour, we tested whether the

observations lie at a similar distance (as defined by the area-weighted root mean square) from the multi-model mean as the models do themselves: the distances from the multi-model mean to the three sets of observations have rank 5, 10 and 17 respectively in the set of 25 distances based on the 24 models and the observations themselves. While this does not by itself provide strong evidence that the models can be considered as statistically indistinguishable from the truth, it also does nothing to undermine the hypothesis.

a. Nearer neighbours

When testing for nearer neighbours, we find that, for SAT, the observations have no nearer neighbour among the model ensemble. For PPT and SLP respectively, 1 and 5 of the models are closer than the multi-model mean. These results appear compatible with those presented by Gleckler et al. (2008). Given the sample size of 24 in each case, these figures represent a frequency of around 8% of the sample overall (with a range of 0–21% across the three data types). We can also perform a leave-one-out validation of the nearer neighbour analysis, using each model as ‘observations’ in turn. This shows that the results obtained for the real data are entirely unremarkable: on using each model in turn as a surrogate data set and checking for nearer neighbours among the remaining 23 models, we find that for the three data sets, 11, 6 and 5 respectively of the 24 models had no nearer neighbour (and one model had no nearer neighbour for *any* data set). The average number of nearer neighbours for each model is 2, 1 and 4.3, or about 9%, 4% and 19% of the sample, for each climatic field in turn. Therefore, the numbers of nearer neighbours to the data are compatible with the paradigm of a statistically indistinguishable ensemble. Furthermore, since the multi-

model ensemble members are exchangeable by definition, reference to the dark blue lines on Figure 1 suggests an effective dimension of the global fields in the range of 4–14, though it is not clear how precise a diagnostic this approach can provide.

A joint analysis of all three data sets combined, equally weighted according to the error on the multi-model mean, finds that in this case the ensemble contains no model which is nearer to the data than the multi-model mean is. The equivalent leave-one-out analysis is again consistent with this result with 6 of the models having no nearer neighbour, and an average number of nearer neighbours per model being only 1.2, or 5% of the sample. Interpreting these values through Figure 1 suggests a dimension of about 10–13 for the combined data set, which lies towards the upper end of the values obtained for each data set individually, but probably not as high as their sum (which we might expect were the fields to vary independently).

As a further test of the theory, we randomly select subsets of 4 models and average their outputs, to generate a large number of pseudo-models drawn from a distribution which has the same mean as the underlying sampling distribution of the models, but with its width reduced by a factor of 2. As predicted by the cyan line in Figure 1, a much higher proportion of these pseudo-models are closer to the observations, than the multi-model mean is: 15%, 22% and 35% of the samples are nearer neighbours for SAT, PPT and SLP respectively. Leave-one-out validation generates comparable values of 24%, 21% and 31%. These figures are again consistent with an effective dimension of around 4–12 for the three data sets.

These results all appear broadly consistent with the concept of a statistically indistinguishable ensemble, and suggest a dimension of the order 4–14 for the fields of climatological variables that are used here. The leave-one-out validation generates results which are com-

patible with those for the real data. However, the accuracy and reliability of this approach for estimating the effective dimension of the ensemble of models is not clear.

This analysis of nearer neighbours shows that the effective dimension of the problem is a critical parameter for quantitative analysis of ensemble performance. Therefore, we next consider some other approaches for its estimation.

b. EOF analysis and effective dimensions

Annan and Hargreaves (2010) assumed a dimension of 40 for global fields of climate data, based in part on a decorrelation length scale of $O(1-2000\text{km})$, estimated from semi-variograms of model errors, and also consistent with estimates of around 25 degrees of freedom for a hemisphere of synoptic data (Bretherton et al. 1999; Jolliffe and Primo 2008). However, such an analysis does not take account of the fact that the differences in modelled climatologies are not really linked directly to synoptic-scale variations in the atmospheric state, but rather depend on the underlying physical parameterisations. For each model, the basic physical parameterisations are the same across the globe and thus similar inter-model differences may be expected to persist over widely dispersed areas with similar climates, which may reduce the effective dimension substantially. When assuming 40 dimensions, the rank histograms of Annan and Hargreaves (2010) were found to be significantly non-uniform, implying a lack of reliability for the model ensemble, in the sense discussed in that paper. On the other hand, their leave-one-out analysis suggested a much lower alternative value of around 5 dimensions. For this value, the non-uniformity of the histograms would be statistically insignificant and we could conclude that the rank histogram analysis provides

no evidence that the ensemble is unreliable.

