

# Addressing Interdependency in a Multimodel Ensemble by Interpolation of Model Properties

BENJAMIN M. SANDERSON

*National Center for Atmospheric Research,\* Boulder, Colorado*

RETO KNUTTI

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

PETER CALDWELL

*Lawrence Livermore National Laboratory, Livermore, California*

(Manuscript received 21 May 2014, in final form 10 March 2015)

## ABSTRACT

The diverse set of Earth system models used to conduct the CMIP5 ensemble can partly sample the uncertainties in future climate projections. However, combining those projections is complicated by the fact that models developed by different groups share ideas and code and therefore biases. The authors propose a method for combining model results into single or multivariate distributions that are more robust to the inclusion of models with a large degree of interdependency. This study uses a multivariate metric of present-day climatology to assess both model performance and similarity in two recent model intercomparisons, CMIP3 and CMIP5. Model characteristics can be interpolated and then resampled in a space defined by independent climate properties. A form of weighting can be applied by sampling more densely in the region of the space close to the projected observations, thus taking into account both model performance and interdependence. The choice of the sampling distribution's parameters is a subjective decision that should reflect the researcher's prior assumptions as to the acceptability of different model errors.


## 1. Introduction

At the time of writing, the Working Group I contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC AR5) has been published, summarizing the current best synthesis of projections for future climate change, and many of the studies referenced therein draw from a database of climate simulations that form phase 5 of the Coupled

Model Intercomparison Project (CMIP5). This multimodel ensemble comprises models with markedly different histories and degrees of independence. We use “independence” here to describe the degree of shared formulation and bias between models rather than in a strict statistical sense of orthogonality. Whereas some models have undergone a largely isolated development process, others share significant fractions of their code with other models in the ensemble.

Throughout the different generations of the CMIP, efforts have been made to integrate results into combined projections that represent some consensus view with associated uncertainty (see [Tebaldi and Knutti 2007](#); [Knutti 2008](#); [Knutti et al. 2010a](#); [Knutti 2010](#)). Many of these methods (and even a simple multimodel mean projection) are only appropriate if each model represents an independent estimate of future climate change. However, the reality of the CMIP ensembles is that some model pairs are closely related ([Masson and Knutti 2011](#); [Knutti et al. 2013](#); [Pennell and Reichler](#)

---

 Denotes Open Access content.

---

\* The National Center for Atmospheric Research is sponsored by the National Science Foundation.

---

*Corresponding author address:* Benjamin Sanderson, National Center for Atmospheric Research, 1850 Table Mesa Dr., Boulder, CO 80305.  
E-mail: bsander@ucar.edu

DOI: 10.1175/JCLI-D-14-00361.1

2011) and that some models exhibit more skill than others in reproducing past and present climate (Gleckler et al. 2008; Reichler and Kim 2008; Knutti et al. 2013).

The CMIP ensembles have been used on multiple occasions to propose “emergent constraints” or relationships between unknown climate parameters and observable quantities (e.g., Hall and Qu 2006; Fasullo and Trenberth 2012; Sherwood et al. 2014). However, the lack of independence of CMIP ensemble members can potentially complicate the interpretation of such studies (Caldwell et al. 2014). The sample size in CMIP3 is small (on the order of 20 models), and if one considers that many models share components and ancestry, then one could argue that it is effectively much smaller (Pennell and Reichler 2011; Annan and Hargreaves 2011; Jun et al. 2008). It is therefore difficult to demonstrate the significance of a single correlation that exists in the multimodel ensemble. Caldwell et al. (2014), Knutti et al. (2010b), and Abe et al. (2009) show that the correlation between the present-day and future climate patterns exhibited in models is often not significant. Potentially worse, the very presence of replicated models in the archive could potentially create artificial correlations (Caldwell et al. 2014), and the screening of predictors will likely find correlations that have no physical basis (DelSole and Shukla 2009; Masson and Knutti 2013). Including the CMIP5 models increases the sample size but does not necessarily solve the problem if there is sufficient common code in successive generations of each model.

It is self-evident that model replication has the potential to bias both multimodel means and emergent constraints (or correlations between observables and unknowns), but many more sophisticated Bayesian studies (e.g., Greene et al. 2006; Furrer et al. 2007; Tebaldi and Sansó 2009) also make the assumption that model members are independent estimates of future change, making it imperative that the degree of interdependency is quantified. Moreover, if it is found that model and code replication in the CMIP archives is commonplace, then strategies must be found for addressing these issues.

This assumption of model independence leads to greater confidence with an increasing number of models, which has led some to state that such methods implicitly consider the ensemble of future projections to be centered around truth (Knutti et al. 2010a). Others suggest a conceptual framework where individual models are indistinguishable from truth (Annan and Hargreaves 2010). This approach would consider the model ensemble to represent a sample from a distribution, which in an ideal case is the same distribution from which the real climate is drawn. Rougier et al. (2013) present a similar argument, that if one considers models in an ensemble as exchangeable and equally plausible, then

one need only make limited subjective judgements about whether a randomly drawn sample should lie closer to the true system state or to the ensemble mean in order to make robust predictions from the ensemble.

Bishop and Abramowitz (2013) propose a slightly different variant of the indistinguishable interpretation, that the observed climate is drawn from a set of potential “replicate earths” that represent different realizations of plausible internal variability of the climate system. Furthermore, they propose that an imperfect ensemble could be transformed to be centered on a best estimate of the true distribution of replicate earths with a linear transformation of the existing ensemble, which is optimized to be as close as possible to the observed data and weighted by error independence.

Finally, such questions of conceptual model paradigms are not necessarily absolute or fundamental (in that ensemble spread can be partly an artifact of model tuning approaches); Sanderson and Knutti (2012) suggest that a small subset of high performing models are indistinguishable from truth and a number of outlier models create some truth-centeredness in the ensemble as a whole. They also highlight that the statistical properties can change over time, such that the degree of truth-centeredness of the ensemble for the present day tells us little about the interpretation of the ensemble for the future.

The present study outlines a method for combining results from a multimodel ensemble, without relying on potentially spurious correlations between observables and model response and without making prior assumptions about a conceptual model framework. This technique can use both model output and observations but must satisfy two major requirements. First, the technique must allow for the selection of desirable characteristics such as a low climatological error compared to observational products, or a skillful historical transient projection (we focus less on the exact form of the metric, leaving this decision to the individual researcher depending on the projection or process in question). Second, we seek to reduce the bias in the multimodel projection arising from multiple closely related models present within the ensemble. This would be manifested in an extreme case when the same model is submitted twice to the ensemble.

In this first of two studies on this topic, we address this question for univariate or bivariate quantities, such as climate sensitivity or the temperature and precipitation change in a single region; producing continuous distributions for such quantities which are insensitive to model replication and with the potential to include model quality information. The method as we present it does not “solve” the problem of model interdependency, since our resampling is still sensitive to some of the properties of the

