The DOI for this manuscript is doi: 10.1175/JCLI-D-14-00362.1

A Representative Democracy to reduce interdependency in a multi-model ensemble

Benjamin M. Sanderson *

National Center for Atmospheric Research, Boulder CO

Reto Knutti

Institute for Atmospheric and Climate Science, ETH, Zurich, Switzerland

Peter Caldwell

Lawrence Livermore National Laboratory, Livermore CA

* *Corresponding author address:* Benjamin Sanderson, National Center for Atmospheric Research, 1850 Table Mesa Dr, Boulder CO, 80305, USA

E-mail: bsander@ucar.edu

ABSTRACT

The collection of Earth System Models available in the CMIP5 archive represents, at least to some degree, a sample of uncertainty of future climate evolution. The presence of duplicated code as well as shared forcing and validation data in the multiple models in the archive raises at least three potential problems; biases in the mean and variance, the overestimation of sample size and the potential for spurious correlations to emerge in the archive due to model replication. Analytical evidence is presented to demonstrate that the distribution of models in the CMIP5 archive is not consistent with a random sample, and a weighting scheme is proposed to reduce some aspects of model co-dependency in the ensemble. A method is proposed for selecting diverse and skillful subsets of models in the archive which could be used for impact studies in cases where physically consistent joint projections of multiple variables (and their temporal and spatial characteristics) are required.

1. Introduction

Today's Earth System Models (ESMs) are great testament to collaborative scientific thinking. Millions of lines of computer code represent the pinnacle of understanding of the intricate coupled interactions of the Earth's land, oceanic, cryospheric, and atmospheric systems. Unlike the more simple atmospheric models of the past, few people (if any) now understand the models in their entirety and so the models themselves have become vehicles of a scientific consensus which we use to project future climates which cannot directly be validated for decades to come. For some parts, such as the representation of the equations of fluid flow, understanding is mature and thus (relatively) uncontentious. But other components, such as the effect of a changing climate on ecosystem dynamics, are sufficiently complex that any computational code must inevitably make significant approximations in order to even represent the bulk behavior of the system in any tractable fashion.

A given model is thus more than a computer program, it is a collection of axioms and beliefs about which processes might be important for evaluating how our environment might change, and how those processes should be represented, and as such, each model is a self-consistent entity. The challenge arises, however, when one wishes to combine the results of many models to attain some more comprehensive understanding of the uncertainties present in their individual implementation. Given a set of models of the climate system, assessing the value of adding another model clearly requires a consideration of whether the model is fit for purpose (e.g. the validity of its axioms, forcing data and tuning protocols). We would argue also that it is important to assess if the model provides new information; to measure how independent is the new model from those in the original set. In an extreme case, adding an exact duplicate of a model already in the set would not add value, rather it would bias any combination of model results towards the results of the duplicated model (Caldwell et al. 2014).

The latest Coupled Model Inter-comparison Project (CMIP5, Taylor et al. 2012) is the largest archive of climate data the world has seen to date. Such Multi Model Ensembles

(MMEs) have often been referred to as 'ensembles of opportunity' (Tebaldi and Knutti 2007), because the range of models represent some sample of the systematic choices which developers face in the course of representing the climate system in the form of computer code. But, as has been noted before (Knutti 2010), this sample is far from perfect.

Firstly, the models available may vary in their ability to resolve certain processes which might be observed in the Earth System. For any given process, a researcher may find relevant observations to rank models for their purposes but the output of the ESMs is sufficiently high dimensional that any ranking is unlikely to be universal (Santer et al. 2009). In contrast to weather forecast models, ESMs can also rarely be validated out of sample and so there remains a risk that empirical components of ESMs can be calibrated using the only available observations, and although this might be a pragmatic approach it leaves little opportunity for assessing and contrasting model performance (Sanderson and Knutti 2012).

A second problem lies in the lack of independence of models, where independence is not meant in a statistical sense but in a more loose sense of models sharing ideas for parameterizations and simplifications or sharing actual computer code, and therefore being biased in similar ways relative to reality. At the time of writing, 61 models are listed in the Earth System Grid database. This doesn't necessarily mean that each of these models provides an independent estimate of future climate change. Indeed, some of these co-dependencies are trivial and can be accounted for by considering models submitted with different resolutions (for example, MPI-ESM-MR and MPI-ESM-LR, see Knutti et al. 2013). Most institutions also produce model variants with a range of different configurations, with options for interactive atmospheric chemistry or carbon cycle (CMCC-CESM and CMCC-CM, for example). Finally, different institutions can share model components, for example the FIO-ESM model shares its atmosphere, ocean, sea ice and land surface code with CCSM4, but adds a surface ocean wave parameterization. Submodel replication is common throughout the ensemble, for example in the models considered for this study over 25 percent use some variant of the Community Atmosphere Model (CAM3, CAM3.5, CAM4 or CAM5) to represent atmospheric processes. The GFDL MOM ocean model is similarly popular (MOM2.2, MOM4.0 and MOM4.1). Table 2 shows a broad illustration of shared model components in the CMIP5 models considered for this study.

This extensive model replication in the CMIP5 and its predecessors is not a problem *per se*, in fact it seems natural to copy successful parts and build on the work of others, and it requires enormous effort to develop entirely new model components. Hence, each institution understandably focuses on certain aspects but copies other components. But model replication presents a number of issues for model ensemble analysis. The first is simply a matter of representation: the Assessment Reports of the Intergovernmental Panel on Climate Change (IPCC) have often used the multi-model mean of the CMIP ensembles to represent a consensus view of model projections of future climate, but clearly this mean will be biased towards models which are highly replicated within the ensemble. Similarly, model agreement on the sign or magnitude of a change in future climate is often taken to imply confidence in a result (Tebaldi et al. 2011, Knutti and Sedáček 2013), but if models are highly replicated within the ensemble, such agreement becomes less significant.

Another issue lies in the possible effect of replicated models in studies which attempt to constrain aspects of future climate change. If a researcher discovers a correlation between an observable quantity and some unknown climate parameter in a multi-model ensemble (such as in Fasullo and Trenberth 2012 or Qu and Hall 2013), the statistical significance of that correlation would be inflated if some points are repeated. This argument is developed in Caldwell et al. (2014) who show that although a data-mining approach will yield more strong correlations between Climate Sensitivity and potentially observable fields than one would expect to see by chance in CMIP5, this may be attributable in part to model co-dependencies.

This is the second in a series of papers examining interdependency in the CMIP ensembles. In Sanderson et al. (submitted), we developed a distance metric which enabled both models and observations to be represented as points in a multi-dimensional space. We then showed that model properties could be interpolated within this space, allowing a resampling of model properties in a manner which was less sensitive to model replication and could take into account a measure of performance in reproducing observations. However, the approach of Sanderson et al. (submitted) is also unable to provide full spatial and temporal variations in quantities. For example, a farmer may not want an estimate of the change in average rainfall, but a set of representative summers with full spatial and temporal information, and the corresponding temperature, sunshine and wind data. For such cases it may be better to use the raw or bias corrected model output directly, but that requires selecting a set of models to use.

It has been proposed before that subsets of larger ensembles may produce more statistically robust results, Evans et al. (2013) investigated this concept using subsets of a multi-physics ensemble of weather forecasting models. Perhaps the simplest approach to achieve this might be to take a single model from each institution, but there are numerous issues with this. Firstly, although there are often similarities between models published by single institutions, such a crude approach would eliminate cases where significantly different models were produced by the same group. There are several examples of the latter case, the GISS-E2 model, for example is published with two structurally different oceans. Furthermore, several groups (CESM, GFDL, UKMO amongst others) publish both a 'bleeding edge' model and a legacy model to the archive, where there might be significant structural changes between the releases. Finally, an institution-based pruning approach would not help identify models from different institutions which share a large fraction of their code.

It could be argued that one could account for many of these problems through careful consideration of model lineages, by documenting the basic parameterizations shared by different models or by assessing the fraction of common code between different models. This, however, would be a considerable undertaking - and the results would require a comprehensive understanding of each model's code. Firstly, although some models document and publish their code-base in full before submitting simulations to the CMIP archive, this practice is far from universal. A model could in theory be defined by summarizing the parameterizations, their values and other structural assumptions which have been employed in that model, but assessing the relative importance of each of those parameterizations in terms of model climatology or response to external forcing would require good prior intuition of the relationships between the parameterizations and the process to be studied, which might be possible in some but not necessarily all cases. Such an approach would clearly be worthwhile, and could greatly aid in the interpretation of differences in climate change projections, but it would be a monumental undertaking.

An alternative approach is to utilize output from the models themselves to establish codependencies. This approach has been demonstrated with some promise by Masson and Knutti (2011) and Masson and Knutti (2013), who used inter-model distances derived from spatial patterns of climatological temperature and precipitation to establish a hierarchical clustering of models which resembles a tree showing structural relationships one might expect from considering model lineages. As noted in Masson and Knutti (2011) and Sanderson et al. (submitted), the distribution of inter-model distances shows recognizable structure, with models from the same institution and models with common heritage generally exhibiting similar patterns of mean state bias. However, the aforementioned studies did not establish any quantitative assessment of inter-model distance, which we attempt to address here.

To this end, we formalize an approach to use model similarity information to select models based on their skill and independence. This does not eliminate model inter-dependency, but allows us to select a subset of models where the most glaring examples of model replication are no longer present. In Section 2.a, we establish a method for identifying near-neighbors in a climate model ensemble, in Section 2.d, we use model similarity information to produce a weighting scheme which accounts for both model skill and model interdependence. Section 2.e shows how this framework can be used to select a subset of models from an archive of climate models. Finally, Section 3.b demonstrates this method using the CMIP5 multi-model archive.

2. Method

2.a. *Processing model output*

In this study, as in our accompanying paper Sanderson et al. (submitted), we produce a matrix of inter-model distances in an EOF space derived from 30 year mean climatological output from each model's historical simulation conducted for CMIP5. The details of the construction of the distance matrix are identical to that of Sanderson et al. (submitted). We use the 'historical' and 'rcp85' experiments, and the 'r1i1p1' simulations in each case. In the special case of CCSM4, we also consider the sensitivity of the technique to internal variability by repeating the analysis with all available simulations in the CMIP5 archive (r1i1p1, r1i2p1, r1i2p2, r2i1p1, r3i1p1, r4i1p1, r5i1p1 and r6i1p1 for the historical runs and r1i1p1, r2i1p1, r3i1p1, r4i1p1, r5i1p1 and r6i1p1 for the RCP8.5 simulations).

The input data for this study is both processed, and used to conduct an EOF analysis in a similar fashion to Sanderson et al. (submitted). Minor differences in the inter-model distances occur because the former study considers both CMIP3 and CMIP5 models, which slightly changes the exact form of the EOFs. For each model, a number of monthly, gridded diagnostic variables are considered to represent the climatology of the model. For each available model in the CMIP3 and CMIP5 ensembles, monthly climatologies are obtained from a single historical simulation by averaging monthly mean fields for the time period 1970-2000. Data is obtained for five 2 dimensional fields (surface air temperature (TAS), total precipitation (PR), outgoing top-of-atmosphere shortwave radiative flux (RSUT), outgoing longwave top-of-atmosphere flux (RLUT), sea level pressure (PSL)) and two three-dimensional fields (atmospheric temperature (T) and relative humidity (RH)). Three dimensional fields are zonally averaged. Corresponding observational monthly mean climatologies are obtained by averaging available years for each field type, as shown in Table 1.