These alternative values for effective dimension would also justify radically different interpretations of our nearer-neighbour results. Referring to the dark blue line in Figure 1, a choice of 40 dimensions suggests that if the ensemble really was statistically indistinguishable from the observations, we should almost certainly find zero nearer neighbours for any data set. Therefore the proportion we obtained, of 8%, would require that the ensemble is too narrow by a factor of 2 or more. Conversely, the lower figure of 5 dimensions should result in around 15–20% of models being nearer neighbours to the data, and in this case the lower observed frequency would imply that the ensemble is instead rather too broad. Therefore, the effective dimension of the climate fields is an parameter of fundamental significance in the analysis and interpretation of the multi-model ensemble, and now we consider this in more detail through an EOF decomposition of the ensemble of model climatologies.

With a sample size of 24, there are 23 EOFs, but, as anticipated, the variance of the inter-model differences is concentrated in the first few EOFs. These generally represent large-scale patterns such as latitudinal variation and land-ocean contrasts. Applying Equation 4 of Bretherton et al. (1999) to the EOF analyses of the three data sets in turn, the effective dimension can be estimated at 4.6, 7.5 and 3.3 respectively for the three variables SAT, PPT and SLP in turn. When all three data sets are combined (inversely weighted according to their standard deviations), the effective dimension of 7.6 barely exceeds that obtained for the precipitation.

The leading N_{ef} EOFs represent roughly 80% of the total variance of the model ensemble in each case, but the first N_{ef} EOFs only represent 40–65% of the variance of the three types of observations. In fact the full set of EOFs only account for 54–87% of the observational

variances, implying that the observations contain substantial variability that is not described by the multi-model ensemble. This would appear to cast doubt on our hypothesis of statistical indistinguishability. However, these results are in fact consistent with what we obtain through leave-one-out validation: when one model is withheld, the EOFs of the remaining 23 of the models only explain on average 69%, 57% and 89% of the variation of the SAT, PPT and SLP of the withheld models respectively. Furthermore, for 4, 11 and 6 of the models respectively, the proportion of their variance that is unexplained is greater than it was for the observations in the original analysis. These figures indicate that the results for the observations are only on average marginally worse than results for the models themselves, and lie well within the ensemble range.

One possible explanation of this is that the ensemble size may be too small to adequately represent the underlying distribution from which it was sampled. Equation 14 of Bretherton et al. (1999) suggests that for a sample of 24 and true effective dimension of 4–11, estimation of N_{ef} will have a low bias of around 15–30% (larger for the higher dimension), with additional uncertainty of about 10% (at one standard deviation) around that value. According to this mean bias formula, we would expect distributions with dimension 6, 11 and 4 to roughly match the figures derived from the CMIP3 data.

This can also be verified by looking at how the effective dimension of subsets of models changes with the sample size. Figure 2 shows the mean effective dimension of different sized subsets of models, for the three variables separately. The effective dimension increases steadily with sample size, and seems some way from saturation at the full sample size of 24 for all of the three cases. This figure also shows equivalent results generated by samples from the isotropic distribution $N(0, 1)^d$ considered in Section 3a (the other families of distributions

generate similar results). It seems that underlying effective dimensions of 6, 11 and 4 do provide reasonably good fits to the observationally-derived results. This confirms our guess: the samples behave as if they are drawn from a space with a somewhat larger dimension, which the finite sample size of 24 is inadequate to fully describe. Encouragingly, these results also match those obtained by the nearer neighbour analysis in Section 4a.