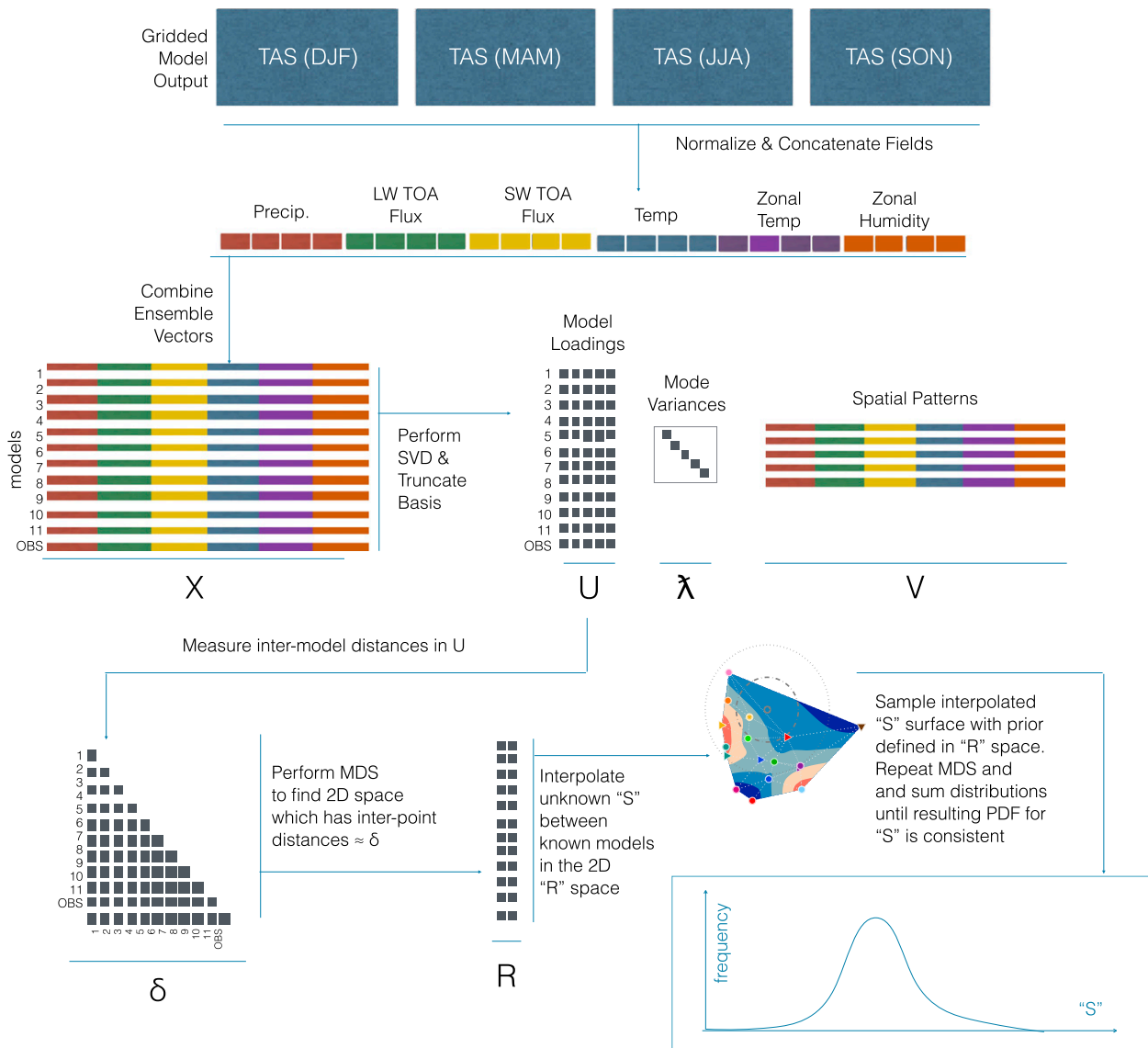


FIG. 1. Graphical representation of the methodology for this study.

original ensemble (model replication, and the presence of poor models). However, we do present a method that allows the researcher to explore the use of different sampling strategies for an ensemble of opportunity such as CMIP. In an accompanying paper (Sanderson et al. 2015), we propose a discrete method more suited to high-dimensional, multivariate gridded data.

## 2. Methods

### a. Data preparation

Our analysis assesses model quality and inter-dependence using output from model simulations of present-day climatology of a number of model

diagnostics summarized for easy reference in Fig. 1. One could also use transient metrics to assess the models, and the impact of such a change will be addressed in a further study. We do, however retain information for each month from the model output, so the model climatology includes a representation of both seasonality and annual mean state. Our approach combines a large number of gridded model outputs (listed in Table 1) into a high-dimensional metric. Using the recent historical mean state removes the necessity for concurrent data for all components of the metric and allows for easier comparison of the models in CMIP5 and its predecessor, CMIP3. It also avoids some issues with poorly constrained or missing radiative forcing components in some models.

TABLE 1. Observational datasets used as observations in Fig. 2.

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 W m <sup>-2</sup>
RSUT	CERES-EBAF	NASA (2011)	2000–05	25.8 W m <sup>-2</sup>
RLUT	CERES-EBAF	NASA (2011)	2000–05	3.32 mm day <sup>-1</sup>
<i>T</i>	AIRS*	Aumann et al. (2003)	2002–10	0.28 K
RH	AIRS*	Aumann et al. (2003)	2002–10	12.12 %

It is necessary to reduce the dimensionality of this problem to a small number of statistically independent variables. This is achieved, as in Sanderson and Knutti (2012), using a multivariate EOF analysis. We use historical climate simulations of the satellite era and variables are chosen such that observations or reanalysis products are available for the same time period and that the same spatial and temporal resolution. We perform an EOF analysis such that the observations are treated as an ensemble member, and thus both models and observations can be represented as points in the same orthogonal space.

This projection allows the computation of a simple model–observation discrepancy for the recent historical mean state of the climate (hereafter stated as simply “bias”) for each model. This bias term can be used to produce a model weighting, but its exact form depends on the choice of variables used in the multivariate metric. For the purposes of the present study, we have endeavored to produce a number of univariate and multivariate metrics from monthly mean, gridded latitude/longitude data from radiative fluxes, surface temperature, and precipitation, as well as zonal mean temperature and humidity data on model levels.

We now construct a distance metric used to evaluate both intermodel and model–observation distances. We construct this metric in a space defined by a multivariate EOF basis set. 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 1 January 1970–31 December 1999. In the case of CMIP3, we use the 20c3m experiments, using “run1” from each simulation (with the exception of CSIRO Mk3.0, for which the first available run is “run2”). For CMIP5, we use the “historical” experiment and the r1i1p1 simulations in each case. In the 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).

Data were downloaded from the Earth system grid for five two-dimensional fields [surface temperature (TS), 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 [sensitivity of results to the choice of variables is presented in section 3e(2)].

We then prepare the data in the same fashion as Sanderson and Knutti (2012), and we repeat the critical steps (listed in the supplementary material of that paper) here for convenience. Data from each model and dataset are regridded onto a 2.5° by 3.75° latitude/longitude grid, and zonal vertical fields are regridded onto a 2.5° latitude grid at 17 pressure levels. For each variable, values are area weighted and for vertical fields, weighted by the pressure difference between the top and bottom of the corresponding level.

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 multimodel ensemble. We derive a single normalization factor (a scalar) for each variable type from the observational fields (see Table 1). For two-dimensional fields, we calculate the intermonthly variance of tropical grid cells and take the average gridcell variance over the tropics to obtain a single normalization factor for each variable. For three-dimensional fields, we take the intermonthly 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 for each variable are calculated from the observations only, and the corresponding output from each model is divided by the same factor (see Table 1 for global normalization values).

The data are then prepared for the EOF analysis; the elements of each two- and three-dimensional field are then each reformed into a one-dimensional vector. If any elements of the vector in any single model or in the observations are missing, the corresponding elements are removed from all models. Each of the field vectors is then normalized by the number of remaining elements, and the second and third fields are concatenated into a

single vector length  $n$  (where  $n = 358\,248$  when all fields are utilized). The  $\mathbf{m}$  vectors are combined to form a matrix  $\mathbf{X}$  size  $m$  by  $n$  (where  $m$  is 51, comprising 50 CMIP model vectors and one observational vector). 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  $\Delta\mathbf{X}$ . The method effectively treats the observations as an additional ensemble member, so the observed data are included in the multimodel mean. The analysis is also repeated with a number of different subsets of the entire set of variables [section 3e(2)]. In these cases, the matrix  $\Delta\mathbf{X}$  is formed using only that subset, and the analysis continues in the same fashion.

