

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 focussing on the performance of the multi-model mean. We present an explanation of this phenomenon in the context of the statistically indistinguishable paradigm, and also provide a quantitative analysis of the main factors which control how likely the mean is to out-perform the models in the ensemble, both individually and collectively. Our analyses lend further support to the usage of 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

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 (Meehl et al. 2007). 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 primarily 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 (Knutti et al. 2010a,b).

One phenomenon that has been observed during comparisons with observational data is that the multi-model mean tends to have a lower RMSE than most, if not all, individual models (Lambert and Boer 2001). This result has been repeatedly replicated, and indeed highlighted, in subsequent research (Section 8.3, Randall et al. 2007; Gleckler et al. 2008; Reichler and Kim 2008; Pierce et al. 2009). One hypothesis that has been extensively explored in recent years is the paradigm of models being considered as independent samples from some distribution which is centred on the truth, as in this case the multi-model mean could be expected to converge to the truth as more models are added to the ensemble (Tebaldi and Knutti 2007, and references therein). However, this hypothesis is convincingly refuted by analysis of the model outputs (Knutti et al. 2010b). Therefore, a claim of truth-centredness can hardly be invoked as a basis for the good performance of the multi-model mean, and a convincing explanation for this phenomenon has yet to be presented.

In this paper, we present such an explanation. Our analysis runs primarily along lines motivated by geometrical considerations. Our analysis is presented within the framework of a statistically indistinguishable ensemble (Annan and Hargreaves 2010; Knutti et al. 2010a), that is, that the truth and models are considered as being drawn from the same distribution. This paradigm is commonly adopted in numerical weather prediction (Toth et al. 2003) where it dates back to the origins of predictability studies (Lorenz 1969). In these situations, uncertainty may be primarily attributed to the sampling of initial conditions within a single model framework. In contrast, in the context of climate change prediction over longer time scales, uncertainty in the climate system is primarily attributed to parametric and structural uncertainties. Therefore the interpretation of this paradigm here is that the range of different parameterisations and model structures sampled by the ensemble of climate

models is representative of our uncertainty regarding the behaviour of the real climate system. While this assumption is clearly optimistic (and surely not exactly true) it is implicit in the use of the ensemble spread as a direct representation of our uncertainty, such as Figures SPM.5 and SMP.7 of the IPCC AR4 Summary for Policymakers (Solomon et al. 2007). Alternatively, the ensemble may either be too narrow, such that reality lies outside its range with high probability, or it may instead be relatively broad, with reality close to the mean. Determining which of these cases is more likely would have important implications for our confidence in climate model outputs. Thus it is of fundamental importance to understand how such ensembles can be expected to behave in comparison with observational data. In Section 2, we present a theoretical investigation into the properties of an ensemble mean in comparison with observations, firstly showing why it will always outperform most of the ensemble members and further investigating the conditions which determine the probability with which it will outperform every sample in the ensemble. We demonstrate that the performance of the mean can be readily explained in terms of some fundamental parameters of the sampling distribution of the ensemble. We apply these theoretical insights to an analysis of the CMIP3 ensemble in Section 3, reconciling the observed behaviour of the ensemble with the properties of its distribution. Our analysis also raises some questions about the adequacy of the sample size. We summarise our results in Section 4. The Appendix contains a theoretical analysis of pairwise error correlations, explaining how results of this nature (which have been presented by others for climate model ensembles) can be interpreted in the context of the statistically indistinguishable paradigm.

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

We start with the elementary observation that for any ensemble, its mean will inevitably outperform a typical ensemble member in a comparison with any observational data set. 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:

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

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 sample mean. We note that this simple algebraic identity (which appears to have folkloric status within the wider ensemble prediction community, e.g. Stephenson and Doblas-Reyes (2000), but is perhaps not so widely known elsewhere) makes no appeal to any properties of the errors in the underlying physics of the models, nor how the ensemble was generated, and does not rely on any probabilistic interpretation of sampling distributions or use of an expectation operator, rather applying immediately to any finite sample. Furthermore, it does not even depend on where the observations happen to lie relative to the ensemble, and

therefore is not contingent on the oft-cited property that the errors of different models have a tendency to cancel. On the contrary, this result holds even if the signs of the errors are the same for all ensemble members.