Data from each model and dataset are regridded onto a 2.5 by 3.75 degree latitude longitude grid, and zonal vertical fields are regridded onto a 2.5 degree latitude grid at 17 pressure levels. For each variable, values are area weighted. Vertically resolved fields are also weighted by the pressure difference between the top and bottom of the corresponding level. In order to usefully concatenate the multivariate field for EOF analysis, the variables must be normalized for each to represent a similar amount of variance in the multi-model ensemble. We normalize each observable field using values obtained from the observations. For 2 dimensional fields, we calculate the inter-monthly variance of tropical grid-cells and take the average over the tropics to obtain a single normalization factor for each variable. For 3 dimensional fields, we take the inter-monthly variance of zonally averaged fields in the tropics between 700 and 400 hPa, and then average the variances over the spatial domain to obtain the normalization factor. Normalization factors are calculated from the observations only, and the fields from each model are divided by the same factor (shown in Table 1). Each field is then reformulated into a single vector. If any elements of the vector in any single model or in the observations are missing, those particular elements are removed from all models. Each field vector is then normalized by the number of remaining elements, and the 2d and 3d fields are concatenated into a single vector length n (where n=358,248 when all fields are utilized). Each of the m vectors are combined to form a matrix X^{20c} (size m by n, where m is 36, comprising 36 CMIP5 model vectors). The ensemble mean value is calculated by averaging the m rows of the matrix, and this is subtracted from each row to yield the anomaly matrix ΔX^{20c}, such that

$$\Delta X^{20c} = X^{20c} - \overline{X^{20c}}. \tag{1}$$

The analysis is also repeated with a number of different subsets of the entire set of variables. In these cases, the matrix ΔX^{20c} is formed using only that subset, and the analysis continues in the same fashion.

The process is repeated to produce a similar matrix to represent the climate change between the historical simulation (1970-2000) and the RCP8.5 simulation (2070-2100). In this second analysis, the anomaly between the two 30 year periods is taken to form the matrix ΔX^{21c}. The future analysis is also repeated with a number of different subsets of the entire set of variables. In these cases, the matrix ΔX^{21c} is formed using only that subset, and the analysis continues in the same fashion.

2.b. Principal Component Analysis

We conduct a principal component analysis on the resulting matrix formed by combining the climatology vectors from each participating model, such that the EOF loadings define a t-dimensional space (where t is the truncation length of the Principal Component Analysis) in which inter-model and observation-model Euclidean distances may be defined. The use of the EOF pre-filter combines fields which are trivially correlated (such as adjacent grid-cells) into a single mode. The results of the analysis do change in a subtle fashion with truncation length, and we discuss this sensitivity further in Section 3.c.1, but for the initial analysis we use a truncation length of $t = 9$. This truncation length effectively provides enough degrees of freedom to represent some subtle differences between related models in the resulting distance metric, but not so many as to introduce excessive random noise into the calculation.

The PCA analysis on any ΔX can be performed by singular value decomposition and truncated to t modes, such that:

$$\Delta X^{20c} = U^{20c} \lambda^{20c} V^{20c^T}, \tag{2}$$

for the present day case ($20c$), and

$$\Delta X^{21c} = U^{21c} \lambda^{21c} V^{21c^T}, \tag{3}$$

for the future case ($21c$). U^{20c} and U^{21c} (sized m by t) are matrices of model loadings, V^{20c} and V^{21c} (sized n by t) are spatial patterns of ensemble variability while λ^{20c} and λ^{21c} (sized t by t) are diagonal matrices representing the variances associated with each mode.

The inter-model distances can then be measured in a Euclidean sense in the loadings matrices U^{20c} and U^{21c}, such that the distances between 2 models i and j can be expressed as:

$$\delta_{ij}^{20c} = \left(\sum_{l=1}^{t} \left(U^{20c}(i,l) - U^{20c}(j,l) \right)^2 \right)^{1/2}, \tag{4}$$

for the present day and

$$\delta_{ij}^{21c} = \left(\sum_{l=1}^{t} \left(U^{21c}(i,l) - U^{21c}(j,l) \right)^2 \right)^{1/2}, \tag{5}$$

for the future. Model-observation distances $\delta_{i(obs)}^{20c}$ which can obviously only be calculated for the present day case are created using a climatological vector from an observational dataset X^{obs} prepared in the same fashion as X^{20c}:

$$\Delta X_{(obs)n}^{20c} = X_{(obs)} - \overline{X^{20c}} \tag{6}$$

where $\overline{X^{20c}}$ is the multi-model mean of X^{20c}, length n. This observational anomaly vector can be projected onto V^{20c} to form an observational loading vector $U_{(obs)}$ (length t). The distance between each model and the observations can be then calculated in a similar fashion:

$$\delta_{i(obs)}^{20c}(i) = \left(\sum_{l=1}^{t} \left(U^{20c}(i,l) - U_{(obs)}^{20c}(l) \right)^2 \right)^{1/2}, \tag{7}$$

Finally, we calculate the variability expected in an initial condition ensemble by taking $n_{ic} = 8$ (historical) or $n_{ic} = 6$ (future) member CCSM4 ensemble for both the historical simulation and RCP8.5. In each case, the data is processed in the same fashion as for the multi-model case to create an n_{ic} by n matrix, X_{ic}^{20c} and X_{ic}^{21c}. We then take anomalies from the CMIP5 ensemble mean:

$$\Delta X_{ic}^{20c} = X_{ic}^{20c} - \overline{X^{20c}}, \tag{8}$$

and

$$\Delta X_{ic}^{21c} = X_{ic}^{21c} - \overline{X^{21c}}. \tag{9}$$

These can also be projected onto V^{20c} and V^{21c} to form loading vectors $U_{(ic)}^{20c}$ and $U_{(ic)}^{21c}$ (size n_{ic} by t). The distance between initial condition ensemble members can be then calculated as before for the multi-model case.

2.c. Forming Random ensembles

In order to compare inter-model distances in the CMIP5 archive with distances expected by chance, we create a set of 10^5 matrices of random data with the same dimensions as U^{20c} and U^{21c} (where m is 36). Each random distribution represents inter-point distances for all possible pair-wise combinations m points (703 distances, in this case). Our results are not sensitive to further increasing the number of random cases.

Each row of one of these random matrices is populated with draws from a Gaussian PDF with variance equal to that from the rows of U^{20c} and U^{21c} (all of the rows have equal variance in each case). As a result, data in these random matrices is independent in directions corresponding to both the EOF number and the model number. We desire matrices with an independent model dimension in order to test the likelihood that CMIP5 output was drawn from a set of independent models. Having independence in the field direction is appropriate because the columns of U^{20c} and U^{21c} are independent by construction.

Our assumption that the t dimensional normal distribution is representative of an independent ensemble of climate projections is subject to some caveats; we are making the effective assumption that a normal distribution of models in the space defined by U^{20c} or U^{21c} is plausible, and that there are no parts of that space which might represent an unphysical climate state. There are some justifications for this assumption; the random distributions are compared with the loading matrices U^{20c} and U^{21c}, which are themselves orthogonal basis sets defined by multi-model variability. As such, we are making the assumption that if there are physical relationships between variables in the model output data (say between adjacent grid-cells or between surface temperature and outgoing longwave radiation for example), then any correlation between these would be represented as a single mode in the EOF analysis. Thus, any linear relationships which exist in the original data are effectively preserved in the random ensemble also. However, a strong nonlinear relationship between two variables in the CMIP5 archive could not be represented in a single EOF mode, and might be represented in two or more modes. In this case, then there would be some of the space which should physically off-limits. Hence, by using normally distributed data to define the random ensembles and their associated length scale for inter-point distances, we make the assumption that multi-model variability can be appropriately described by a linear basis set. Although one could potentially consider designing a random sample which fitted a high-dimensional distribution to the existing ensemble to account for nonlinear relationships between modes, the increase in complexity, the lack of samples in the original ensemble and the necessary subjective parameterization of such a distribution means this is impractical for the present study.

2.d. Weighting for Uniqueness

In this section, we seek to use the relationships derived in the Section 2.b to define a weighting scheme which would effectively down-weight closely related model pairs the ensemble, which we can assess using the expectation values for near neighbor distances in the random ensembles proposed in Section 2.c. Our scheme should also provide the capability to down-weight models which exhibit low fidelity in a desirable metric.

The limiting cases of such a scheme are easy to define. We consider the models, as before, to be represented as points in a space defined by the loadings of the model in an ensemble-wide EOF analysis. In the extreme case, if the distance between two models is exactly zero then the models are considered identical and each member of the pair should be given half the weight that they would otherwise have (equivalently, a statement that adding an identical model to an existing ensemble member should not change the results).

We propose a simple functional form for model similarity which satisfies the requirements for a given model pair $[i, j]$, separated by a distance δ_{ij}^{20c} or δ_{ij}^{21c}:

$$S(\delta_{ij}^{20c}) = e^{-\left(\frac{\delta_{ij}^{20c}}{D_u}\right)^2} \quad (10)$$

$$S(\delta_{ij}^{21c}) = e^{-\left(\frac{\delta_{ij}^{21c}}{D_u}\right)^2}, \quad (11)$$

where D_u is a free parameter, a 'radius of similarity', such that model pairs separated by less than this value are considered similar. The distance is squared so that the metric tends to unity for values $<< D_u$. The smallest reasonable value for D_u would be the expected distance between two identical models exhibiting different realizations of internal model variability, given this represents a case where the model structure is identical. As D_u is increased from this value, increasingly distant pairs of models are considered similar. In the extreme case, as D_u approaches the largest inter-point distances (i.e. the largest values of δ_{ij}^{20c} or δ_{ij}^{21c}) in the ensemble, then only the models with the largest biases would exhibit a value of S of close to unity and all other members would be down-weighted.

In Section 2.c, we derived D_u empirically by considering the nearest neighbors one would expect to find by chance in a t dimensional normal distribution of equal population, variance and dimensionality as U. This is achieved in practice by considering the randomly generated distributions from the Section 2.a. We define D_u to be the 50th percentile of nearest-neighbor distances in the 10^5 randomly generated ensembles.

One can thus obtain a value for the effective repetition of model i in the ensemble:

$$R_u(i)^{20c} = 1 + \sum_{j \neq i}^{m} S(\delta_{ij}^{20c}) \quad (12)$$

$$R_u(i)^{21c} = 1 + \sum_{j \neq i}^{m} S(\delta_{ij}^{21c}), \quad (13)$$

for the the past and future cases respectively, where m is the total number of models. We then propose a uniqueness weighting for model i by taking the inverse of the number of models similar to i:

$$w_u(i)^{20c} = \left(R_u(i)^{20c}\right)^{-1} \tag{14}$$

$$w_u(i)^{21c} = \left(R_u(i)^{21c}\right)^{-1}. \tag{15}$$

for the the past and future cases respectively. If desired, a weighting scheme could also consider model quality, a model should be given increasingly less weight the further that model lies from the point representing the observations in the EOF space. In the limiting case, the model weight should tend to zero as the distance of the model to the observations tends to infinity. These attributes are satisfied by the following construction for w_q, the model quality weighting:

$$w_q(i) = e^{-\left(\frac{\delta^{20c}_{i(obs)}}{D_q}\right)^2}, \tag{16}$$

where $\delta^{20c}_{i(obs)}$ is the Euclidean distance between the EOF loading for model i and the loading of the observed climatology projected onto the same EOF basis set. This is only calculated for the historical data where observations are available. D_q is a 'radius of model quality', and is a free parameter in the weighting scheme. As $D_q \to +\infty$, then $w_q \to 1$ for all models, and the quality weighting has no distinguishing effect. As the value of D_q is reduced, models closer to the observations are increasingly up-weighted. The smallest reasonable value for D_q would be the smallest observational bias seen in the ensemble (i.e. $min(\delta_{i(obs)})$). In the extreme case as $D_q \to 0$, the majority of the weight is placed on the single best performing model.

To explore the sensitivity to this parameter, we consider two values for D_q: a 'wide' choice where D_q is equal to the mean inter-model distance in the CMIP5 ensemble and a 'narrow' choice which is half of this value. Expressing D_q in terms of the CMIP variance has the disadvantage that the variance itself can be influenced by both model quality and reproduction, but this decision is a matter of practicality. We present the values of D_q as subjective, effectively as a statement that relative skill, rather than any absolute measure, should define whether we accept or reject a model. In effect, the 'wide' case describes a situation where only the models with the largest biases in the ensemble are down-weighted, while in the narrow case a distinction is made between the 'average' and 'best' performers. It might be desirable to let internal or natural variability define D_q, but as we show in Section 3.a, this would lead to a situation where $\delta_{i(obs)}^{20c} >> D_q$ for all i, which given Equation 16, would place the majority of the weight on the model with the lowest value of $\delta_{i(obs)}^{20c}$.

2.e. Eliminating interdependent models

If the researcher's goal is simply to produce a multi-model average which is less susceptible to bias by model replication, then simply weighting each model by the appropriate value of w_u would suffice. This approach could be used directly for calculating a central estimate of combined multi-model projections.