### b. Accounting for intermodel similarity

#### 1) COMPUTING PAIRWISE DISTANCES

It is desirable to minimize the impact of correlated fields in  $\Delta\mathbf{X}$ , and so we perform an EOF analysis to reduce the data to a small number of orthogonal components. The use of the EOF prefilter combines fields that are trivially correlated (such as adjacent grid cells) into a single mode. A singular value decomposition (SVD) is performed on  $\Delta\mathbf{X}$  and truncated to  $t$  modes to obtain the dominant modes of multivariate ensemble variability such that

$$\Delta\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T, \quad (1)$$

where  $\mathbf{U}$  is an orthogonal matrix of model loadings (size  $m$  by  $t$ ) whose columns are the eigenvectors of the model covariance matrix  $\Delta\mathbf{X}(\Delta\mathbf{X})^T$ ,  $\mathbf{\Lambda}$  (size  $t$  by  $t$ ) are the eigenvalues of  $\Delta\mathbf{X}(\Delta\mathbf{X})^T$ , and  $\mathbf{V}$  (size  $n$  by  $t$ ) are the eigenvectors of the field covariance matrix  $\Delta\mathbf{X}(\Delta\mathbf{X})^T$ . The dimensions are sorted by decreasing eigenvalue, such that the basis set can be truncated to a smaller number of modes  $t$  (where  $m$  modes define a complete basis that can be used to reconstruct the original data, so for a truncated case  $t < m$ ).

The model loadings  $\mathbf{U}$  now define a  $t$ -dimensional space (where  $t$  is the truncation length of the SVD) in which intermodel and observation-model Euclidean distances may be defined [see section 3e(1) for justification]. The intermodel distances can then be measured in a Euclidean sense in the loadings matrix, such that the distances  $\delta_{ij}$  between two models  $i$  and  $j$  can be expressed as

$$\delta_{ij} = \left\{ \sum_{l=1}^t [\mathbf{U}(i,l) - \mathbf{U}(j,l)]^2 \right\}^{1/2}. \quad (2)$$

For the present-day cases, the model–observation distance  $\delta_{i(\text{obs})}$  is calculated using the row of  $\mathbf{U}$  corresponding to the observations.

#### 2) MULTIDIMENSIONAL SCALING

Multidimensional scaling (MDS) is a technique that can be used to take the distance matrix  $\delta$  and create, if possible, a distribution of points in a two-dimensional space ( $R$ ) that exhibit approximately the same interpoint distances as those derived from the original  $t$ -dimensional space. The input to the multidimensional scaling step is the distance metric  $\delta_{ij}$ , where  $i$  and  $j$  refer to different models in the ensemble, or to the observational point (making  $m$  points in total). Euclidean distances  $\delta_{ij}$  are calculated from the  $\mathbf{U}$  values [section 2b(1)], which are concatenated to form a matrix dimensions  $m$  by  $t$ , where the final row corresponds to the observations. We apply a metric MDS algorithm to solve for  $\mathbf{R}$ , which minimizes a performance function or *stress* for the distribution of model dissimilarities. We require a solution where  $\mathbf{R}$  has dimensions  $m$  (where  $m$  is the number of samples, 51) by  $p$  (where  $p$  is 2, our desired dimensionality). Approximate intermodel distances in  $\mathbf{R}$  can be described as follows:

$$d_{ij}(\mathbf{R}) = \left[ \sum_{s=1}^p (x_{is} - x_{js})^2 \right]^{1/2}. \quad (3)$$

The algorithm thus minimizes a “metric stress” term  $\sigma_1$ , taken here as Kruskals Stress-1 formula (Borg and Groenen 1997):

$$\sigma_1(\delta, \mathbf{R}) = \left\{ \frac{\sum_{i=1}^m \sum_{j=1}^{i-1} [\delta_{ij} - d_{ij}(\mathbf{R})]^2}{\sum_{i=2}^m \sum_{j=1}^{i-1} d_{ij}^2(\mathbf{R})} \right\}^{1/2}, \quad (4)$$

which is equal to the RMSE in the replication of the intermodel distance distribution divided by the sum of squared distances. The optimal solution is obtained with the SMACOF (scaling by majorizing a convex function) algorithm (de Leeuw 1977). Again, a row of the matrix  $\mathbf{R}$  corresponds to observations.

The solution is approximate, so to reduce the dependency of later results on any one specific optimization, we repeat the process for 20 instances with randomized initial values for  $\mathbf{R}$ . A separate interpolation is conducted for each of the 20 cases, and the distributions for the interpolant are averaged to form the final result. This was found to be sufficient to produce smooth and reproducible distributions in the later part of the study. Section 3c(2) examines the accuracy of this assumption in more details.



### c. Ensemble interpolation

Interpolation is the process of estimating the values of unknown points lying within the convex hull of a set of known data points. While there are numerous techniques whereby the interpolation in the reduced space might be achieved, but we perform a Delaunay triangulation in two dimensions (Delaunay 1934) using the points defined by each model. To interpolate in this space, we use “natural neighbor interpolation” (Sibson 1981), which has the advantage of producing a continuous interpolating surface between sampled points. Once the interpolant is generated in the MDS space, it can be used to resample that space with different prior distributions.

To perform the resampling procedure, we create a  $p =$  two dimensional distribution of points in the MDS space and use the interpolant detailed above to predict a distribution of the unknown variable corresponding to that distribution. We then consider a number of idealized prior distributions defined in the MDS space that we can use to produce distributions for the variable of interest. We first consider a uniformly sampled prior where the space is sampled regularly in the two dimensions, with  $10^3$  regularly spaced intervals in between the maximum and minimum values of  $\mathbf{R}$  in each dimension, making  $10^6$  domains in total. If a domain lies outside the convex hull of the ensemble, it is rejected and if a domain lies within the ensemble, the corresponding unknown value is calculated using the interpolation scheme detailed above. The resulting distribution is normalized by the number of domains within the convex hull to form a single distribution. The process is repeated for  $N = 20$  instances, with different random seeds to initiate the MDS algorithm. The final distribution is the sum of the distributions from each instance, divided by the number of instances. Note that although the MDS step makes the interpolation process tractable, a uniform sample in the MDS space does not necessarily lead to a uniform sample in the space defined by  $\mathbf{U}$ , as the MDS result effectively defines a surface that is folded to pass through points in a higher-dimensional space.

The process for implementing the Gaussian truth-centered prior is slightly different. A  $10^6$  member two-dimensional normal distribution is created, centered on the values of  $\mathbf{R}$  corresponding to the observations. The standard deviation of the distribution determines how tightly the observations should constrain the result, and should ideally reflect uncertainty in the positioning of the observed point. Uncertainty could arise from natural variability or errors in the observations or reanalysis. If we consider model internal variability to be a proxy for natural variability [although Haughton et al. (2014)

suggest that this might be a slight underestimate], considering these terms alone was found to produce a prior sufficiently narrow that the majority of models were eliminated from consideration (in other words, most models are inconsistent with observations within the range of internal variability and observation uncertainty). Hence, in order to provide a weaker constraint on the space, we choose a prior with a standard deviation equal to that of the CMIP5 archive by calculating the variance of  $\mathbf{R}$  for the rows corresponding to models in CMIP5.

The use of the CMIP5 ensemble variance to define the prior is clearly an arbitrary decision, forced by the lack of alternative. The variance itself can clearly be influenced by model replication or by the presence of very poor models in the archive. Hence, to resample the ensemble in this fashion does not address fully the issue of model interdependency. However, using this length scale allows us to resample the space at a scale that keeps the influence of the (relatively) better performing models while reducing the effect of the (relatively) poorer performing models. Sensitivity to this value is illustrated by repeating the calculation with a distribution with half the variance of  $\mathbf{R}$ . Corresponding values for the unknown quantity are calculated for each of the  $10^6$  points using the interpolation scheme described above.

This basic process can be used to interpolate simultaneously for a large number of model properties. Hence, for each of the  $10^6$  “metamodels,” interpolated values are calculated, along with loadings for EOFs that can be used to estimate the metamodel’s position in the EOF space, and thus calculate its bias from observations. Because the interpolation is based upon a Delaunay triangulation, each metamodel can also be expressed as a combination of three or fewer component models from the original CMIP archive.