This result is, however, limited in scope and does not address the perhaps more interesting phenomenon of the multi-model mean frequently outperforming *all* models in the ensemble. This phenomenon is not guaranteed to occur and 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 a better representation than every model for the surface air temperature and precipitation in both summer and winter seasons, but that a few individual models have a better representation of sea level pressure. Gleckler et al. (2008) analysed the CMIP3 ensemble, examining 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 (visual inspection of their Figure 3 suggests something of the order of 2% of models outperform the multi-model mean, across a wide basket of comparisons). Here we will call a model that is closer to a set of observations than the multi-model mean is (in terms of having a lower RMS error), a ‘nearer neighbour’ for those data.

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 $m_i - M$ (see Figure 1 for illustration). 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 investigate how the probability of an individual ensemble member being a nearer neighbour for the data may be affected by various factors relating to the sampling distributions of both the ensemble members and the data. Initially we test the perfect, statistically indistinguishable case where synthetic “models” and “data” are drawn from the same distribution. As in Annan and Hargreaves (2010), we use the multivariate d-dimensional standard Normal $N(\mu = 0, \sigma^2 = 1)^d$, but here we explore 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. We use an ensemble size of 25, but the results are very insensitive to this choice.

The results, plotted as the solid dark blue line in Figure 2, were calculated by repeated Monte Carlo sampling of both ensemble and observations. The Appendix of Palmer et al. (2006) presents essentially the same result, but using the true mean of the sampling distri-

bution in contrast to our use of the sample mean of a finite ensemble. It is immediately apparent that there is a strong dependence on the dimension of the sampling distribution. This can be understood through the fact that the expected value of $\|m_i - O\|$ is a factor $\sqrt{2}$ greater than that of $\|O - M\|$ by construction, and in high dimensions the sampling uncertainty around these expected values will shrink in relative terms, reducing the overlap in their sampling distributions and decreasing the probability that a sample from the latter may be greater than one from the former. Note, however, that these vectors are not probabilistically independent, so this argument is not rigorous.

The results in Figure 2 can easily be used to calculate the probability of the mean being better than all models, or more generally the probability that k models from a sample of n are better than the mean, through the binomial distribution. This probability is given by $p^k(1 - p)^{n-k} \binom{n}{k} = p^k(1 - p)^{n-k} n! / k!(n - k)!$ where p is the probability (plotted in the figure) of a random model being better than the mean. For example, with a dimension of 10 and sample size of 25, the probability of the mean being better than all models is $(1 - 0.08)^{25} \simeq 12\%$.

We next relax the assumption of a statistically indistinguishable ensemble by scaling the sampling distribution of the models by a constant which is either smaller than, or greater than, unity, while keeping the sampling distribution of the observations fixed. These two choices correspond to the case of a uniformly underdispersive or overdispersive ensemble — that is, an ensemble that is respectively either too narrow, such that the observations will tend to lie in the tails or outside the ensemble range, or too broad, such that the observations are closer to the ensemble mean than a typical model will be. The base case of the statistically indistinguishable ensemble is contrasted in Figure 2 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 compared to the sampling distribution of the observations (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. Figure 1 illustrates graphically why this is the case: the region, within which a model would be a nearer neighbour to the data, grows as the data move further away from the ensemble mean. 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.

For a truth-centred ensemble (magenta line in Figure 2), which can be considered as an extreme case in which the ensemble is infinitely overdispersive, it is obviously impossible for a model to be closer to the observations than the true mean of its sampling distribution is. In this extreme case, the rôle of the sampling error on the ensemble mean becomes critical. For our ensemble size of 25, this probability is around 1% for a dimension of 3, and rapidly becomes vanishingly small at higher dimensions.