We should note that Bretherton et al. (1999, Equation 1) presents an alternative formula for estimating effective degrees of freedom. However, this formula is very inaccurate for small sample sizes, with relative uncertainties as high as 30% (again at one standard deviation) for a sample size of 24, with this formula itself depending on various approximations. When applied to the CMIP3 model results, this alternative method generates very low estimates for the effective dimension, ranging from 1.2 to 2.2 for the three data types. These results are impossible to reconcile with the nearer neighbour analysis, the percentages of variance explained, or the estimates using the other formula, so are not considered credible.

In summary, our nearer neighbour analysis of the CMIP3 database supports our previous suggestion that the ensemble can reasonably be interpreted as statistically indistinguishable from the truth (Annan and Hargreaves 2010). Furthermore, it suggests that the effective dimension of the ensemble of climatological fields is around 4–11, depending on the variable in question. These results are backed up with EOF analysis. With these values, the non-uniformity for the rank histograms of Annan and Hargreaves (2010) is no longer significant at the $p < 5\%$ level.

Our conclusions appear to differ somewhat from those of Jun et al. (2008a), who also presented an eigenvalue analysis of the CMIP3 ensemble. One possible cause of this may relate to their use of a localised approach which, by focussing on the finer scales, may pose

a stiffer challenge to models which in several cases barely exceed (and therefore cannot fully resolve) the resolution of the gridded observations. More importantly, we would not interpret their results observations as inconsistent with the statistically indistinguishable paradigm, since their single data point (their Table 3) lies at the 10th percentile of the rank histogram. Of course, the statistically indistinguishable paradigm is not axiomatically true, and indeed we consider it is likely to be falsifiable through sufficiently detailed analysis. However, it appears to generally hold up fairly well under a wide range of analyses, and may therefore be considered a reasonable starting-point for use of the multi-model ensemble. Some further results derived from the paradigm of a statistically indistinguishable ensemble are presented in the Appendix.

5. Conclusions

By using some simple geometrically-inspired methods, we have explained why the multi-model mean has such good performance, firstly in having a lower root mean square error than the average of the individual models, and secondly in terms of how likely it is to outperform the best of the models. While the former result is a trivial algebraic result, the latter depends strongly on the relative widths, and effective dimensions, of the sampling distributions. Our analysis here further supports the notion of a statistically indistinguishable ensemble, as the results obtained with observational data are consistent in all respects with those generated by leave-one-out validation.

We have shown that the individual data fields have an effective dimension of around 4–11 for the three climatic variables considered. The upper value does not increase when

all data types are considered together, but might be expected to if future changes were also considered, especially in light of the weak relationship between present and future climate (Whetton et al. 2007; Abe et al. 2009). The EOF analysis, and calculation of the effective dimension of subsets of models, also shows that the sample size is too small to fully characterise the distribution from which it is drawn. Thus we might expect a larger set of models (constructed with alternative plausible physical parameterisations and numerical methods) to introduce some additional patterns of climate which are significantly distinct from those already obtained.

Acknowledgments.

This work was supported by the S-5-1 project of the MoE, Japan and by the Kakushin Program of MEXT, Japan. We acknowledge the modeling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled Modelling (WGCM) for their rôles in making available the WCRP CMIP3 multi-model dataset. Support of this dataset is provided by the Office of Science, U.S. Department of Energy.

APPENDIX

Some corollaries of the statistically indistinguishable paradigm

a. How can we interpret pairwise error correlations?