However, some issues associated with model co-dependence cannot be solved by weighting alone. For example, the potential bias associated with regression-based predictions of unknown climate parameters can only be addressed by removing the interdependent models. This can be achieved in a pure statistical fashion (see Caldwell et al. 2014) but the interpretation of such constructions is not always intuitive.

We propose here a less formal approach which should be readily reproducible for a variety of purposes where it is desired to remove the most blatant model codependencies. Our method is a step-wise model elimination, where the models with the highest co-dependencies are removed first.

The simplest approach here would be to recursively remove a member of the closest near-neighbor pair until the remaining ensemble conforms to a plausible random distribution in the n dimensional EOF space. Since better models are replicated more, however, such an approach preferentially eliminates the models clustering closer to observations while models with large biases would be preserved. This has a significant detrimental effect on the mean performance of the remaining ensemble. Instead, we propose a strategy which considers both model performance and model independence when creating an ensemble subset.

Firstly, we introduce a bulk quantity which describes the ensemble characteristics, the 'independent ensemble quality score':

$$S_m^{20c} = \sum_i^m w_u^{20c}(i) w_q(i) \quad (17)$$

$$S_m^{21c} = \sum_i^m w_u^{21c}(i) w_q(i), \quad (18)$$

for historical and future cases, where w_u^{20c}, w_u^{21c} and w_q are described in Section 2.d as the individual model weights corresponding to model i. Using the product of the two weights is a subjective decision, and other functional forms could potentially be explored. However, as we now demonstrate, this simple combination of the uniqueness and quality weights addresses our goals to remove the influence of exactly replicated models and of very poor models.

This can be illustrated as follows for the historical simulation: If an independent model is added to the ensemble, $w_u^{20c}(i)$ equals 1 for model i, and so S_m will increase by the model quality score, $w_q(i)$. The increase is large for a high performing model, and approaches zero for a very poor model. However, if two identical models i and j are added to the ensemble together $w_u^{20c}(i)$ and $w_u^{20c}(j)$ each equal 0.5, and so S_N will still only increase by $w_q(i)$.

If we start with an N member ensemble, we eliminate a single member by considering the maximum possible ensemble quality score for each combination of $N-1$ members. The excluded model j is removed from the ensemble and the process is repeated until an appropriate stopping criterion has been reached. We can assess the effective number of models remaining at any point by considering the 'number of effective models', for both historical and future cases:

$$n_{eff}^{20c} = \sum_{i}^{m} w_u^{20c}(i) \tag{19}$$

$$n_{eff}^{21c} = \sum_{i}^{m} w_u^{21c}(i), \tag{20}$$

each representing the sum of the uniqueness weights for the remaining models in the ensemble.

The approach outlined here is quantitative but subjective, with a number of free parameters. In order to demonstrate its utility, we consider a case study of the CMIP5 ensemble, where we can objectively demonstrate that we can use the algorithm to produce a subset of CMIP5 models which provides comparable model diversity, improved mean model performance and reduced model replication in comparison to the original model sample.

3. Results

3.a. CMIP5 Ensemble Properties

The initial dataset from which we draw our conclusion is the matrix of pairwise distances between models in the CMIP-5 archive, δ^{20c} and δ^{21c} which are calculated from U^{20c} and U^{21c} matrices. This matrix is represented graphically in Figure 1 for the all-variable case using both present day climatological fields calculated from 1970 to 2000 in historical simulations, and the anomalies from those fields in the RCP8.5 simulation between 2070 and 2100. In both cases, recognizable structure relating to model genealogy is visible in the inter-model distance field.

We can compare, in a bulk sense, the distribution of distances in the matrices to that one might expect from a purely random distribution. The distributions for the CMIP5 derived matrix and the random distributions are plotted in Figures 2(a) and (b) for a number of different variable choices.

The random distributions have the same variance as the original CMIP5 distributions by design because each dimension of the random psuedo-ensembles is normally distributed with the same variance as the original CMIP5 case in each dimension of U_{20c} and U_{21c}. Because we consider a large number of pseudo-random normally distributed ensembles, we can produce best estimates and confidence intervals for the distribution of inter-model distances one would expect if the models were normally distributed in the space defined by U_{20c} and U_{21c}. If the CMIP5 distribution falls outside of this range, this implies that the models in CMIP5 are distributed in a non-normal fashion in the space.

We find there are some significant deviations in the CMIP5 distribution from what one would expect in a purely random case. Firstly, there are a number of model pairs which lie closer to each other in the EOF space than ever occurs by chance in the random samples (less than 50 percent of the expected mean inter-point distance for the random case). However, there is also an absence of models at intermediate distances (between 50 and 90 percent of the mean inter-point distance), relative to the random distributions. This indicates that the distribution of CMIP5 models in the EOF space has a rather heterogeneous, clustered distribution - with families of closely related models lying close together but with significant voids in-between model clusters. These features are especially clear in the future case, where the distances are measured in terms of (2070-2100) anomalies from the (1970-2000) climate mean state. We also show the histogram of inter-model distances in initial condition CCSM4 ensemble, demonstrating that inter-model distances due to internal model variability alone are an order of magnitude smaller than the mean inter-model distances seen in the CMIP5 archive.

The responsible model pairs can be explicitly plotted. Figure 3(a) shows model pairs which are closer together than the expected nearest-neighbor distances in the random distributions, using all variables. Many of these samples correspond to identical models from the same institution submitted at a different resolution (IPSL-CM5A-MR/LR, MPI-ESM-LR/MR for example). Other model pairs relate to changes in model configuration which do not influence the set of atmospheric diagnostics considered here (HadGEM2-AO and HadGEM2-ES for example share the same atmospheric, ocean and ice models, but the former lacks treatment of the carbon cycle which has little effect in these concentration driven simulations). Finally, there are some cases where models from two institutions share a large fraction of code-base, and this is reflected in their proximity in EOF space (HadGEM2-AO and ACCESS1-0 or FIO-ESM and BNU-ESM, for example). Several other model pairs are plotted with dotted lines. These, to a lesser degree, still occur closer together than one might expect by chance (for the models joined by a black line, one such pair would be expected by chance in a 36 member ensemble). These connections can also be related to common model components (for example, NorESM and CCSM4 share atmosphere and land surface, MPI-ESM and CMCC-CSM5 share atmospheric code). We also include the observational point in the same analysis in Figure 3(a), which shows that none of the models in the CMIP5 archive are considered closer to the observations than would be expected by chance. In the later part of the study, where we prune similar models from the archive, this give us some confidence that similar models are not being removed because they are all converging on the 'true' climate. We can repeat the analysis for future changes in the same variables (Figure 3(b)), which show a similar close relationships to present day case. Using specific fields produces similar (but non-identical) relationships (Figure 3(c-e)). The all-variables case shows that all close relationships would be expected from a genealogical perspective. However, when one uses single variables (PR especially), there are some unexpected results (e.g. MIROC and CAM5 are considered close). We attribute this to the difficulty of representing inter-model precipitation variability in a low dimensional basis set (although models from different centers may in some cases share parameterizations).

3.b. Stepwise model elimination

There are various arguments to support the hypothesis that the CMIP5 ensemble is biased by the inclusion of common components, some of which are featured more frequently than others. One can make this argument from a consideration of the models themselves (see Introduction and Table 2), or by examining the spatial distribution of models in orthogonal dimensions derived from model output. We have proposed a method of model removal which maximizes a metric reflecting both model diversity and fidelity. The iterative model elimination process is illustrated for the CMIP5 ensemble in Figure 4.

The plot shows the consecutive removal of models from the set of 36 considered in this study until a single model remains. The process is demonstrated by eliminating interdependent models as judged by the simulation of present day climatology. The model quality weights w_q are obtained using the mean state climatology from the models as compared to the observations. Model uniqueness is calculated as in Section 2.e after each iteration.

We demonstrate the sequence of model removal in Figure 4 (for present day similarities, all variables and a 'wide' quality radius). The figures show the order in which models are removed from the archive to achieve the maximum independent ensemble quality. If the removed model is closer than D_u (a function of the number of models remaining) to any other remaining model, then that model is shown to merge with its nearest neighbor. However, if the model is further than D_u from any other model, the model branch is shown as terminating in the diagram.

We have not yet fully discussed an appropriate point to stop trimming models. This question is ultimately subjective, and the conclusion is somewhat dependent on the specific needs of the researcher. However, Figure 5 shows some changing characteristics of the remaining ensemble as the ensemble size is decreased, and these can be used to recommend ensemble subsets for different scenarios. In essence, a first phase of eliminating models just removes redundant data, a second improves the characteristics of the ensemble by removing poor models and partly redundant ones. Going beyond that potentially worsens the ensemble mean bias representation.

Figure 5(a) shows how n_{eff} varies as models are removed from the archive as described in Section 2.e. The actual number is dependent on the choice of D_u, the radius of similarity.

Two choices of D_u are illustrated, using either the 50th percentile of nearest-neighbor distances in the set of 10^5 random ensembles (as was used in Section 2.d) or, for comparison, the 90th percentile. Using all the models in the archive, n_{eff} is 15.5 using the larger value for D_u, or 22.5 using the smaller value (using present day climatology metrics of similarity). The removal of the first 10 models has little effect on n_{eff} (especially using the larger value of D_u). The removal of the remaining models results in a monotonic decrease in n_{eff}.

As was indicated by Figure 5(a), most of the early model eliminations have little effect on n_{eff}. Figure 4 shows that many of the initial removals represent models (CCSM4 to CESM1-BGC, HadGEM2-ES to HadGEM2-AO, GFDL-ESM2M to GFDL-ESM2G) which are largely structurally identical, at least in terms of their long term atmospheric climatology - differing only in the presence of an active carbon cycle which would not influence the diagnostics used in this study. It is thus largely random which member of the pair is eliminated. In this regime, there is a strong inverse relationship between model quality weights (w_q) and uniqueness weights (w_u), as shown in Figure 6(a).

The second broad class of eliminations is models with strong connections, often from the same institutions but with some differing components. In these cases, the model with the higher value quality weighting (w_q) is generally preserved (for instance, GISS-E2-H and GISS-E2-R which differ in their ocean components). In this regime, the inverse relationship between the model quality weight and uniqueness weights is weaker (Figure 6(b)), as the clear duplicates have already been removed. Note that the uniqueness weights now refer to uniqueness within the remaining subset, and not within the full CMIP5 archive.

The final stages of removal (approximately the final 20 models) do result in a reduction in the number of effective models, illustrated by the termination of the model path. As shown in Figure 5(b), in this regime - the distribution of inter-model distances are now consistent with what one might expect from a purely random sample. Each family of closely related models is now represented, to a large extent, by its own 'champion'. Figure 6(c) shows that when only 10 models remain, the relationship between w_u and w_q is rather weak, with all remaining models having comparable uniqueness weights.

Our value judgment for an appropriate stopping criterion is thus dependent on the application. If one wishes to only remove near-identical models, one should stop trimming when the number of effective models n_{eff} begins to significantly decrease. However, if one wishes to produce the best-performing ensemble mean simulation of the mean state, it is more logical to also remove the worst performing models such that the RMSE error of the sub-ensemble mean is minimized.

3.c. Sensitivity to initial choices

The algorithm as described in Section 2.e requires several assumptions and we explore the sensitivity of the results to those choices in this section. Figure 7 shows the models which are retained in the analysis with a range of different initial variable and parameter choices. In each case, the analysis is repeated and there is a stepwise removal of models based on maximising the ensemble quality score. On each line of the plot, we show which models remain when the smallest inter-point distance in the remaining archive is first greater than 50% (unfilled symbols) or 10% (filled symbols) of purely random distributions of the same population, variance and dimensionality (regions marked by mid grey and dark dray shading in Figure 4). Thus, we can explore the sensitivity of the retained models to our initial assumptions.

Firstly, there is the choice of which variables are used to derive the inter-model distance matrix. To address this, we repeat the analysis with a variety of individual fields, as well as the multivariate example discussed in the previous section. The analysis is repeated for zonal mean temperature and humidity (TQ), gridded precipitation (PR), gridded Top of Atmosphere shortwave and longwave fluxes (TOA), Gridded surface air temperature (TAS) and all variables combined (ALL). Secondly, we explore the 'radius of model quality' D_q introduced in Equation 16. The analysis is repeated for two values, a 'wide' value where D_q is equal to the mean inter-model distance in the CMIP5 ensemble and a 'narrow' choice which is half of this value. The latter 'narrow' case effectively increases the role of the model quality metric, such that models with a low quality score are removed earlier in the algorithm, unlike in the 'wide' case, where highly interdependent models are removed first. Finally, we construct the model uniqueness weightings w_u using the inter-model distances derived from the 30 year mean 1970-2000 present day data in the 'present' case, but use the anomaly between 2070-2100 and 1970-2000 for the 'future' case.