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 variation can generally be described by a relatively small number of empirical orthogonal functions (EOFs). Therefore, we now present results from experiments which use two families

of sampling distributions which are anisotropic in the metric space, to investigate how this anisotropy may affect the properties of the ensemble mean and in particular the probabilities of finding nearer neighbours. In order to provide a compact and meaningful comparison across different families of sampling distribution, we use the concept of effective degrees of freedom as an indication of the number of EOFs that contribute significantly to the overall variance of the ensemble. To calculate the effective degrees of freedom, we use the formula 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 of the total variance, then the number of effective dimensions N_{ef} is defined as $N_{ef} = 1/\sum f_i^2$. For the isotropic case presented above, the number of effective dimensions according to this formula equals the number of true dimensions d . In our first set of experiments, we use a distribution in which the eigenvalues of the covariance matrix λ_i drop off geometrically, $\lambda_i/\lambda_{i-1} = k$ for some ratio $k < 1$ (this being a typical spectrum in geophysical applications). 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 coefficient is scaled so as to be sampled from $N(\mu = 0, \sigma^2 = k^i)$. The dotted lines in Figure 2 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 an alternative case where the eigenvalues of the covariance matrix decrease as $\lambda_i \propto 1/i$. 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 true dimensions of the sampling distribution. For this

spectrum 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 2. This distribution of eigenvalues is not typical of geophysical applications, but is included to test the sensitivity of our results to a wide range of scenarios. For a more rapid decay such as $\lambda_i \propto 1/i^2$, 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 eigenvalue spectrum irrelevant to our investigations.

From these experiments, which sample a wide range of eigenvalue spectra, 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 most realistic geometric case having probabilities that generally lie between the other two cases. 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.

3. CMIP3 analysis

The theoretical framework developed in Section 2 is now applied to outputs from the ensemble of CMIP3 models. We first explore the nearer neighbour phenomenon and then continue with a more detailed analysis of the effective dimension of the ensemble, based on an EOF decomposition. We note 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, whereas the multi-model mean represents the simplest assumption of “model democracy” where each ensemble member receives equal weight. Therefore, the

good performance of the multi-model mean, and the existence (or otherwise) of models which outperform the multi-model mean, would appear to 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 address 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, especially for these choices of relatively well-observed and well-understood variables. We also note that to the extent that the observational error vector is orthogonal to the multi-model differences arising through differences in parameterisations, such errors cannot in fact affect whether a particular model is a nearer neighbour for the data.

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 generally comparable to those presented by Lambert and Boer (2001) and Gleckler et al. (2008), although differences in data sets and model ensembles preclude an exact match. Given the sample size of 24 in each case, our 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 found with the observational data appear compatible with the paradigm of a statistically indistinguishable ensemble. Furthermore, since the multi-model ensemble members are exchangeable by def-

initiation, reference to the dark blue lines on Figure 2 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 2 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 2, 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 now try 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 $O(1-2000\text{km})$. This value was estimated from semi-variograms of model errors, and is 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 variability in the atmospheric state, but rather depend on the underlying physical parameterisations. For any one model, the basic physical parameterisations are the same across the globe and thus for any pair of models, their differences may be expected to persist over widely dispersed areas with similar climates, reducing the effective dimension substantially. The leave-one-out analysis of Annan and Hargreaves (2010) suggested a much lower alternative value of around 5 dimensions for the model fields examined. For this value, the non-uniformity of the rank histograms in that paper would not be statistically significant.