The pairwise correlation between the errors of ensemble members $\text{corr}_{i \neq j}((m_i - O), (m_j - O))$ has been studied in various ensemble analyses (Jun et al. 2008b; Collins et al. 2010). The correlation of two vectors can be expressed as their dot product divided by their norms. Therefore, if m_i and m_j are vectors of model outputs as before, then the correlation of their errors is equal to $\frac{(m_i - O) \cdot (m_j - O)}{\|(m_i - O)\| \|(m_j - O)\|}$. The numerator can be expanded similarly to in Equation 1, arriving at $(m_i - M) \cdot (m_j - M) + (m_i - M) \cdot (M - O) + (m_j - M) \cdot (M - O) + \|M - O\|^2$. The three dot product terms will be zero on average over i, j , and will also typically be relatively small compared to the last term if the dimension of the problem is large, due to the near-orthogonality of the vectors as explained previously. Therefore, the numerator will vary around, and generally be quite close to, the single term $\|M - O\|^2 = \sigma_O^2$, where σ_O is the distance of the observations from the multi-model mean. The two norm terms in the denominator can be expanded along similar lines, to get $\sqrt{\|m_i - M\|^2 + 2(m_i - M) \cdot (M - O) + \|M - O\|^2}$ and the equivalent for m_j . Here only the cross product is again zero on average, and usually small. The first term is always positive, and represents the squared distance of the model from the multi-model mean. As i and j vary, the denominator will vary around $\sigma_m^2 + \sigma_O^2$ where σ_m is the root mean square distance of the ensemble members from the multi-model mean. While in contrast to Equation 1, this analysis cannot be considered a strict proof, we can still reasonably expect the pairwise correlations to generally cluster around the value

$$\frac{\sigma_O^2}{\sigma_m^2 + \sigma_O^2}.$$

In the case of a statistically indistinguishable ensemble where the observations and models are similar distance from the multi-model mean, the pairwise correlations should therefore be clustered around 0.5. For the case where the observations are sampled from a distribution which has half or double the width of that of the models, the correlations will be clustered around 0.2 and 0.8 respectively. This is confirmed experimentally in the idealised cases shown in Figure 1 of Annan and Hargreaves (2010). Figure 3 of Knutti et al. (2009) presents some pairwise correlations for the CMIP3 ensemble of a little less than 0.5 on average, which is consistent with the rank histogram analysis of Annan and Hargreaves (2010) indicating the model spread to be a little on the broad side for these variables. Collins et al. (2010) presents a comprehensive correlation analysis of a wide range of climatic variables, for several different ensembles of the Hadley Centre model and also for both the slab ocean and fully coupled versions of the CMIP3 multi-model ensemble. They find that the correlations for the CMIP3 ensembles appear to be generally clustered reasonably close to the value 0.5, but rather higher values are frequently found for most of the single model ensembles based on HadCM3/HadSM3. One exception to this trend is for a subset of runs of HadSM3 in which model parameters were deliberately chosen to explore previously untested regions of parameter space without reference to the quality of the model results. This suggests that these single model ensembles are generally clustered relatively far from the observations and are unlikely to be reliable in the sense of Annan and Hargreaves (2010). However, further analysis would be required in order to use this correlation analysis as a robust diagnosis of ensemble performance.

b. An application of statistically indistinguishable versus truth-centred interpretations: climate sensitivity estimated from the CMIP3 ensemble

It may not be widely recognised to what extent the two paradigms of statistically-indistinguishable and truth-centred ensembles, which have both been used in analysis of multi-model ensembles, lead to radically different interpretations of ensemble outputs. Therefore, here we illustrate this point with a simple calculation based on the climate sensitivity (i.e., the equilibrium global mean surface air temperature response to a doubling of atmospheric CO₂). Table 8.2 of Solomon et al. (2007) contains the climate sensitivities of 19 GCMs, a subset of the 24 models considered in this paper. If we adopt the truth-centred paradigm in which each model estimate is assumed to lie equiprobably and independently above or below the true value, then a simple combinatorial argument implies that between 6 and 14 samples (inclusive) should fall on each side of the truth, with at least 90% probability. This implies a central “very likely” confidence interval of 2.7–3.4C for the climate sensitivity. This analysis would of course represent a strong contradiction with the assessment of climate sensitivity that was actually presented in Chapters 9 and 10 of the same book.