We find that variable choice has little impact on the final choice of model subsets. Although in some cases, the choice of model from a given institution can change, the overall number of models retained is similar for each of the variable choices. The use of the 'narrow' radius of model quality, however, significantly decreases the number of retained models with respect to the 'wide' value. This can be explained by considering that the narrow setting increases the ratio of the model quality weighting for models lying close to the observations, and those far away. In the 'narrow' regime, the ensemble quality score is best maximised by removing the poorly performing models earlier in the analysis, and thus after the interdependent remaining models have been removed, the number of remaining unique models is smaller than in the 'wide' case.

3.c.1) EOF truncation choices

Some subjective decisions are required in the interpretation and subsequent usage of the PCA conducted in Section 2.a, and we discuss these at greater length here. In previous studies like Masson and Knutti (2011), the inter-model distances were calculated without the PCA stage, simply calculating distances in the space defined by the anomaly matrices, ΔX^{20c} and ΔX^{21c}. For the purposes of this study, and its companion studies (Sanderson et al. submitted), it is neccessary to decrease the dimensionality (and co-dependence) of the data in order to establish prior expectations of near-neighbor distances.

In this study, as in Sanderson et al. (submitted), the inter-model distances are calculated with the truncated set of 9 modes. The resulting inter-model distance matrix calculated with

U^{20c} truncated to 9 modes has a 0.93 correlation with the matrix one would calculate using the full-field matrix ΔX^{20c}, but using the orthogonal basis set allows us to form random matrices with which to compare the results (Figure 2).

For smaller values of t, only the leading patterns of model difference are retained, which results in large inter-model distances between different model families (e.g. CESM and GFDL models) and very small distances between models in the same family (e.g. CESM-CAM5 and CESM-CAM4). With such few degrees of freedom, very small intermodel distances cannot be ruled out by chance in the random ensembles, and so no models can be excluded from the ensemble (see Figure 8 for truncation values of 3 or less. The analysis produces very similar results, and the minimum number of retained models, for values of t between 8 and 12 (see Figure 8), with relatively little sensitivity to variable choice (not shown). For values of t of 15 or greater, the higher order modes increasingly represent subtle and often noisy differences between models in the archive, which inflates the distance between the near-neighbors in the ensemble. Hence, once again we see fewer models ruled out.

To test the sensitivity of the inter-model distance matrix to variable choice, we also repeat the EOF analysis with a number of different subsets of diagnostic variables. The resulting correlation depends significantly on which exact variable is retained. The inter-model distances calculated using gridded surface temperature only ('TAS') are highly correlated with the multi-variate case (R=0.95, untruncated). Top of atmosphere radiative fluxes (RAD, R=0.85 untruncated), Total Precipitation (PR, R=0.66 untruncated), and zonally averaged vertical temperature and humidity (QT, R=0.42 untruncated) are increasingly poorly correlated with the full field multi-variate case. This implies that some fields, such as surface temperature have sufficient information to render a multi-variate approach unnecessary.

With a truncation length of 9, which we used for the bulk of this study, the resulting distance matrix remains highly correlated to the full field distance matrix, but the influence of covariant fields and models is reduced (see Caldwell et al. 2014 for an extensive discussion of these issues).

3.d. Ensemble Mean Performance

The results of Section 3.b suggest that eliminating the strongest interdependent models to leave a plausibly random distribution would leave between 10 and 25 of the 36 CMIP5 models considered here (depending on variable and parameter choices). Trimming the ensemble to its more independent subset does not worsen the fidelity of the climatological mean result, and removing the poorer performing outliers (models with large biases) can actually improve it, as we show in this section.

We can first examine how the multi-model mean of present day climatology compares against observations. Figure 5(c) considers the Root Mean Square Errors (RMSE) of various weighted and unweighted multi-model means calculated using the same multi-variate climate state vectors described in Section 2.a and the observations listed in Table 1. We illustrate this using the 'ALL' variable case, with the 'wide' radius of model quality and present day derived inter-model distances. We also compare with the average RMSE seen when a completely random sample (without replacement) of the same size is taken, as compared to the detailed technique outlined in Section 3.b.

If one considers only the far left of the plot, where all 36 models are retained, weighting the models by uniqueness actually increases the RMSE. This is largely to be expected - as we have seen in Figure 6(a) that the best performing models have the lowest uniqueness weights. It also suggests that a mean of the CMIP5 ensemble is already weakly weighted towards the better performing models. If we explicitly weight the model mean towards models which lie closer to the observations in the EOF space, the RMSE can be reduced significantly.

As the first 10 (highly interdependent) models are removed from the archive, the simple mean RMSE increases slightly while the random draw RMSE remains constant, likely because the high-performing models have less representation when the duplicates have been pruned. The uniqueness weighted mean also becomes more similar to the simple mean case (u_w is now more consistent across the ensemble). Between 28 and 12 models remaining, the simple RMSE decreases significantly and when 20 models remain, the subset outperforms the

RMSE of the random sample. The lowest RMSE values occur with between 12 and 5 models remaining. Removing any further models increases the RMSE of the simple multi-model mean. With 5 or fewer models remaining, all models have a high value of both w_u and w_q, so weighting by uniqueness or quality has little effect. In all cases, any further removal of models (below 5) significantly increases the RMSE, a fact which is likely attributable to the Cauchy-Schwartz inequality (Annan and Hargreaves 2011).

4. Discussion and Conclusions

The present study considers how one might remove potential biases which might arise from shared components in the CMIP5 archive of climate models, and its predecessors. We also propose some simple diagnostics which might be used to identify interdependent models using model diagnostic output, and a possible strategy to choose a model subset to maintain model diversity without replication and to incorporate model quality information into this decision.

This study represents a proof of concept; the choice of diagnostics used in this study are of course arbitrary, to some degree, though the results of which models are interdependent do seem to be relatively resilient to changes in variable and time period (see Figure 3, Pennell and Reichler 2011 and Knutti et al. 2013). However, we do assume that a model's mean state climatology can be used to assess both its skill and independence. Clearly, if our final goal is to assess the plausibility of a model's future simulations then the mean state simulation is not a perfect assessment of model skill, although it could be argued that it is a neccessary condition and as such a weighting strategy based on present day climatology can be justified in the absence of any additional information.

Certainly, which model exhibits the highest quality score is very much dependent on the specific metrics in which the researcher might be interested (Santer et al. 2009), and it is far beyond the scope of this study to conduct an exhaustive comparison of possible model metrics. In this study, we have focussed primarily on diagnostic output from the atmospheric model, and our results are thus liable to be most sensitive to common component in that model. As such, the results of this study should be interpreted as illustrative of a potential method for reducing the effects model interdependency, and not as a prescriptive list of models which should be used for future studies. Most studies based on CMIP5 could easily use such a framework, but the value judgements of future researchers should be embedded into the choice of metric used to assess model similarity and quality.

We assess the likelihood of near-neighbor models occurring by chance using a large number of random distributions of the same dimensionality as the truncated orthogonal set of EOF loadings we derive from the original ensemble. The random sample is not a proxy for the space which might be attainable by the real climate, rather it is a proxy for the distribution of models represented in an orthogonal basis set defined by multi-model variability. As such, we are making the assumption that if there are physical relationships between variables in the model output data (say between surface temperature and outgoing longwave radiation), then any correlation between these would be represented as a single mode in the EOF analysis. However, if there exists a strong nonlinear relationship between two variables in the CMIP5 archive then this relationship could not be represented in a single EOF mode, and might be represented in two or more modes. In this case, then the distribution of models in the space could be more complex than a simple Gaussian. One could imagine designing a random sample which fitted a high-dimensional distribution to the CMIP5 ensemble to account for such nonlinearities, but the increase in complexity, the lack of samples in the original ensemble and the neccessary parameterization of such a distribution means this is impractical.

We also assume, by drawing random samples using the variance defined by the original ensemble, that none of the CMIP5 members can be ruled out *a priori*. One could imagine a situation where an arbitrarily poor model was included in the ensemble which would increase the variance represented in each mode such that any realistic models would look self similar and would be down weighted by the uniqueness weighting. Therefore, the method only makes sense if there is some level of base confidence that none of the models in the archive are completely unrepresentative of the true system. But, we would argue that this is true of any analysis which uses the CMIP5 archive and that even a simple mutli-model mean is subject to a sanity check of the participating models.

Caveats aside, this study illustrates some interesting characteristics of the CMIP5 archive and potential issues which might arise from treating this archive as a random sample of possible climate models. There is extensive replication of model code in the archive, primarily within institutions but also in some cases between institutions (see Table 2). This should come as little surprise, a quick examination of AOGCM makeup in the CMIP5 models indicates that some individual components are used by over 25 percent of the archive. But, we show in this study (like in Masson and Knutti 2011 and others) that many of those similarities can be identified also through a simple analysis of model output. A more detailed discussion of shared model components is given in the supplementary material of Knutti et al. (2013).

Similarities in diagnostic output are not always predictable from a consideration of model construction alone. One can find examples of cases with significant changes in code-base, but with minor changes in diagnostic similarity. For example, CCSM4 and CESM1-CAM5 have significantly different aerosol schemes, dynamics, cloud microphysics and yet our results show the two models as very strongly related when considering the distribution of inter-model distances. This indicates that tuning strategies and non-atmospheric components may play a significant role in diagnostic model similarity, even when primarily atmospheric output is used to assess inter-model distance. This implies that although the diagnostic output is a useful indicator of model similarities, those similarities may not be a function of shared code alone. The climateprediction.*net* (Stainforth et al. 2005) and QUMP (Murphy et al. 2007) experiments, for example show that considerable diversity in model behavior is achievable through parameter perturbation alone with an identical codebase.

There are several possible additional factors which might influence diagnostic similarity. Firstly, the tendency for various generations of models from a single institution to exhibit strong similarities in spite of extensive model component changes (see Figure 2 in Sanderson and Knutti 2012 with reference to CESM, GFDL or Hadley Centre models) indicates that some elements of model calibration tend to cluster models from a given modeling center. The reasons for this clustering have multiple possible candidates which could lie in institutional policy or regional focus (institutions might be more concerned with their model's performance in the region's climate). Standard metrics used to judge model performance during the model development process or preferred observational datasets may also vary from institution to institution. Secondly, models rarely change all components at the same time, so we would posit that evaluating when a model is 'new' is a subjective matter. Finally, the CMIP5 protocol allows for some flexibility in the way that models implement external forcings - so different groups, even with identical models, can choose to represent the historical and future boundary conditions in different ways to produce differences in the simulated climate. Knutti et al. (2013) see similar relationships in control simulations, but one cannot exclude the possibility that the control simulations themselves might also include common assumptions on boundary conditions.

In summary, we confirm earlier arguments that models are not independent, some are essentially duplicates, and the effective number of independent models based on this method is less than half of the actual number of models, consistent with earlier studies (Jun et al. 2008, Annan and Hargreaves 2011, Sanderson and Knutti 2012). Some models are closer to observations than others (Gleckler et al. 2008, Knutti and Sedláček 2013). We believe that our method, and results do not strongly hinge on the way in which one interprets the ensemble as 'truth centered' (Knutti 2010), 'indistinguishable from truth' (Annan and Hargreaves 2011, Rougier et al. 2013) or neither (Sanderson and Knutti 2012, Bishop and Abramowitz 2013). One could imagine a hypothetical ensemble following any of these frameworks, and by duplicating some of its members, bias would be introduced in the ensemble distribution.

By evaluating our ensemble subset performance in terms of ensemble mean performance, we do not necessarily advocate a truth centered ensemble, as the ensemble mean would also be the best estimate of future change in the indistinguishable case.