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 2, a

choice of 40 dimensions suggests that if the ensemble really was statistically indistinguishable from the observations, we should very likely ($p > 70\%$) find zero nearer neighbours for all three data sets. Therefore the proportion we obtained, of 8%, would suggest that the ensemble is too narrow by a factor of 2 or more. Conversely, the lower figure of 5 dimensions should result in at least 15% of models being nearer neighbours to the data, and in this case the much lower observed frequency would imply that the ensemble is instead rather too broad (though probably not by as much as a factor of two). 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 address this in more detail through an EOF decomposition of the spread of the ensemble of model climatologies around the multi-model mean. In accordance with the data we are using for comparison, we consider the EOFs of the space spanned by each set of the (area-weighted) two-dimensional data fields in turn.

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. 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 alone.

The leading N_{ef} EOFs represent roughly 80% of the total variance of the model ensemble in each case (in reasonable accordance with the exponential spectrum), but when we project the observational anomalies onto these EOFs, we find that these first few EOFs only

represent 40–65% of their variance. Even the full set of EOFs only account for 54–87% of the observational variance for the three data sets, implying that the observations contain substantial variation that is orthogonal to the space spanned by the multi-model ensemble. These results might appear to cast doubt on the hypothesis of statistical indistinguishability. However, the results are actually consistent with what we obtain through leave-one-out validation: when one model is withheld and used in place of observations, 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 model respectively, and in many cases, the proportion of variance that remains unexplained is greater than was found in the case of the real observations.

One plausible explanation of this result is that the ensemble size may simply be too small to adequately sample all the important directions of the underlying distribution. The estimate of N_{ef} obtained via the method used here is necessarily bounded above by the sample size irrespective of the effective dimension of the underlying sample distribution, and even for a moderately large sample size the estimate will generally have a low bias. Equation 14 of Bretherton et al. (1999) suggests that for a sample of 24 and true effective dimension of 4–11, the estimation of N_{ef} will have a low bias of around 15–30% (with the larger value applying to the higher dimension), with additional uncertainty of about 10% at one standard deviation around that value. Based on this estimate of mean bias, the N_{ef} values that we found for CMIP3 would correspond to distributions which have actual dimensions of 6, 11 and 4 respectively.

This interpretation is supported when we check how the effective dimension of the ensemble varies when we use subsets of different numbers of models. Figure 3 shows the estimated

effective dimension of different sized subsets of models (averaged over 1000 randomly-selected subsets of each size), for the three variables separately. The effective dimension increases steadily with sample size, and appears to still be some way from convergence at the full sample size of 24 for all of the three cases. This figure also shows equivalent results generated from random samples from the isotropic distribution $N(0, 1)^d$ used in Section 2a (the other families of distributions generate similar results). It seems that underlying effective dimensions of 6, 11 and 4 do indeed provide good fits to the observationally-derived results. This supports our suspicion: the CMIP3 ensemble behaves as if the models were drawn from a space with a rather larger dimension, which the finite sample size of 24 is inadequate to fully describe. Thus, it is not surprising that the observations contain a significant component which lies outside the space spanned by this modest sample. Encouragingly, the underlying effective dimensions estimated via this method are similar to those obtained by the nearer neighbour analysis in Section 3a.

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 estimated to be 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.

To summarise this analysis, our nearer neighbour analysis of the CMIP3 database sup-

ports our previous hypothesis that the statistically indistinguishable paradigm provides a reasonable basis for explanation of the properties of the CMIP3 ensemble (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. Using these values, the non-uniformity for the rank histograms of Annan and Hargreaves (2010) is no longer significant at the $p < 5\%$ level. As noted in Section 2, the proportion of nearer neighbours found by Gleckler et al. (2008) over a wide range of comparisons appears to be rather lower than observed here, at approximately 2% of models. This could suggest either that the ensemble is generally overdispersive when examined over their broader perspective, or alternatively the effective dimension of the ensemble for these different data sets could be higher than for the three data types considered here.