## 3. Results

### a. Intermodel similarity

We first present results showing the distribution of pairwise distances between models in both CMIP3 and CMIP5, plotted in Fig. 2. The distances are Euclidean distances in the nine-dimensional space defined a multivariate EOF analysis described in section 2b(1).

Several properties of the ensemble are apparent from Fig. 2; some (in particular older) models (GISS-E-H, IAP FGOALS-g1.0, INM-CM3, NCAR PCM1; expansions of all model names are available online at <http://www.ametsoc.org/PubsAcronymList>) appear to be outliers with large distances to all other models and to the observations. Models with common heritage

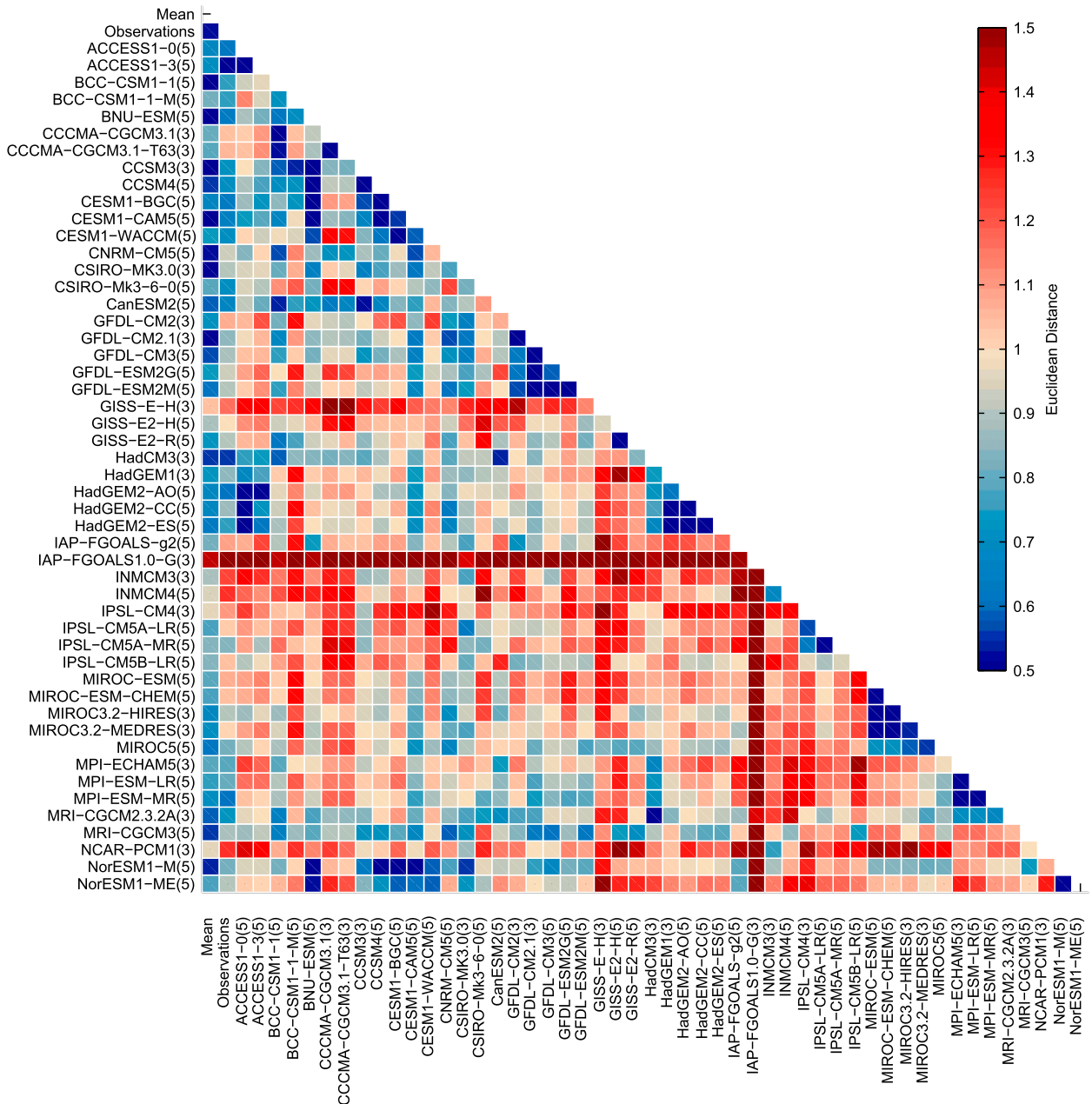


FIG. 2. A graphical representation of the intermodel distance matrix for CMIP3, CMIP5, the multimodel mean, and a set of observed values. Each row and column represents a single climate model (or observation). Each box represents a pairwise combination, where warm colors indicate a greater distance. Distances are measured as a fraction of the mean model bias in the combined CMIP3 and CMIP5 ensembles.

(GFDL, CCSM/CESM, MIROC, etc.) tend to be closer to each other than they are to other models in the ensemble. Models from different institutions with common components (such as CCSM4 and NorESM1) are correctly identified as near-neighbors in the ensemble. The intermodel distances are found to be relatively robust to changes in EOF truncation length and to changes in the

diagnostic fields used [see section 3e(1)], and very similar to those found by Knutti et al. (2013), in which the supplementary material discusses the intermodel relationships at greater length.

This similarity information is clearly relevant to the issue of some models in the CMIP ensembles being overrepresented, and one could potentially use it to

down-weight highly replicated models within the ensemble [see Sanderson et al. (2015) for further thinking on this]. However, a weighting approach would still result in a discrete and to some degree arbitrary sample of climate properties, where in some cases a continuous distribution would be preferable. A possible approach might be to interpolate values directly within the truncated EOF space, but this is infeasible because the number of models is comparable to dimensionality of the space, which effectively reduces the problem to a linear regression with global predictors of unknown variables. Furthermore, robust emergent constraints on future climate behavior in multimodel ensembles are rare and may be spurious (Caldwell et al. 2014; Huber et al. 2011).

### *b. Multidimensional scaling*

Our proposed solution is to use multidimensional scaling (MDS) to represent the models and observations as points on a two-dimensional surface, the distribution of which approximately preserves the intermodel distances in the EOF space. This has a number of advantages; first, it allows the construction of a continuous, smooth interpolant of climate properties between models in the archive. This interpolant is not dependent on any global relationships or emergent constraints existing to relate observable quantities to unknown quantities because it is simply a surface fitted through all ensemble points without requirements for monotonic behavior over the domain. The process does make the assumption, however, that model states can be locally interpolated. For example, if one model has a state  $p(1)$  and a similar model has a state  $p(2)$ , the process assumes that  $p(3) = [p(1) + p(2)]/2$  is also a possible model. Although it might not be possible to construct the model  $p(3)$  (just as it would not be possible to produce a model that behaves like the CMIP multimodel mean), it is a useful construct because it gives us the possibility of experimenting with different sampling strategies.

If there are no global relationships between the predictor state (in this case, the position in the MDS space), and the predicted quantity, preferentially resampling the surface more densely close to the observations will produce a similar distribution to resampling the surface uniformly, which allows us to conclude that the diagnostics used to construct the MDS space are not useful for constraining the variable we might be interested in. However, if such relationships *do* exist, and the relevant predictor is included in the calculation of the MDS space, then sampling the interpolated surface close to the observed point will result in a tighter constraint on the variable of interest than the uniformly sampled case.