There are of course different ways to account for model performance and interdependence. In the companion paper (Sanderson et al. submitted), we proposed a method to produce probabilistic estimates that are largely insensitive to model duplicates and can consider model performance. However, when high dimensional data and/or spatially and temporally consistent fields are required (e.g., for impact models), a fully probabilistic method becomes unwieldy and might even hinder the development of tractable impact analyses (Dessai and Hulme 2004). Bishop and Abramowitz (2013) also proposes an alternative technique where models in the archive are subject to a linear transformation, where the weighted mean of transformed models is calculated to be optimally close to an observed climate. This transformation and weighting can then be extrapolated for future projections. This method has the advantage that the resulting transformed models have independent errors, and weight future projections by climatological skill. However, the transformed models are not, themselves physically self-consistent and there is a potential for simulations to be over-fitted to historical data in a manner which could potentially result in overconfident future projections. In comparison, the method we present here preserves a subset of self-consistent physical models (for both present day and future projections), and although they might not be independent in the strict sense of orthogonality, this subset can be simply used for almost any application or analysis.

We thus propose that there is significant utility in spanning the potential uncertainty in future climate by representing spread with an appropriate subset of models. This study introduces weights which assess model uniqueness and model climatology fidelity. We find that the two were inversely related such that the models with the best simulations of the present day climate were also least unique. A part of this is possibly due to the fact that models have been calibrated by the observations, and will thus appear to cluster around those observations (and each other). But, a closer examination reveals that a large fraction of the high-scoring models' lack of uniqueness can be explained by other models which have duplicated some, or all of their code. When these duplicates are removed, this strong inverse relationship is weakened (but not entirely eliminated).

This property of the ensemble is clearly to some extent contingent on the choice of metrics used, but it does raise a potentially interesting property of the ensemble; that the best performing models might also be the most promiscuous. This situation implies that the ensemble as a whole is already strongly weighted towards the better performing models. We show that if the models are weighted to reward their uniqueness, then the RMSE of the ensemble mean is increased. Thus, through a mechanism of quasi natural selection, the climate community has created an ensemble of models which has already up-weighted its climatologically best performing members. In other words, relying on model democracy is to some degree upweighting skilled model structures without deliberately thinking about it or discussing it, by the mechanism of duplication of well-proven code.

This could be seen as an argument in support of keeping the entire ensemble when performing an analysis, and at least some justification that the multi-model mean result is a defensible best estimate. But, it is at best an accidental property that is not guaranteed to remain in future ensembles, and may not at all be visible for more specific questions or metrics. Whether a model is extensively duplicated is not a pure function of its quality or fidelity. A sub-model with open source code and few restrictions on its use is more likely to be utilized by another group than another model with a closed-source policy. However, a model which is jointly used by a large number of groups also has a large development pool invested in improving that model. Duplication within institutions depends also on funding and the available computing resources. One could make the argument that the CMIP5 ensemble distribution and the social and intellectual landscape of the climate community are surely related, but certainly not in any simple fashion.

A question also remains of whether the original CMIP5 ensemble is sufficient to assess systematic uncertainty in future climate change. This question could easily form a study in itself, but our results are somewhat informative in this matter. Firstly, the number of truly independent models in the archive is significantly less than the number of submitted models, when gauged by model output. Hence, adding another model to the existing archive has most value if the developers introduce novel components and assumptions. It is true that exploring different configurations of existing components through sub-model exchange or parameter perturbation can certainly modify model behavior, and we would argue that such experiments should continue in order to fully explore the inherent uncertainties in the existing model set.

However, this uncertainty is conditional on the number of independent models available to us, and establishing whether the current set is sufficient is a question which might not be a useful, because there is not a convenient space in which systematic model assumptions can be defined. For example, the current CMIP5 ensemble might have n fundamentally different convection schemes, each with its own advantages and biases, but nobody would argue that this constituted a "full set". Where there is approximation and parameterization, there are potentially limitless ways to address this. And because nobody can know the behavior of the $n + 1^{th}$ model, the question of ensemble adequacy cannot be answered in a strict sense. Within the ensemble we have, we can tractably experiment with subsetting to assess how many models are required to have confidence in the distribution of future climate change formed by the full set, but we can never know if the $n+1^{th}$ model will adopt different assumptions or resolve a new process to place its projection outside of the existing distribution.

We argue that a joint consideration of model similarity and quality metrics allows the researcher to make use of a more quantitatively defensible sample of simulations available in the CMIP archives, either through weighting or by model elimination (in itself, an extreme form of weighting) to produce a best estimate of combined model projections. Our approach for achieving this can be controlled with a small number of subjective but clearly defined parameters, which can potentially mitigate some of the arbitrary sampling issues which arise from relying on model democracy, and can be tailored to specific questions by choosing appropriate metrics and datasets.

It should be noted in this discussion that the CMIP5 archive is not a full representation of the uncertainty space for GCM projections. Rather, it is a collection of intended 'best possible models', the final iterations of their respective tuning processes as model developers calibrate their parameterization choices to best represent the observed climate properties which they find most important, although there may be other acceptable configurations (Mauritsen et al. 2012). Clearly, these choices and targets will vary from model to model, but the fact that there are implicitly a near-infinite number of rejected parameter configurations for each model must be remembered when trying to interpret the significance of the spread of simulations in the archive. In a practical sense, we ignore these rejected configurations because we do not have access to them. In addition, there is some evidence to suggest that the model diversity one can attain by structural changes significantly exceeds that of parameter changes in currently available Perturbed Parameter ensembles (Yokohata et al. 2013). Nevertheless, it should be remembered that both the CMIP5 ensemble (and by definition our subsets of that ensemble) is already a subset of all possible model configurations which have been chosen by model developers.

There are some cases where we would argue it is essential to eliminate interdependent models, such as when a correlation found in the multi-model ensemble is used as a constraint on a climate parameter (such as for climate sensitivity in Fasullo and Trenberth 2012, or for high latitude surface albedo feedbacks in Hall and Qu 2006). The presence of closely related, or even identical models in the archive would tend to artificially inflate the significance of any correlation simply because identical models would exhibit similar values for both the predictor and for the unknown quantity (Caldwell et al. 2014). Removing the obvious interdependent models as shown in this study would certainly be better than assessing a correlation based on the entire archive, but a method for achieving this in a strict statistical sense is presented in Caldwell et al. (2014).

There is a danger that as models improve, the better models have the potential to converge on the 'true' climate state, which might lead to their elimination if interdependent models are removed. We show in Figure 3 that this is unlikely to be the case for CMIP5, given none of the models lie close enough to the observations to be influenced by the uniqueness weighting. However, one could imagine if a small group of models make a real advance which removes a long-standing systematic bias (for example, as some models begin to explicitly resolve convection), then it would be neccessary to accept a higher level of similarity among the better performing models (i.e. the uniqueness weighting u_w could no longer be independent of the skill weighting u_s).

Proposing a subset of models to consider for a less biased analysis could be seen as overly prescriptive, but our aim is not to focus on the exact set of models which should be used for future studies, rather to establish a framework in which researchers could make their selection based upon metrics which are most relevant to their question. We would argue that although the collection of models which arise from the 'ensemble of opportunity' is often seen as sacrosanct, the democratic policy of one-model, one-vote is no longer a logical one in the increasingly complex family tree of models available to the researcher. A subset of 10-20 models that are reasonably independent and perform well for the criteria that are judged to be relevant is very likely to be more skillful than the full ensemble. Giving equal weight to all models which have completed a simulation of interest is, albeit implicitly, adopting a weighting scheme which rewards model components which are highly replicated. This weighting scheme might fortuitously have the property of rewarding the most skilled components but, we would argue, this property should be demonstrated and the decision how to incorporate it should be made consciously.

5. Acknowledgments

We acknowledge the World Climate Research Programme's Working Group on Coupled Modeling, which is responsible for CMIP, and we thank the climate modeling groups for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Inter-comparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. Portions of this study were supported by the Office of Science (BER), US Department of Energy, Cooperative Agreement No DE-FC02-97ER62402. The National Center for Atmospheric Research is sponsored by the National Science Foundation. We would also like to thank our anonymous reviewers for their extensive and insightful comments.

REFERENCES

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.

Annan, J. and J. Hargreaves, 2011: Understanding the CMIP3 Multimodel Ensemble. *Journal of Climate*, **24 (16)**, 4529–4538.

Aumann, H. H., et al., 2003: Airs/amsu/hsb on the aqua mission: Design, science objectives, data products, and processing systems. *IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING*, **41 (2)**, 253.

Bishop, C. H. and G. Abramowitz, 2013: Climate model dependence and the replicate earth paradigm. *Climate Dynamics*, **41 (3-4)**, 885–900.

Brohan, P., J. Kennedy, I. Harris, S. Tett, and P. Jones, 2006: Uncertainty estimates in regional and global observed temperature changes: a new dataset from 1850. *J. Geophys. Res.*, **111 (D12)**, D12 106.

Caldwell, P. M., C. S. Bretherton, M. D. Zelinka, S. A. Klein, B. D. Santer, and B. M. Sanderson, 2014: Statistical significance of climate sensitivity predictors obtained by data mining. *Geophysical Research Letters*, **41 (5)**, 1803–1808.

Dessai, S. and M. Hulme, 2004: Does climate adaptation policy need probabilities? *Climate policy*, **4 (2)**, 107–128.

Evans, J. P., F. Ji, G. Abramowitz, and M. Ekström, 2013: Optimally choosing small ensemble members to produce robust climate simulations. *Environmental Research Letters*, **8 (4)**, 044 050.

Fasullo, J. T. and K. E. Trenberth, 2012: A less cloudy future: The role of subtropical subsidence in climate sensitivity. *Science*, **338 (6108)**, 792–794.

Gleckler, P., K. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res*, **113**, D06 104.

Hall, A. and X. Qu, 2006: Using the current seasonal cycle to constrain snow albedo feedback in future climate change. *Geophys. Res. Lett.*, **33**, L03 502, doi:10.1029/2005GL025127.

Jun, M., R. Knutti, and D. W. Nychka, 2008: Local eigenvalue analysis of CMIP3 climate model errors. *Tellus A*, **60 (5)**, 992–1000.

Knutti, R., 2010: The end of model democracy? *Climatic Change*, **102 (3-4)**, 395–404.

Knutti, R., D. Masson, and A. Gettelman, 2013: Climate model genealogy: Generation CMIP5 and how we got there. *Geophysical Research Letters*, **40 (6)**, 1194–1199.

Knutti, R. and J. Sedáček, 2013: Robustness and uncertainties in the new CMIP5 climate model projections. *Nature Climate Change*, **3 (4)**, 369–373.

Masson, D. and R. Knutti, 2011: Climate model genealogy. *Geophys. Res. Lett*, **38 (8)**, L08 703.

Masson, D. and R. Knutti, 2013: Predictor screening, calibration, and observational constraints in climate model ensembles: An illustration using climate sensitivity. *Journal of Climate*, **26 (3)**, 887–898.

Mauritsen, T., et al., 2012: Tuning the climate of a global model. *Journal of Advances in Modeling Earth Systems*, **4 (3)**, M00A01.

Murphy, J., B. Booth, M. Collins, G. Harris, D. Sexton, and M. Webb, 2007: A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. *Philosophical Transactions A*, **365 (1857)**, 1993.

NASA, 2011: CERES EBAF Data Sets. *Available online*, URL `http://eosweb.larc.nasa.gov/PRODOCS/ceres/level4_ebaf_table.html`.

Pennell, C. and T. Reichler, 2011: On the effective number of climate models. *Journal of Climate*, **24 (9)**, 2358–2367.

Qu, X. and A. Hall, 2013: On the persistent spread in snow-albedo feedback. *Climate Dynamics*, 1–13.

Rougier, J., M. Goldstein, and L. House, 2013: Second-order exchangeability analysis for multimodel ensembles. *Journal of the American Statistical Association*, **108 (503)**, 852–863.

Sanderson, B., R. Knutti, and P. Caldwell, submitted: Addressing interdependency in a multi-model ensemble by interpolation of model properties. *Journal of Climate*.

Sanderson, B. M. and R. Knutti, 2012: On the interpretation of constrained climate model ensembles. *Geophysical Research Letters*, **39 (16)**, L16 708, doi:DOI:10.1029/2012GL052665.

Santer, B., et al., 2009: Incorporating model quality information in climate change detection and attribution studies. *Proceedings of the National Academy of Sciences*, **106 (35)**, 14 778–14 783.