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 perhaps, we would not interpret their results as inconsistent with the statistically indistinguishable paradigm, since their single data point (their Table 3) only lies at the 10th percentile of the rank histogram. Of course, the statistically indistinguishable paradigm is not literally true, rather the question is whether it is a reasonable approximation to adopt in the use of the ensemble. Further testing of the credibility of this approximation is critical as it underpins the use of the ensemble as a measure of uncertainty.

4. Conclusions

We have clarified and explained the performance of the multi-model mean in the context of the statistically indistinguishable paradigm. In particular, we have shown that the observed (small but non-zero) proportion of models which are better than the mean strongly depends on both the effective dimension of the sampling distribution, and the relative widths of the sampling distributions of truth and models.

If the ensemble was substantially underdispersed, such that reality was often found towards the tails of the distribution or outside its spread, then more such models would be expected. If anything, the small number of these models suggests an ensemble that is somewhat overdispersive, but not to such an extent that the truth can reasonably be considered as lying at the ensemble mean. In this latter case, nearer neighbour models would be so rare as to be virtually nonexistent. Our analysis 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. However our results are only directly applicable to the specific comparisons presented here. There is no guarantee that the ensemble will perform so well for other data sets or time periods, especially future changes.

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 significantly when all data types are combined, but might be expected to if future time periods were also taken into account, 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, indicates 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. According to the estimate of Bretherton et al. (1999), if the underlying sample distribution had 10 effective dimensions, we would require 90 models before the effective dimensionality of the sample would be expected to reach 90% of the correct value. It might be worth considering whether model design could be targetted to sample these climates more efficiently than at present.

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. We are grateful to four reviewers, and numerous attendees at several seminars, for their helpful comments.

APPENDIX

How can we interpret pairwise error correlations in the context of the statistically indistinguishable paradigm?

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 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. (2010b) 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. 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). 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. However, further analysis would be required in order to use this correlation analysis as a robust diagnosis of ensemble performance.

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.

- 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., G. Abramowitz, M. Collins, V. Eyring, P. Gleckler, B. Hewitson, and L. Mearns, 2010a: IPCC Good Practice Guidance Paper on Assessing and Combining Multi Model Climate Projections. IPCC Working Group I Technical Support Unit, University of Bern, Bern, Switzerland.
- Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl, 2010b: Challenges in combining projections from multiple climate models. *Journal of Climate*, **23**, 2739–2758.
- Lambert, S. J. and G. J. Boer, 2001: CMIP1 evaluation and intercomparison of coupled climate models. *Climate Dynamics*, **17**, 83–106.

- Lorenz, E., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21 (3)**, 289–307.
- Meehl, G., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. Mitchell, R. Stouffer, and K. Taylor, 2007: The WCRP CMIP3 multimodel dataset. *Bull. Am. Meteorol. Soc.*, **88**, 1383–1394.
- Palmer, T., R. Buizza, R. Hagedorn, A. Lawrence, M. Leutbecher, and L. Smith, 2006: Ensemble prediction: A pedagogical perspective. ECMWF Newsletter 106.
- Pierce, D., T. Barnett, B. Santer, and P. Gleckler, 2009: Selecting global climate models for regional climate change studies. *Proceedings of the National Academy of Sciences*, **106 (21)**, 8441.
- Randall, D. A., et al., 2007: *Climate Models and Their Evaluation*. In: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, chap. 8. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.
- Reichler, T. and J. Kim, 2008: How well do coupled models simulate today’s climate? *Bulletin of the American Meteorological Society*, **89 (3)**, 303–311.
- 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.
- Stephenson, D. and F. Doblas-Reyes, 2000: Statistical methods for interpreting Monte Carlo ensemble forecasts. *Tellus A*, **52 (3)**, 300–322.

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.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Forecast Verification: A Practitioners Guide in Atmospheric Science*, 137–163.