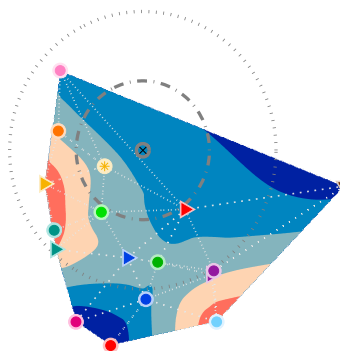
The MDS process uses the matrix of dissimilarities (Fig. 2) to form a distribution of points in a two-dimensional space, where the distance between models best approximates the intermodel distances shown in Fig. 2 [the algorithm used in this study is described in section 2b(2)]. Uncertainty due to the approximations arising from the process can be sampled by conducting multiple realizations. Sample results for a single calculation are shown in Fig. 3, where the subset of points corresponding to CMIP3 and CMIP5 are shown in Figs. 3a and 3b and the combined ensemble calculation is shown in Fig. 3c.

In each case, the pairwise intermodel distances approximate those plotted in Fig. 2 (as do the distances from models to observations). Clearly defined clusters of models correspond to models from different institutions (e.g., all Hadley Centre models or all the GFDL models lie closer to each other than they do to any other model in the CMIP archive). There are some cases where models are effectively submitted more than once; for example, in CMIP3, CCCMA-CGCM3.1 is submitted at two different resolutions and in CMIP5 some models are submitted both with and without interactive chemistry (e.g., HadGEM2-ES and HadGEM2-CC). In each of these cases, the algorithm positions these models at negligible distances from each other. Newer model versions often stay close to their predecessors, supporting the idea of some evolutionary process in which models are improved or changed (mutation), successful concepts or code are shared (cross-breeding), and poor models are eliminated (selection).

The MDS space attempts to represent intermodel differences in the nine-dimensional EOF space on a two-dimensional surface. As such, its dimensions are no longer associated with a single physical pattern. Thus, by building an interpolated surface in this space, we do not require the existence of global relationships between predictors and predictands; rather, we assume only that a model that is close to another model in the MDS space (and therefore also close in the EOF space) is also likely to have a similar value of the predictand, in the absence of any other data. As an example, Fig. 3 demonstrates a linearly interpolated surface for equilibrium climate sensitivity (the equilibrium global mean surface temperature response to a doubling of carbon dioxide, hereafter  $S$ ) using the known values in the CMIP3 and CMIP5 ensembles (see section 2c for details on the interpolation). The MDS process is conducted for the combined CMIP3/CMIP5 ensemble (Fig. 3c); for consistency, the same 2D coordinates are used to represent the models for the CMIP3 (Figs. 3a and 3b, respectively). Although the model coordinates are identical in the combined CMIP3/5 case and the individual CMIP3



(a) CMIP3



(b) CMIP5



(c) CMIP3/5

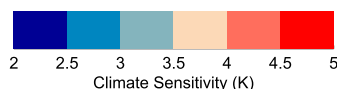
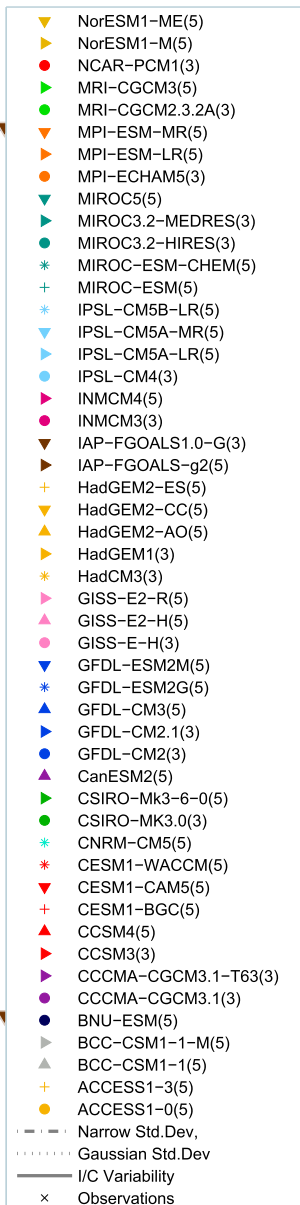
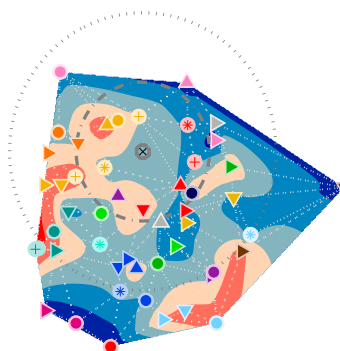


FIG. 3. Graphical representation of sample MDS results for (a) CMIP3, (b) CMIP5, and (c) combined CMIP3/CMIP5 ensembles. The concentric circles represent Euclidean distance from the observational projection in the space, meaning better performing models lie closer to the origin. Ellipses are centered on the observed projected point, while their radii represent the standard deviation of the “Gaussian” and “narrow” distributions in the two dimensions of the MDS space; there is also an ellipse representing the spread one would expect from climate variability alone. Each symbol represents one model in the ensemble, with symbols of the same color generally representing models from the same institution. Numbers in brackets indicate whether a model is in the CMIP3 or CMIP5 archive. Colored shading indicates interpolated values for  $S$ .

and CMIP5 cases, a separate interpolation of climate sensitivity is carried out in each case.

Figures 3a and 3b show qualitatively that CMIP5 contains a relatively larger number of models with a comparable bias to the best performing models in CMIP3, consistent with the results of Knutti et al. (2013). This means that the interpolated surface for  $S$  in Figs. 3b and 3c has more points to constrain it in the region close to the observations than in Fig. 3a.

### c. Univariate projections

#### 1) PROBABILISTIC DISTRIBUTIONS

We propose forming posterior distributions for unknown climate quantities of interest with an appropriate prior defined in the MDS space. We consider some idealized possibilities; a uniform prior, equivalent to correcting for dependence but ignoring bias, which samples the space uniformly within the “convex hull” of the ensemble (the convex hull represents the outer boundaries of the largest possible polygon formed about the available models) and two Gaussian priors centered on the observations (further details in section 2c).

In an ideal case, the radius of model acceptance would be governed by internal variability—and thus the model’s skill score would represent the chance that the model’s climatological state would be produced in by the real world climate variability. However, we find that the variance due to internal variability alone is orders of magnitude smaller than that due to the spread in multimodel bias [a similar result is suggested by Haughton et al. (2014)]. We illustrate this effect in Fig. 3 by projecting all six available members in the CMIP5 archive from CCSM4 onto the MDS space in the same manner as before, and measuring the standard deviation of that projection in the two dimensions. These distances are represented as the smallest gray circle surrounding the observations in Fig. 3. Assuming that CCSM4 variability is a reasonable proxy for real-world climate variability, all models in the CMIP5 and CMIP3 would be an order of magnitude or more farther from the observations than could be explained by internal variability. Thus, any resampling conducted using natural variability as the radius of acceptability would then effectively be a point sample of the interpolated surface, and would rule out all models in the original CMIP archive.

We must therefore adopt a more pragmatic approach, where the width of the distribution is essentially a subjective decision as to what degree the researcher believes that the observations constrain the available models. In this case, we choose the width of the distribution such that the variance of the interpolated distribution is equal to that of the original ensemble. This

emphasizes the relative skill of the different points in the ensemble, without distinguishing too harshly between the weight allocated to the better performing models. This is clearly subjective, so we test the implications of using a tighter constraint by producing a second set of distributions with a variance half that of the original ensemble, and repeating the analysis as before (hereafter referred to as the “narrow” prior).