Stainforth, D. A., et al., 2005: Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, **433 (7024)**, 403–406.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of cmip5 and the experiment design. *Bulletin of the American Meteorological Society*, **93 (4)**, 485–498.

Tebaldi, C., J. M. Arblaster, and R. Knutti, 2011: Mapping model agreement on future cliamte projections. *Geophys. Res. Lett.*, **38**, L23 701, doi:10.1029/2011GL049863.

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)**, 2053–2075.

Yokohata, T., et al., 2013: Reliability and importance of structural diversity of climate model ensembles. *Climate Dynamics*, **41 (9-10)**, 2745–2763.

List of Tables

Table 1. Observational Datasets used as ‘observations throughout. * ”The data used in this effort were acquired as part of the activities of NASA’s Science Mission Directorate, and are archived and distributed by the Goddard Earth Sciences (GES) Data and Information Services Center (DISC).”

Field	Source	Reference	Years	Global normalization
TS	HadCRUT3	Brohan et al. (2006)	1970-2000	2.09 K
PR	GPCP	Adler et al. (2003)	1979-2001	30.1 Wm^{-2}
RSUT	CERES-EBAF	NASA (2011)	2000-2005	25.8 Wm^{-2}
RLUT	CERES-EBAF	NASA (2011)	2000-2005	3.32 mm/day
T	AIRS*	Aumann et al. (2003)	2002-2010	0.28 K
RH	AIRS*	Aumann et al. (2003)	2002-2010	12.12 %

TABLE 2. Submodel components for the 38 CMIP5 models considered in this study.

Model	Atmosphere	Land	Ocean	Ice	Source
NorESM1-ME	CAM4	CLM4	MICOM-HAMOCC	CICE	https://verc.enes.org/ISENES2/models/earthsystem-models/ncc/noresm
NorESM1-M	CAM4	CLM4	MICOM-HAMOCC	CICE	https://verc.enes.org/ISENES2/models/earthsystem-models/ncc/noresm
MRI-CGCM3	MRI-AGCM3	HAL	MRI.COM3		http://www.mri-jma.go.jp/Publish/Technical/DATA/VOL_64/index_en.html
MPI-ESM-MR	ECHAM6	JSBACH	MPIOM		http://www.mpimet.mpg.de/en/science/models/mpi-esm.html
MPI-ESM-LR	ECHAM6	JSBACH	MPIOM		https://www.enes.org/models/system-models/mpi-m/mpi-esm
MIROC5	FRCGC-AGCM	MATSIRO	CCSR-COCO	Bitz/Lipscomb	http://journals.ametsoc.org/doi/full/10.1175/2010JCLI3679.1
MIROC-ESM-CHEM	FRCGC-AGCM	MATSIRO	CCSR-COCO	Bitz/Lipscomb	http://www.wcrp-climate.org/wgcm/WGCM15/presentations/21Oct/KIMOTO_Japan.pdf
MIROC-ESM	FRCGC-AGCM	MATSIRO	CCSR-COCO	Bitz/Lipscomb	http://www.wcrp-climate.org/wgcm/WGCM15/presentations/21Oct/KIMOTO_Japan.pdf
IPSL-CM5B-LR	LMDZ (CM4)	ORCHIDEE	NEMO-OPA	NEMO-LIM	http://icmc.ipsl.fr/index.php/icmc-models/icmc-ipsl-cm5
IPSL-CM5A-MR	LMDZ	ORCHIDEE	NEMO-OPA	NEMO-LIM	http://icmc.ipsl.fr/index.php/icmc-models/icmc-ipsl-cm5
IPSL-CM5A-LR	LMDZ	ORCHIDEE	NEMO-OPA	NEMO-LIM	http://icmc.ipsl.fr/index.php/icmc-models/icmc-ipsl-cm5
INMCM4	INMCM	INMCM	INMCM	INMCM	http://link.springer.com/article/10.1134%2FS000143381004002X
IAP-FGOALS-g2	GAMIL 2.0	CLM3	LICOM2	CICE4_LASG	http://link.springer.com/article/10.1007%2Fs00376-012-2140-6
HadGEM2-ES	HadGAM2 (N96L38)	TRIFFID	HadGOM2		http://cms.ncas.ac.uk/wiki/UM/Configurations/HadGEM2
HadGEM2-CC	HadGAM2(N96L60)	TRIFFID	HadGOM2		http://cms.ncas.ac.uk/wiki/UM/Configurations/HadGEM2
HadGEM2-AO	HadGAM2 (N96L38)	MOSES2	HadGOM2		http://cms.ncas.ac.uk/wiki/UM/Configurations/HadGEM2
GISS-E2-R	GISS	GISS	Russell	Russell	http://data.giss.nasa.gov/modelE/ar5/
GISS-E2-H	GISS	GISS	HYCOM	HYCOM	http://data.giss.nasa.gov/modelE/ar5/
GFDL-ESM2M	GFDL-AM2.1	LM3	MOM4.1	SIS	http://cms.ncas.ac.uk/wiki/UM/Configurations/HadGEM2
GFDL-ESM2G	GFDL-AM2.1	LM3	GOLD	SIS	http://www.gfdl.noaa.gov/earth-system-model
GFDL-CM3	GFDL-AM3	LM3	MOM4.1	SIS	http://www.gfdl.noaa.gov/earth-system-model
FIO-ESM	CAM3.5	CLM3	POP2	CICE4	http://www.wcrp-climate.org/wgcm/WGCM15/presentations/21Oct/WANG_WGCM.pdf
CanESM2	AGCM4	CLASS	NCAR		http://journals.ametsoc.org/doi/pdf/10.1175/JCLI-D-11-00715.1
CSIRO-Mk3-6-0	Gordon	CABLE	MOM2.2	SIS	http://www.bom.gov.au/amoj/docs/2013/jeffrey_hres.pdf
CNRM-CM5	ARPEGE-Climate	ISBA	NEMO-OPA	GELATO	http://www.cnrm-game.fr/spip.php?article126&lang=en
CMCC-CMS	ECHAM5	SILVA	OPA8.2	LIM	http://www.wcrp-climate.org/wgcm/WGCM16/Bellucci_CMCC.pdf
CMCC-CM	ECHAM5	SILVA	OPA8.2	LIM	http://www.cmcc.it/models/cmcc-cm
CMCC-CESM	ECHAM5	SILVA	OPA8.2	LIM	http://www.cmcc.it/models/cmcc-cm
CESM1-CAM5	CAM5	CLM4	POP2	CICE4	https://www2.cesm.ucar.edu/models
CESM1-BGC	CAM4	CLM4	POP2	CICE4	https://www2.cesm.ucar.edu/models
CCSM4	CAM4	CLM4	POP2	CICE4	https://www2.cesm.ucar.edu/models
BNU-ESM	CAM3.5	CLM/BNU	MOM4.1	CICE4.1	http://www.wcrp-climate.org/wgcm/WGCM15/presentations/21Oct/WANG_WGCM.pdf
BCC-CSM1-1-M	BCC_AGCM 2.1	CLM3	MOM4	SIS	http://link.springer.com/article/10.1007%2Fs13351-014-3041-7
BCC-CSM1-1	BCC_AGCM 2.1	CLM3	MOM4	GFDL SIS	http://link.springer.com/article/10.1007%2Fs13351-014-3041-7
ACCESS1-3	UKMO GA1.0	CABLE v1.8	MOM4.1	CICE4.1	https://wiki.csiro.au/display/ACCESS/Home
ACCESS1-0	HadGEM2 r1.1	MOSES	MOM4.1	CICE4.1	http://www.cawcr.gov.au/publications/technicalreports/CTR_059.pdf

List of Figures

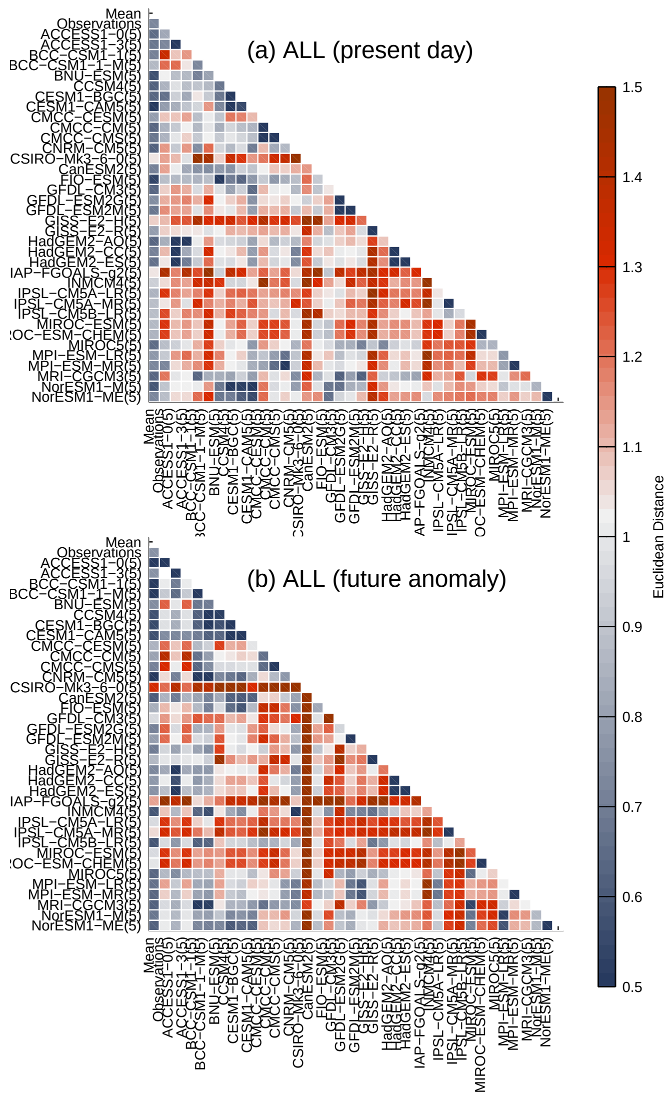

FIG. 1. A graphical representation of the inter-model distance matrix for CMIP5 calculated for ALL variables using (a) 1970-2000 monthly mean climatological fields as defined in Table 1 and (b) changes in the aforementioned fields between (1970-2000) and (2070-2100). Each row and column represents a single climate model (or observation). Each box represents a pair-wise combination, where warm colors indicate a greater distance. Distances are measured as a fraction of the mean inter-model distance in CMIP5.

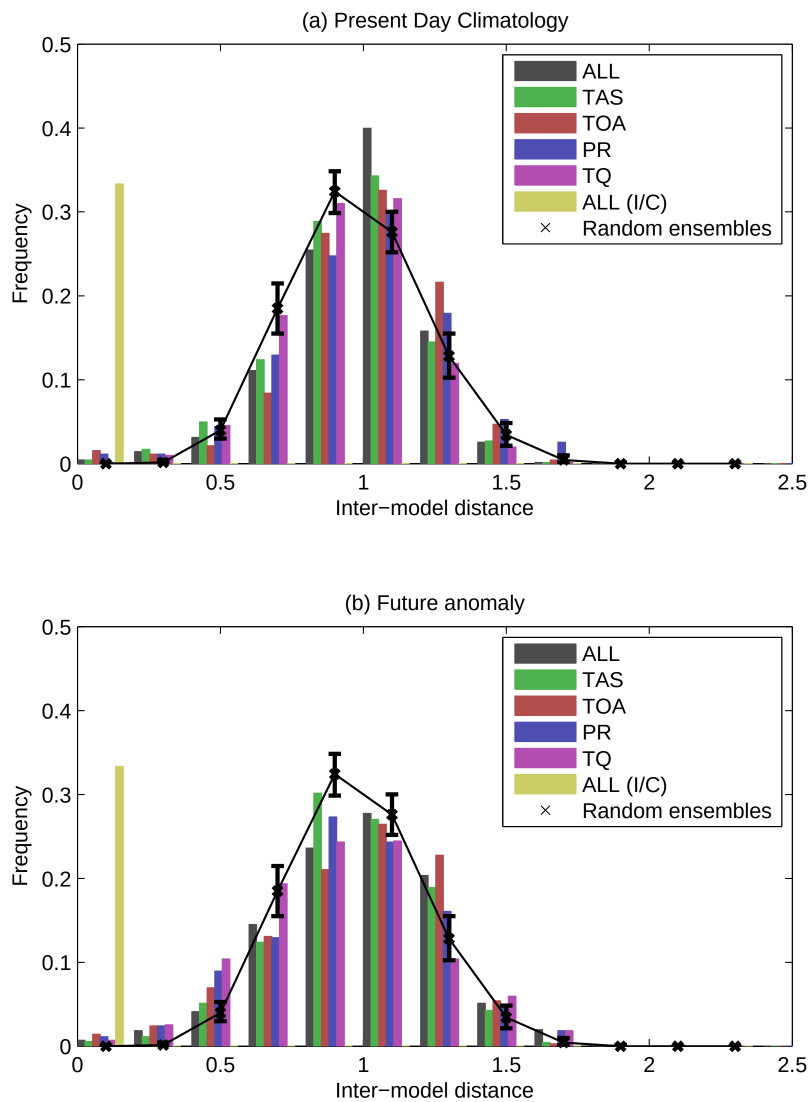

FIG. 2. Histograms of CMIP5 inter-model euclidean distances in the EOF loading space derived from (a) 1970-2000 monthly mean climatological fields as defined in Table 1 and (b) changes in the aforementioned fields between (1970-2000) and (2070-2100), as compared to a sample of 10^5 histograms calculated from randomly sampled distributions. Gray bars show the histogram of inter-model distances in the CMIP5 ensemble in an EOF space constructed with all available variables, while other colors show distances constructed with only a subset of variables; Surface Temperature (TAS), Top of Atmosphere Shortwave and Longwave fluxes (TOA), Total Precipitation (PR) and zonal mean temperature and humidity (TQ). The yellow bars indicate the distribution using all variables from the CCSM4 initial condition ensemble. The box and whisker plots show the range of bin values observed in the random distributions showing the 10th, 50th and 90th percentiles of the distribution.