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 Graphical representation of relationship between model ensemble, multi-model mean and truth. Black crosses indicate models, with ellipses showing isolines of the sampling distribution. M (black diamond) marks multi-model mean, O is observations in the case of statistically indistinguishable (blue), overdispersive (O', red) or underdispersive (O'', cyan) ensembles. Coloured circles indicate domains where a model will be a nearer neighbour to the observation. 28
- 2 The probability that a single model is better than the multi-model mean, as a function of effective dimension. An ensemble of 25 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). Magenta indicates the truth-centred case. 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 $1/i$. 29
- 3 Estimated effective dimension as a function of ensemble size. Coloured lines indicate results from CMIP3 ensemble data as shown, averaged over 1000 model subsets for each point. Black lines indicate equivalent synthetic results from isotropic distributions with 4, 6 and 11 dimensions (from bottom to top respectively). 30

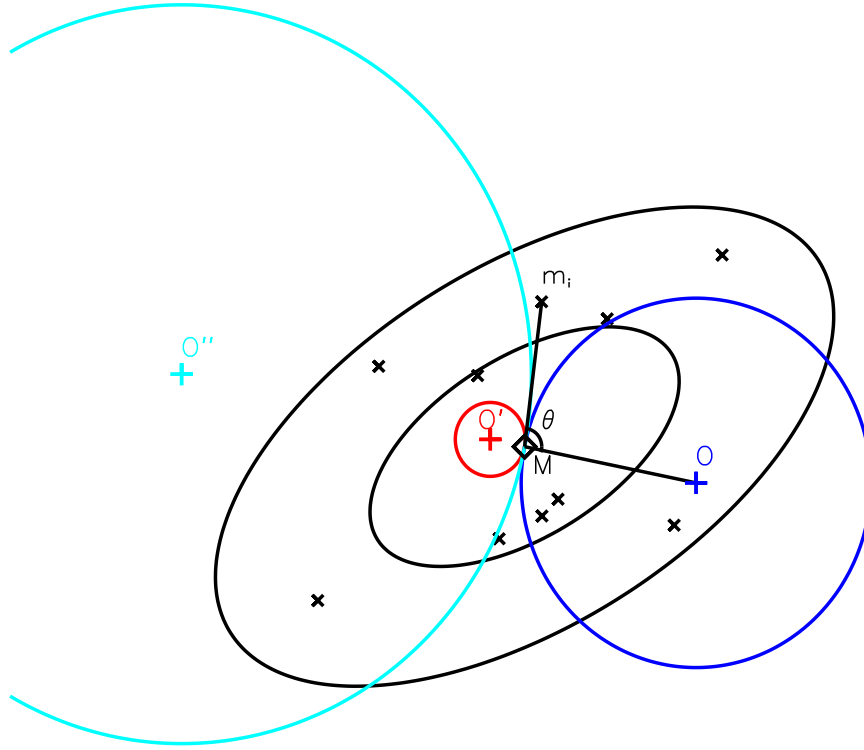


FIG. 1. Graphical representation of relationship between model ensemble, multi-model mean and truth. Black crosses indicate models, with ellipses showing isolines of the sampling distribution. M (black diamond) marks multi-model mean, O is observations in the case of statistically indistinguishable (blue), overdispersive (O' , red) or underdispersive (O'' , cyan) ensembles. Coloured circles indicate domains where a model will be a nearer neighbour to the observation.