We begin by using the interpolant to produce a distribution for a known quantity, in this case global mean surface temperature (TS) between 1970 and 2000. Figure 4a shows the distributions for TS for the CMIP3, CMIP5, and combined ensembles as well as continuous distributions derived from the ensemble interpolation strategy described in the previous section, for three priors: “Gaussian” and narrow (which are both centered on observations and illustrated by the concentric circles in Figs. 3a–c), and “uniform,” which samples the entire interpolated surface  $R$  uniformly. The plot shows that there is a perhaps surprising variation in the historically simulated global mean temperatures in the CMIP3 and CMIP5 archives, with the entire ensemble range spanning almost 3 K. The plot also shows that the CMIP3/5 interpolated distributions derived from the Gaussian and narrow priors have median values within 0.1 K of the observed value, while the uniformly sampled distribution is biased 0.4 K low in TS, similar to the uninterpolated raw output of the combined CMIP3/5 ensemble. Clearly, one would expect the method to perform well in this case because the observed value of TS is part of the combined metric used to create the observational point in Fig. 3. If one considers the distributions derived using variables other than TS, the median of the resampled distribution is farther from the observed value of global mean temperature (but the observed value lies within the 10th and 90th percentiles of the distribution irrespective of the variables used to create the MDS space).

We can also produce distributions for unknown aspects of future climate. For example, distributions for  $S$  using each prior are shown in Fig. 4b. Using either prior, the resampled distributions for  $S$  are smooth, and the median value for climate sensitivity lies between 3 and 3.5 K for all variable choices and model selections (see Fig. 4b). Weighting the combined CMIP3/CMIP5 ensemble up-weights the importance of the low-bias and high-sensitivity models in CMIP5 such as MPI-ESM-LR, CESM1-CAM5, and CanESM2 and down-weights some of the low-sensitivity outliers from CMIP3 such as NCAR PCM1, INM-CM3, and IAP-FGOALS-g1.0. This tends to increase the lower bound on climate sensitivity as the prior becomes narrower (the 10th percentile for climate sensitivity is 2.5, 2.8, and 2.9 K using

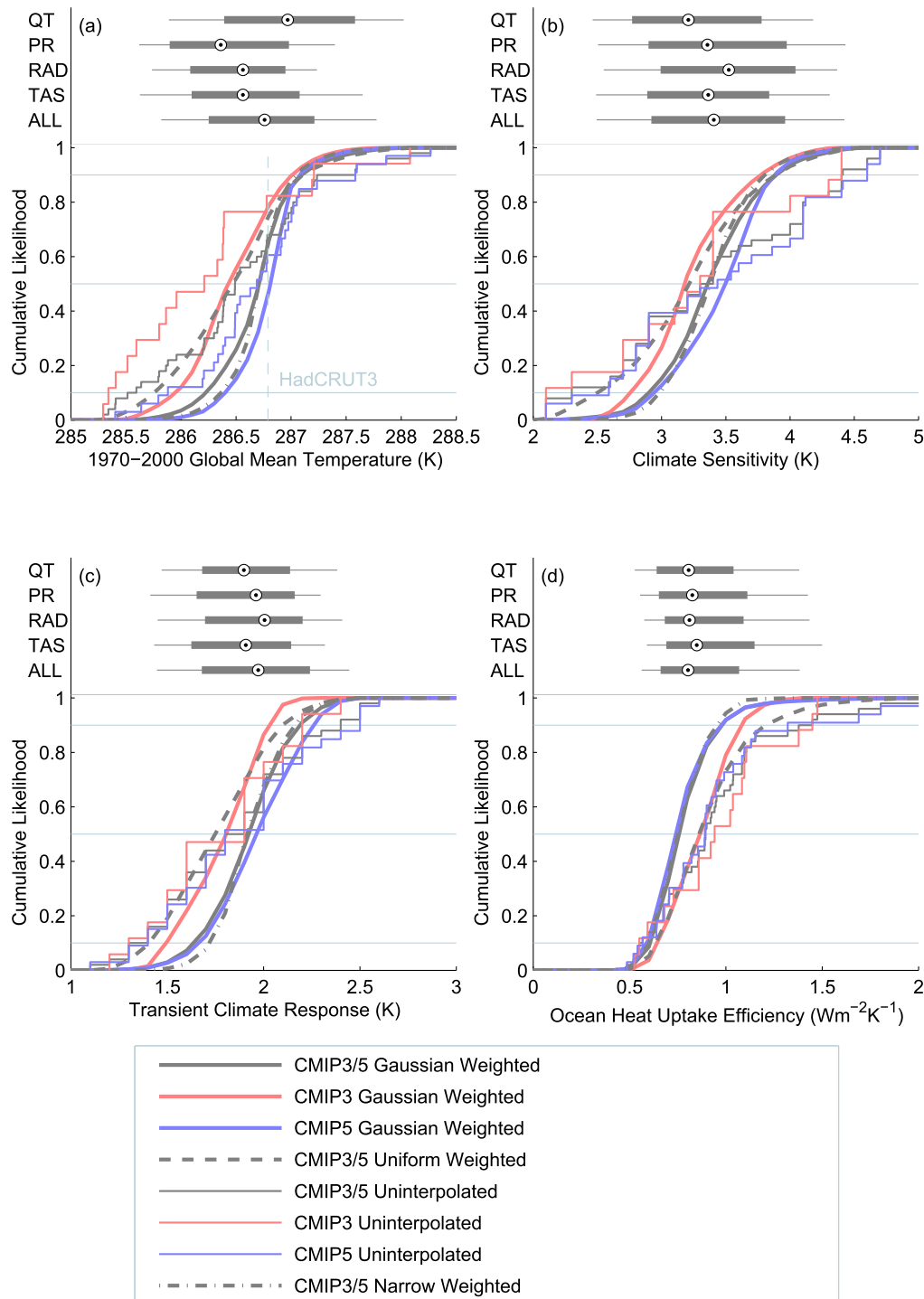


FIG. 4. Cumulative distributions for (a) global mean historical surface temperature, (b) climate sensitivity, (c) transient climate response, and (d) ocean heat uptake efficiency assuming different priors in the interpolated space depicted in Fig. 3, using “ALL” variables. Thick solid curves show distributions for CMIP3 (red), CMIP5 (blue), and the combined ensembles (gray) where the space is sampled with a Gaussian distribution centered on the observed climate. Dashed lines show distributions resulting from uniform prior sampling within the convex hull of the ensemble distribution. Stepped thin lines indicate histograms of the original distributions within each ensemble. Bars and whiskers at the top of the plot indicate the 90th and 99th percentiles of the distribution when the calculation is repeated for different subsets of the full variable set (QT is zonal mean temperature and precipitation, PR is surface total precipitation, TAS is surface temperature, RAD is top-of-atmosphere short-wave and longwave fluxes, and ALL is all variables combined).

the unweighted, Gaussian, and narrow priors respectively). Figure 4c tells a similar story for transient climate response, decreasing the width of prior results in an increase in the lower bounds for TCR (1.3, 1.6, and 1.7 K using the unweighted, Gaussian, and narrow priors, respectively). The upper bound on ocean heat uptake efficiency is constrained toward lower values as the distribution is tightened, with a 90th percentile of 1.2, 0.9, and 0.9 W m<sup>-2</sup> K<sup>-1</sup> for the unweighted, Gaussian, and narrow priors respectively (the lower bound is relatively unaffected).

## 2) ASSESSMENT OF THE METHOD

In the introductory section of this paper, we set out to address two issues with sampling unknown parameters from model ensembles of opportunity, with the inclusion of low-quality outliers and the potential for model replication biasing the ensemble result. In this section, we present some evidence to demonstrate that we have indeed addressed these issues to some extent.

Figure 5a serves two purposes; first it shows that the distances from models to observations after the MDS process (averaged over 20 iterations) are tightly related to the original distances from models to observations in the original EOF space. The plot shows the distance of each model in the ensemble to the observed point in both the original EOF space and the MDS space, showing that the two are highly correlated (although the plot shows that very small distances are slightly underestimated and large distances overestimated relative to their original values). If the distances are taken as a measure of model bias, it can also be seen that the resampling process can effectively eliminate consideration of those models with a large bias when forming a posterior distribution for an unknown quantity. If we believe that the calculated bias is informative, a tight prior centered on observations allows us to satisfy the first requirement, namely that very poor models should not influence our result.