The statistically indistinguishable paradigm, on the other hand, assigns an equal probability of 5% to each of the 20 intervals on the number line (including the semi-infinite ones) demarked by the 19 values. The central 90% “very likely” confidence interval, therefore, is precisely the full range of the values obtained, i.e. 2.1–4.4C. With such a small sample, however, the endpoints of the interval are determined entirely by the outlying samples which therefore makes this analysis rather sensitive to sampling error — for example, swapping the “experimental” MIROC model for the version of this model that was in fact used, would

change the upper bound of that interval to 6.3C. The implied 70% confidence interval of 2.3–4.3C is much more robust in respect of sampling error, and is rather close to the IPCC statement that climate sensitivity was “likely” to lie in the range of 2–4.5C. This analysis, together with others such as Fig TS.30 of Solomon et al. (2007) which is based on counting the level of agreement between models, suggests that the statistically indistinguishable paradigm is much more closely aligned with the actual beliefs and interpretations of climate researchers, whether or not they have clearly expressed it in those terms.

REFERENCES

- Abe, M., H. Shiogama, J. Hargreaves, J. Annan, T. Nozawa, and S. Emori, 2009: Correlation between Inter-Model Similarities in Spatial Pattern for Present and Projected Future Mean Climate. *SOLA*, **5** (0), 133–136.
- Adler, R., et al., 2003: The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979–present). *Journal of Hydrometeorology*, **4** (6), 1147–1167.
- Allan, R. and T. Ansell, 2006: A new globally complete monthly historical gridded mean sea level pressure dataset (HadSLP2): 1850–2004. *Journal of Climate*, **19** (22), 5816–5842.
- Annan, J. D. and J. C. Hargreaves, 2010: Reliability of the CMIP3 ensemble. *Geophysical Research Letters*, **37** (2), L02703.
- Bretherton, C., M. Widmann, V. Dymnikov, J. Wallace, and I. Bladé, 1999: The effective number of spatial degrees of freedom of a time-varying field. *Journal of Climate*, **12** (7), 1990–2009.
- Collins, M., B. B. Booth, B. Bhaskaran, G. R. Harris, J. M. Murphy, D. M. H. Sexton, and M. J. Webb, 2010: Climate model errors, feedbacks and forcings: A comparison of perturbed physics and multi-model ensembles. *Journal of Climate*, doi: 10.1007/s00382-010-0808-0.

- Giorgi, F. and L. Mearns, 2002: Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the “reliability ensemble averaging” (REA) method. *Journal of Climate*, **15 (10)**, 1141–1158.
- Gleckler, P., K. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *Journal of Geophysical Research-Atmospheres*, **113 (D6)**, D06 104.
- Jolliffe, I. and C. Primo, 2008: Evaluating Rank Histograms Using Decompositions of the Chi-Square Test Statistic. *Monthly Weather Review*, **136 (6)**, 2133–2139.
- Jones, P., M. New, D. Parker, S. Martin, and I. Rigor, 1999: Surface air temperature and its changes over the past 150 years. *Reviews of Geophysics*, **37 (2)**, 173–199.
- Jun, M., R. Knutti, and D. Nychka, 2008a: Local eigenvalue analysis of CMIP3 climate model errors. *Tellus*, **60 (5)**, 992–1000.
- Jun, M., R. Knutti, and D. Nychka, 2008b: Spatial analysis to quantify numerical model bias and dependence: how many climate models are there? *Journal of the American Statistical Association*, **103 (483)**, 934–947.
- Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl, 2009: Challenges in combining projections from multiple climate models. *Journal of Climate*, (Accepted).
- Lambert, S. J. and G. J. Boer, 2001: CMIP1 evaluation and intercomparison of coupled climate models. *Climate Dynamics*, **17**, 83–106.
- Smith, R., C. Tebaldi, D. Nychka, and L. Mearns, 2009: Bayesian modeling of uncertainty in

ensembles of climate models. *Journal of the American Statistical Association*, **104 (485)**, 97–116.