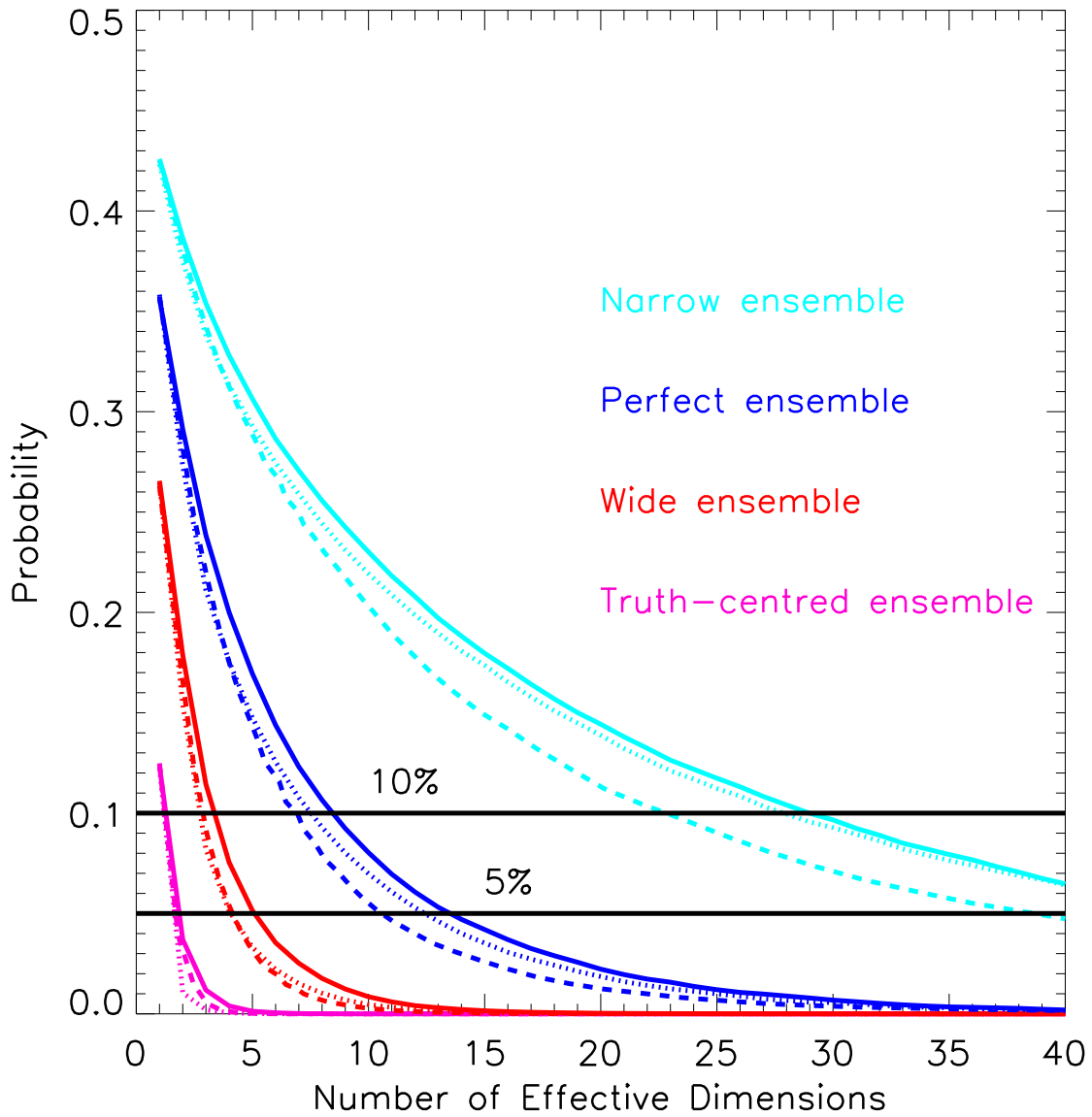


FIG. 2. The probability that a single model is better than the multi-model mean, as a function of effective dimension. An ensemble of 25 models are drawn from the same distribution as the observations (dark blue line), or one that is narrower or wider by a factor of two (cyan and red respectively). Magenta indicates the truth-centred case. 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 $1/i$.