Figure 5b attempts to illustrate how successful our method is for addressing the second requirement; that duplicated models should not influence our result. We assess model replication with a simple *k*-means cluster analysis, using the EOF loading matrix, **U** (sized *m* by *t*), allowing 14 clusters. The models associated with each cluster are shown in Fig. 5b, again showing much of the structure which might be expected from Figs. 2 or 3, with models from the same institution tending to fall in the same cluster. The histograms in Fig. 5b show how much weight is allocated to each different cluster in the original CMIP3/5 ensembles, as well as in the resampled distributions. In the former case, the histogram simply indicates the number of models in each cluster, with some

being highly represented (such as cluster 4, the Hadley Center models, or cluster 8, the GFDL models) and some with only one member (such as cluster 13, IAP FGOALS).

To evaluate the weight of a given model *n* in the resampled ensembles, we use the same method that we use to interpolate *S* or TCR in section 2c, but instead we interpolate a vector that is 1 for the element *n*, and 0 elsewhere. Each interpolated model will then have a value between 0 and 1 that represents the fractional contribution of model *n*. This can be repeated for each value of *n* to show the relative makeup of each interpolated model in terms of its CMIP3 and CMIP5 constituents. Integrating these values over all the models in the resampled Gaussian and uniform priors then yields the contribution of each of the original CMIP3 and CMIP5 models to those resampled ensembles. These are combined according to the clusters defined in Fig. 5b, showing whether the relative contribution of each cluster is increased or decreased by the resampling process.

The results for the uniform sampling strategy show a subtle redistribution of weight amongst the clusters. The weights associated with clusters with large numbers of models in the original ensemble such as 4, 8, and 12 are slightly reduced in the uniformly resampled ensemble, whereas clusters with a low representation (such as 13 and 5) are slightly increased. The effect is subtle, though, and the histogram associated with the original CMIP3/5 ensemble tends to indicate that the weights associated with major modeling centers are all well represented (i.e., the original ensemble is not strongly biased by overreplication of a single model). The weighted Gaussian and narrow ensembles show, as expected, a shifting of weight toward clusters that contain models with a low bias.

Finally, we address the issue of how accurate the method is for predicting *S* for an out-of-sample model, and whether indeed this is necessary. Figure 5c is calculated by repeating the method of section 2c, but omitting a single model when constructing the interpolation surface for *S*. The resulting interpolated surface is then used to predict the out of sample value of *S* for the missing model, plotted as a function of the actual value of *S* in Fig. 5c. The process is repeated for 20 iterations of the MDS process, as before—with the mean result and the spread given by the position of the markers and whiskers in the plot.

The predicted value is correlated with the actual value with a coefficient of 0.66, and there are clearly some occasions where the interpolation fails. For example, if CESM-CAM5 is removed from the ensemble, the predicted sensitivity is considerably less than the actual

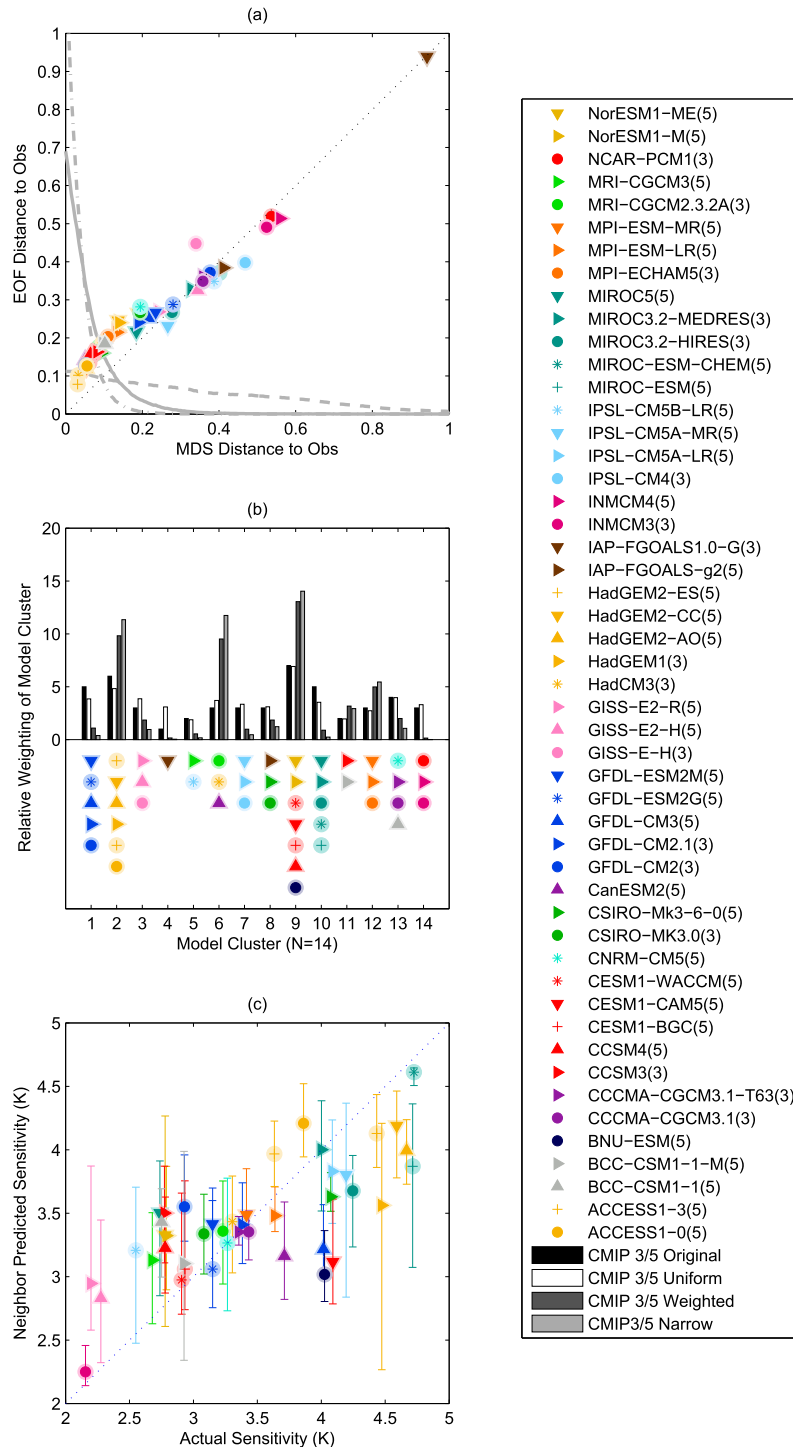


FIG. 5. (a) Comparison of model to observation Euclidean distance in the original nine-dimensional EOF space and the equivalent distance in the two-dimensional MDS space. (b) A  $k$ -means cluster analysis using climatological EOF loadings to group CMIP5 models into 14 clusters. Vertical bars indicate the relative weighting associated with each cluster in the original ensemble, and using different resampling priors. (c) A “leave one out” plot where a model is excluded and the interpolation process used in section 2c is used to predict the model’s climate sensitivity; vertical bars show the range of predicted values over 20 iterations.



value; this can be understood by considering that although CESM-CAM5 has a relatively similar climate to its predecessors (in the context of CMIP variability, at least, shown in Fig. 2), it also has a considerably greater climate sensitivity. Therefore, the “best guess” that the model would have a similar value of  $S$  to its forebears is in this case incorrect.