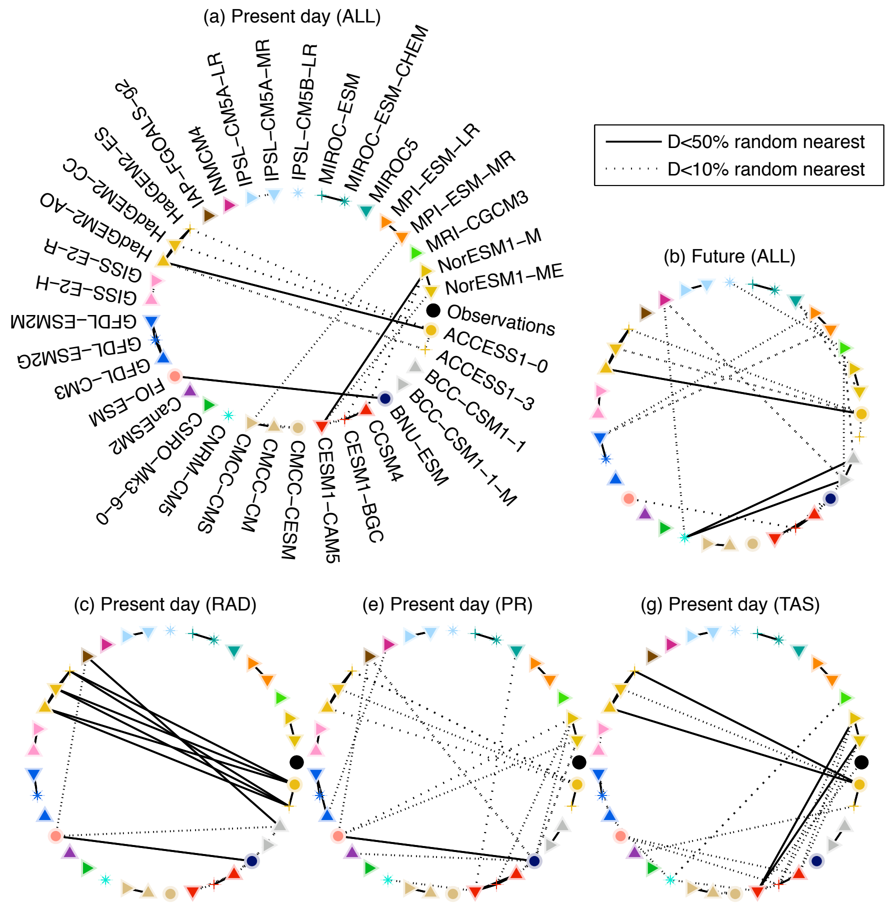

FIG. 3. An illustration of inter-model and observation-model distances in an EOF space defined by (a) 1970-2000 simulated climatology for 'ALL' variables and (b) the anomaly between 1970-2000 and 2070-2100 under the RCP8.5 scenario for 'ALL' variables. Plots are repeated for individual variables, Top of Atmosphere shortwave and longwave fluxes (c), Precipitation (d) and Surface Air Temperature (e). Inter-model lines illustrate where the inter-model distance is less than 50% (dotted) or 90% (solid) of nearest inter-point distances in a randomly generated distribution of with the same dimensionality, variance and population.

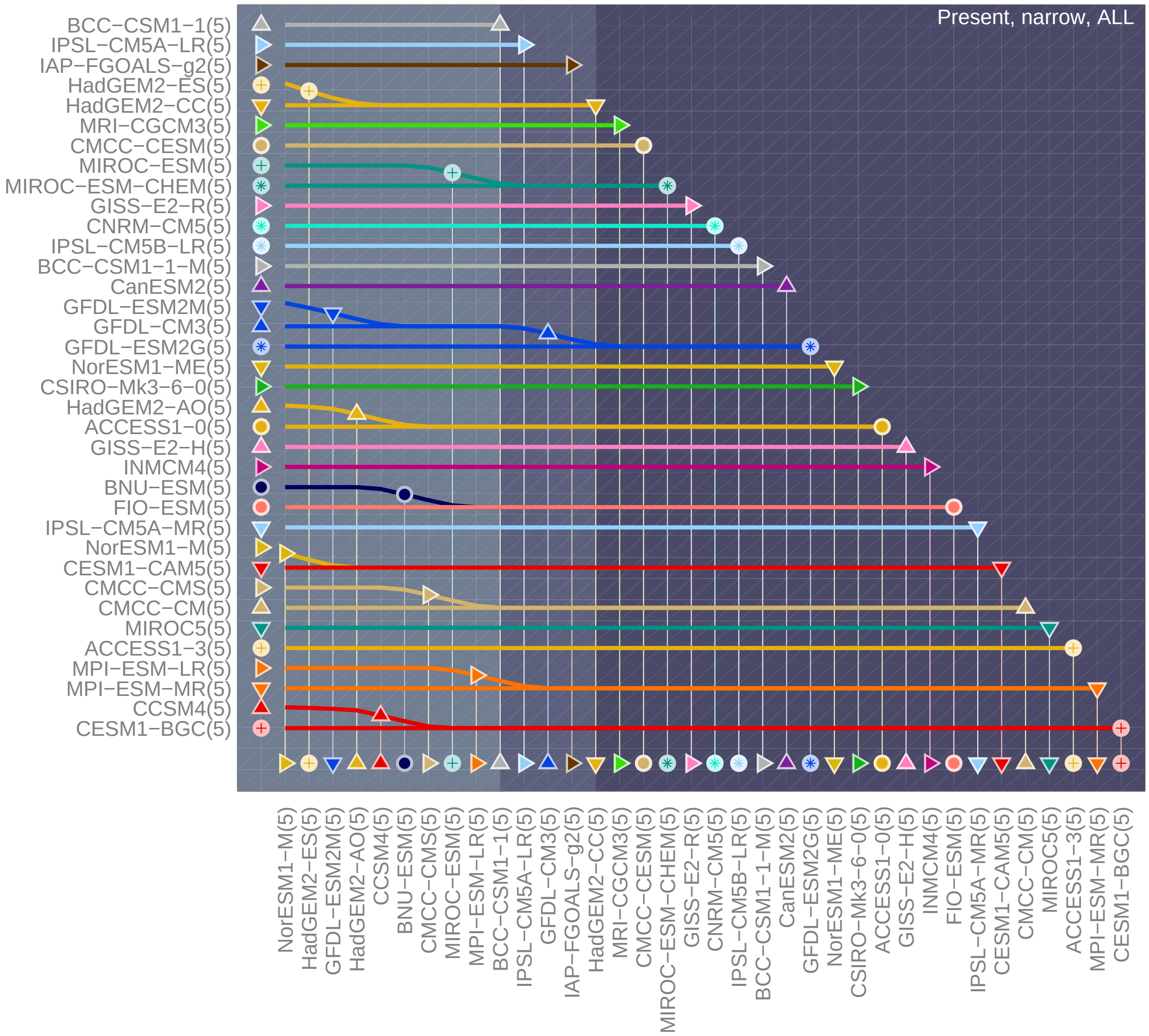

FIG. 4. An illustration of the stepwise model elimination procedure outlined in Section 2.e as applied to the 36 models from the CMIP5 ensemble, using model similarity information from the present day (1970-2000) climatology for 'ALL' variables and the 'wide' quality radius. The full set of models are shown on the left of each plot, and the order of model removal is shown on the bottom axis with the left-most model removed first. If the number of effective models n_{eff} decreases by less than 0.5, then the removed model is shown merging with its nearest neighbor in EOF space. If the number of effective models decreases by more than 0.5, the line is shown as ending - indicating the removal of that model family from the ensemble. Background shading indicates whether the smallest inter-point distance in EOF space using the remaining archive is less than 90% (light grey), 50% (mid grey) or 10% (dark grey) of purely random distributions of the same population, variance and dimensionality.

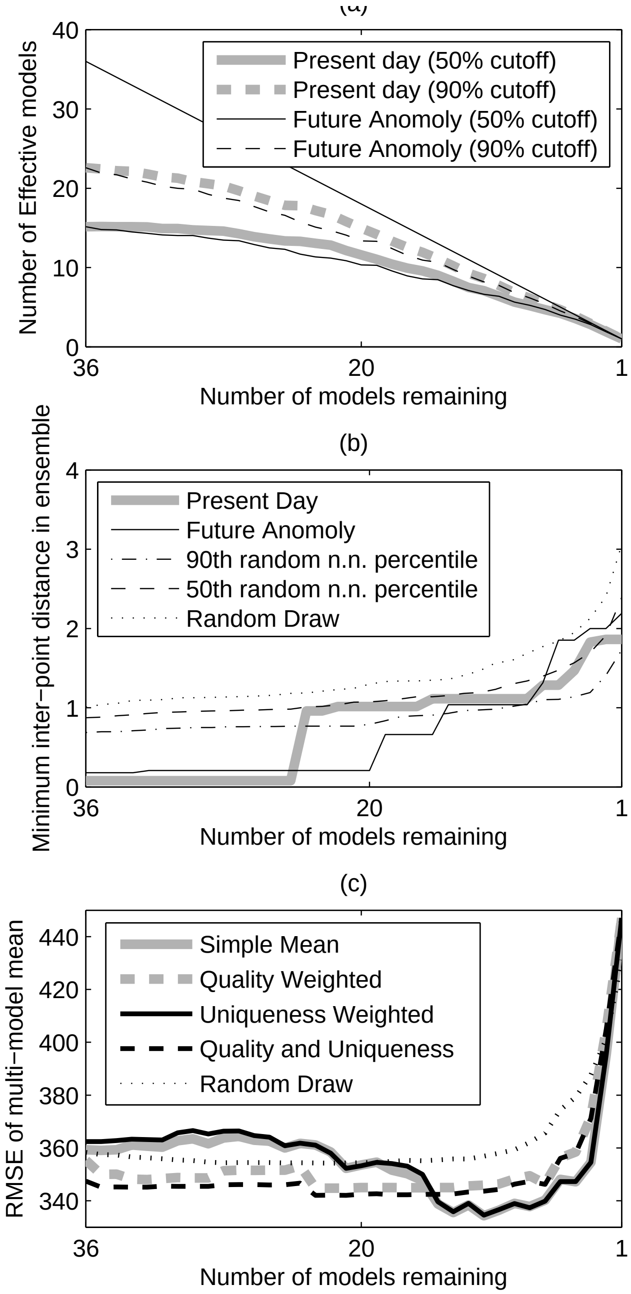

FIG. 5. Plots illustrating the stepwise model elimination following the procedure in Section 2.e. Calculations are conducted using model similarity metrics derived from both present day climatology and from future climate change under RCP8.5. (a) The number of effective models as a function of the number of actual models remaining in the ensemble. The percentile cutoff is the fraction of nearest neighbor distances seen in purely random ensembles used to define the radius of similarity D_u in Equation 10. (b) The nearest-neighbor distance as a function of the number of models remaining. For comparison, the 10th, 50th and 90th percentile of nearest neighbor distances in purely random ensembles of the same dimensionality and variance are shown. (c) RMSE of weighted and unweighted multi-model means as a function of remaining models.