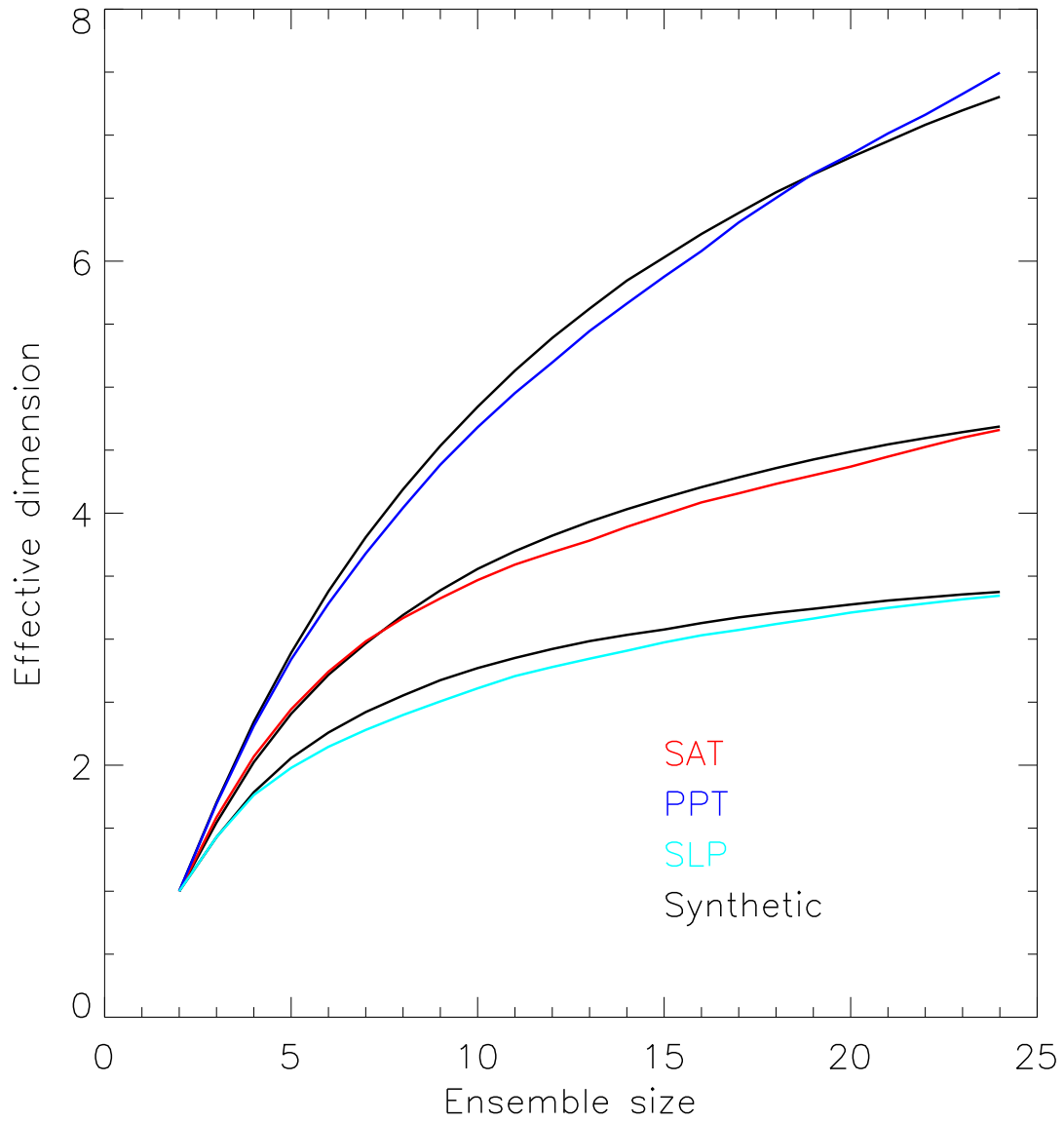


FIG. 3. Estimated effective dimension as a function of ensemble size. Coloured lines indicate results from CMIP3 ensemble data as shown, averaged over 1000 model subsets for each point. Black lines indicate equivalent synthetic results from isotropic distributions with 4, 6 and 11 dimensions (from bottom to top respectively).