Solomon, S., D. Qin, M. Manning, Z. Chen, et al., 2007: Climate change 2007: The physical science basis. Contribution of the Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.

Tebaldi, C. and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **365 (1857)**, 2053.

Whetton, P., I. Macadam, J. Bathols, and J. O’Grady, 2007: Assessment of the use of current climate patterns to evaluate regional enhanced greenhouse response patterns of climate models. *Geophysical Research Letters*, **34 (14)**, L14 701, doi:10.1029/2007GL030025.

List of Figures

- 1 The probability that a single model is better than the multi-model mean, as a function of effective dimension. The models are drawn from the same distribution as the observations (dark blue line), or one that is narrower or wider by a factor of two (cyan and red respectively). 5% and 10% thresholds are indicated for convenience. Solid lines indicate isotropic distributions, dotted lines indicate geometrically-decaying eigenvalues and dashed lines indicate eigenvalues that decrease as $1/\sqrt{i}$. 26
- 2 Estimated effective dimension as a function of ensemble size. Coloured lines indicate results from CMIP3 ensemble data as shown. Black lines indicate synthetic results from isotropic distributions with 4, 6 and 11 dimensions (from bottom to top). 27

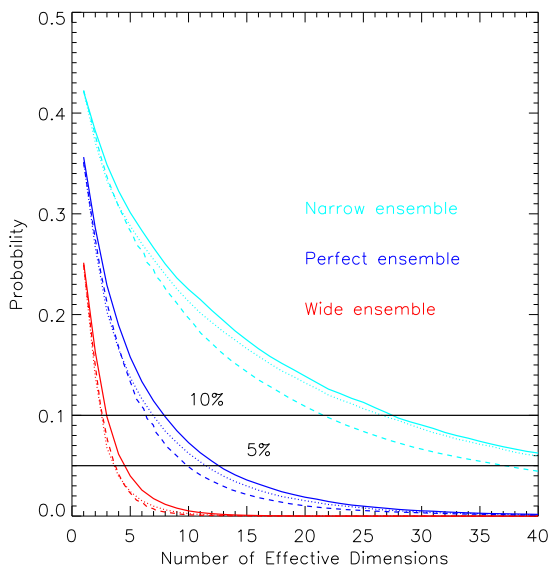


FIG. 1. The probability that a single model is better than the multi-model mean, as a function of effective dimension. The models are drawn from the same distribution as the observations (dark blue line), or one that is narrower or wider by a factor of two (cyan and red respectively). 5% and 10% thresholds are indicated for convenience. Solid lines indicate isotropic distributions, dotted lines indicate geometrically-decaying eigenvalues and dashed lines indicate eigenvalues that decrease as $1/\sqrt{i}$.

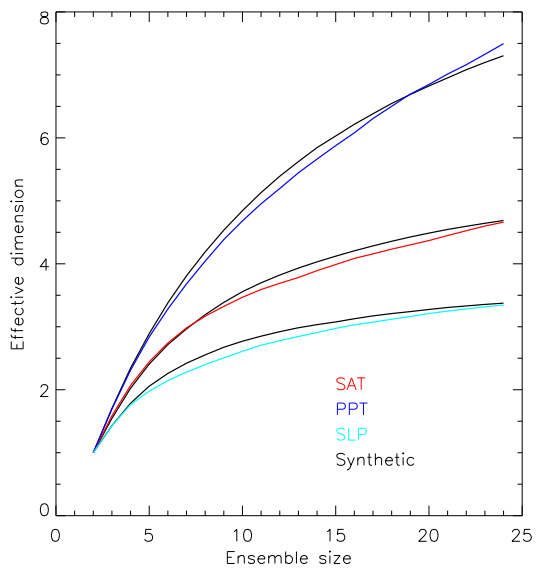


FIG. 2. Estimated effective dimension as a function of ensemble size. Coloured lines indicate results from CMIP3 ensemble data as shown. Black lines indicate synthetic results from isotropic distributions with 4, 6 and 11 dimensions (from bottom to top).