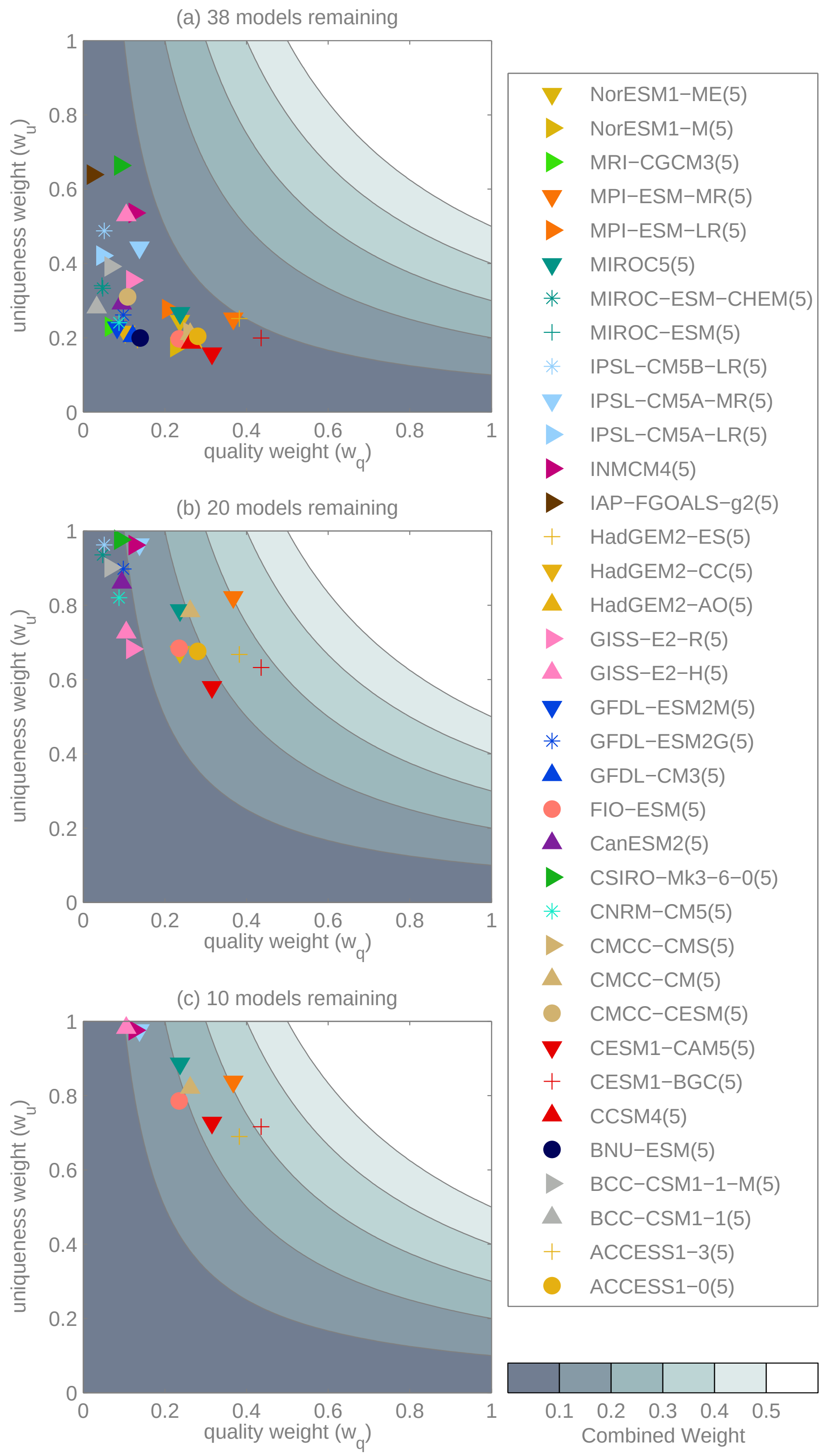

FIG. 6. A plot demonstrating how model uniqueness weights and model quality weights change as models are eliminated in the sequence shown in Figure 4, for (a) 36, (b) 20 and (c) 10 models remaining.

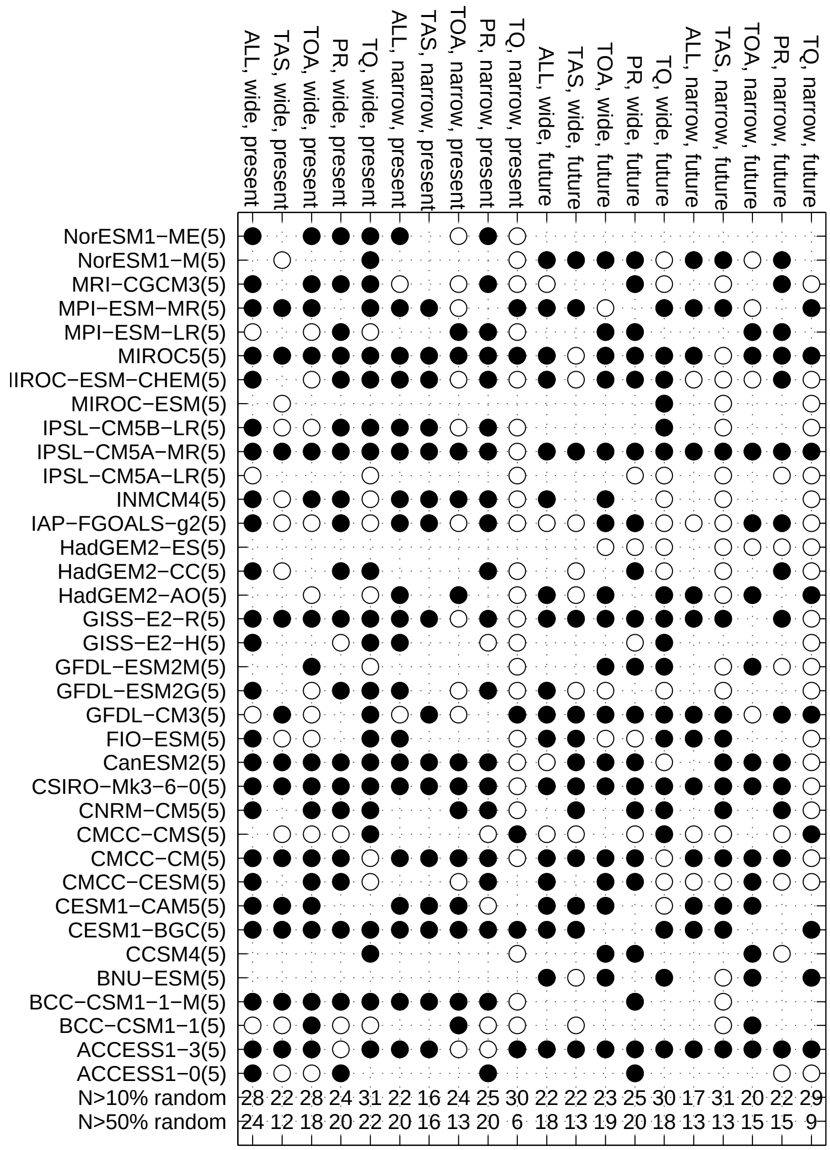

FIG. 7. A plot showing suggested subsets of CMIP5 given model quality scores and co-dependencies derived in a number of ways. Each line in the figure repeats the analysis leading to figure 4 with different assumptions. Plotted are the remaining models where the smallest inter-point distance in EOF space using the remaining archive is greater than 10% (unfilled symbols) or 50% (filled symbols) of purely random distributions of the same population, variance and dimensionality (regions marked by mid grey and dark dray shading in Figure 4). The analysis is conducted with zonal mean temperature and humidity (TQ), gridded precipitation (PR), gridded Top of Atmosphere shortwave and longwave fluxes (TOA), Gridded surface air temperature (TAS) and all variables combined (ALL). D_q, the radius of model quality is set to 'wide' or 'narrow' (the latter increasing the role of model quality metrics in model elimination). w_u, the model uniqueness weighting is shown calculated with the future RCP8.5 data, or the present day data. Numbers at the bottom of the plot indicate the number of retained models for the two conditions where the minimum remaining inter-model distance is greater than the 10th or 50th percentile of random smallest inter-model distances.

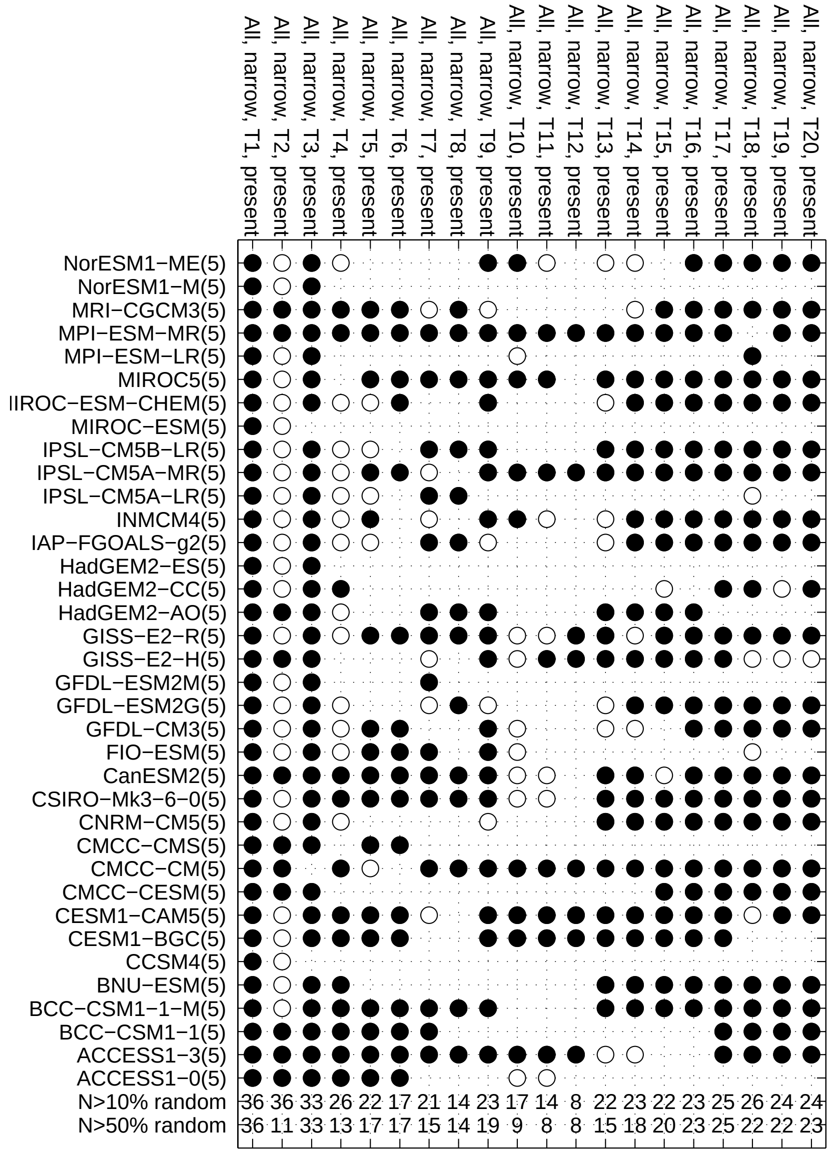

FIG. 8. A plot as in figure 7 showing suggested subsets of CMIP5 with different truncation lengths for the EOF analysis. Plotted are the remaining models where the smallest inter-point distance in EOF space using the remaining archive is greater than 50% (unfilled symbols) or 10% (filled symbols) of purely random distributions of the same population, variance and dimensionality (regions marked by mid grey and dark dray shading in Figure 4).