We would argue, however, that the strength of this technique does not ultimately hinge on its ability to predict an out of sample case, because we do not claim that the exact value of the interpolant at the point in the MDS space associated with the observations is our best estimate of the true value of  $S$ . In the absence of any additional information, we propose that the space in between models can be interpolated with a smooth surface and we can then sample that surface using the prior of our choosing, rather than simply accepting the prior governed by CMIP’s sample of opportunity. Unless two models are positioned at identical positions in the MDS space with different values of  $S$ , an interpolant can be found, irrespective of whether a global relationship between  $S$  and observable quantities exist (even if two models are at identical positions, it can be remedied by including additional diagnostics where the models do differ into the distance metric). Moreover, if there is such a global relationship between the diagnostics that make our distance metric and  $S$ , then this method will utilize it (with an appropriate prior) by excluding regions with a large bias and their associated unlikely values of  $S$ . The resulting distribution still remains a reallocation of weight between values of  $S$  from the original ensemble and should not therefore be interpreted as a PDF for  $S$  because the method is not informative values of  $S$  beyond the original ensemble range, nor can common systematic error (such as overly coarse resolution or common missing processes) or parametric uncertainties (i.e., a consideration of perturbed versions of each of the GCMs) influence the result.

#### d. Multivariate projections

Our analysis can be extended to multivariate projections by creating an interpolated surface for each unknown variable in turn. The space can then be jointly sampled for variables of interest. For CMIP5 simulations, the temperature and precipitation changes were taken as the difference between the 30-yr annual mean values in 2070–2100 in a single realization of RCP8.5 and the values in the historical simulation years 1970–2000. For CMIP3, the future values were taken from years 2070–2100 in the A1B scenario but temperature and precipitation changes were scaled by the ratio of median warming between 1980–2000 and 2090–2100 for RCP8.5

(4.6 K) and SRES A1B (3.4 K) in Rogelj et al. (2012). This is clearly an approximation, the A1FI scenario would be a much closer match to RCP8.5 in terms of net radiative forcing, but was not available for most of the models in the CMIP3 archive.

For each local projection (in this case, precipitation change or temperature change in 2100), a surface is constructed in the  $p =$  two dimensional space as in section 2c. We follow the “weighted” method, creating a  $10^6$  member truth-centered ensemble in the space and interpolating temperature and precipitation changes for each member. Thus, a joint distribution for temperature and precipitation change is constructed. Again, the process is repeated for  $N = 20$  random starting conditions, with the plots in Fig. 6 showing the summation of the results.

In Fig. 6, we show a demonstration of the multivariate capability with temperature and precipitation projections for 2100 under RCP8.5 for a number of different regions. In cases where there exists a strong correlation between temperature and precipitation change in the original ensemble [as in the Arctic (Fig. 6a) or Amazonia (Fig. 6c)], this correlation is preserved in the resampled distribution.

Note, as for the univariate case, these distributions should not be interpreted as PDFs, but rather as resampled histograms of possible model behavior. A full probabilistic treatment would need to additionally account for the impact of common error present in all models (i.e., limited resolution or missing processes) as well as terms accounting for uncertainty in individual model projections arising from physical parameter uncertainty (Stainforth et al. 2005) and initial condition uncertainty (Deser et al. 2012).

#### e. Sensitivity studies

##### 1) EOFs AND TRUNCATION CHOICES

Some subjective decisions are required in the interpretation and subsequent usage of the SVD conducted in section 2b(1), and we discuss these at greater length here. In previous studies like Masson and Knutti (2011), the intermodel distances were calculated without the SVD stage, simply calculating distances in the space defined by the anomaly matrix,  $\mathbf{X}$ . For the purposes of this study it is necessary to decrease the dimensionality (and interdependence) of the data in order to establish prior expectations of near-neighbor distances.

Euclidean distances are measured in the  $\mathbf{U}$  loading matrix, and each mode in that matrix has the same variance over the ensemble. This means that a single set of covarying diagnostics in  $\mathbf{X}$  will not dominate our distance metric (as all those fields would be represented

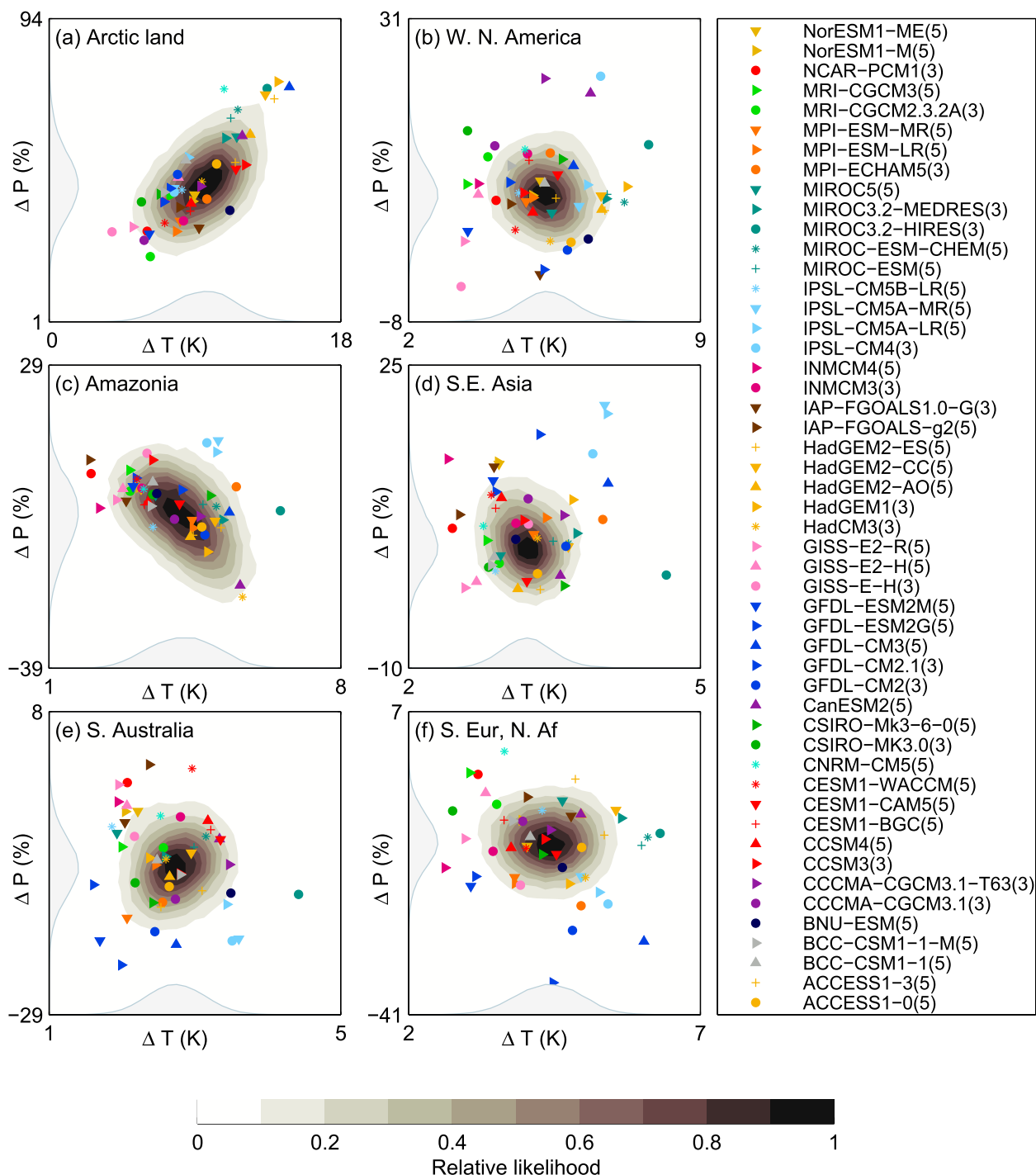


FIG. 6. Regional projections of annual mean temperature and precipitation changes for RCP8.5 (years 2070–2100) and recent climatology (years 1970–2000). For CMIP5 simulations, direct output from the RCP8.5 simulation is used. For CMIP3 simulations, changes are taken from the A1B simulation and scaled by the ratio of forcing in RCP8.5 to A1B. Each point represents a single climate model projection for the respective region (as defined in Fig. 3). The curve on each axis represents the univariate likelihood distribution for temperature and precipitation change independently, whereas the shaded contours indicate joint likelihood derived from a Gaussian prior.

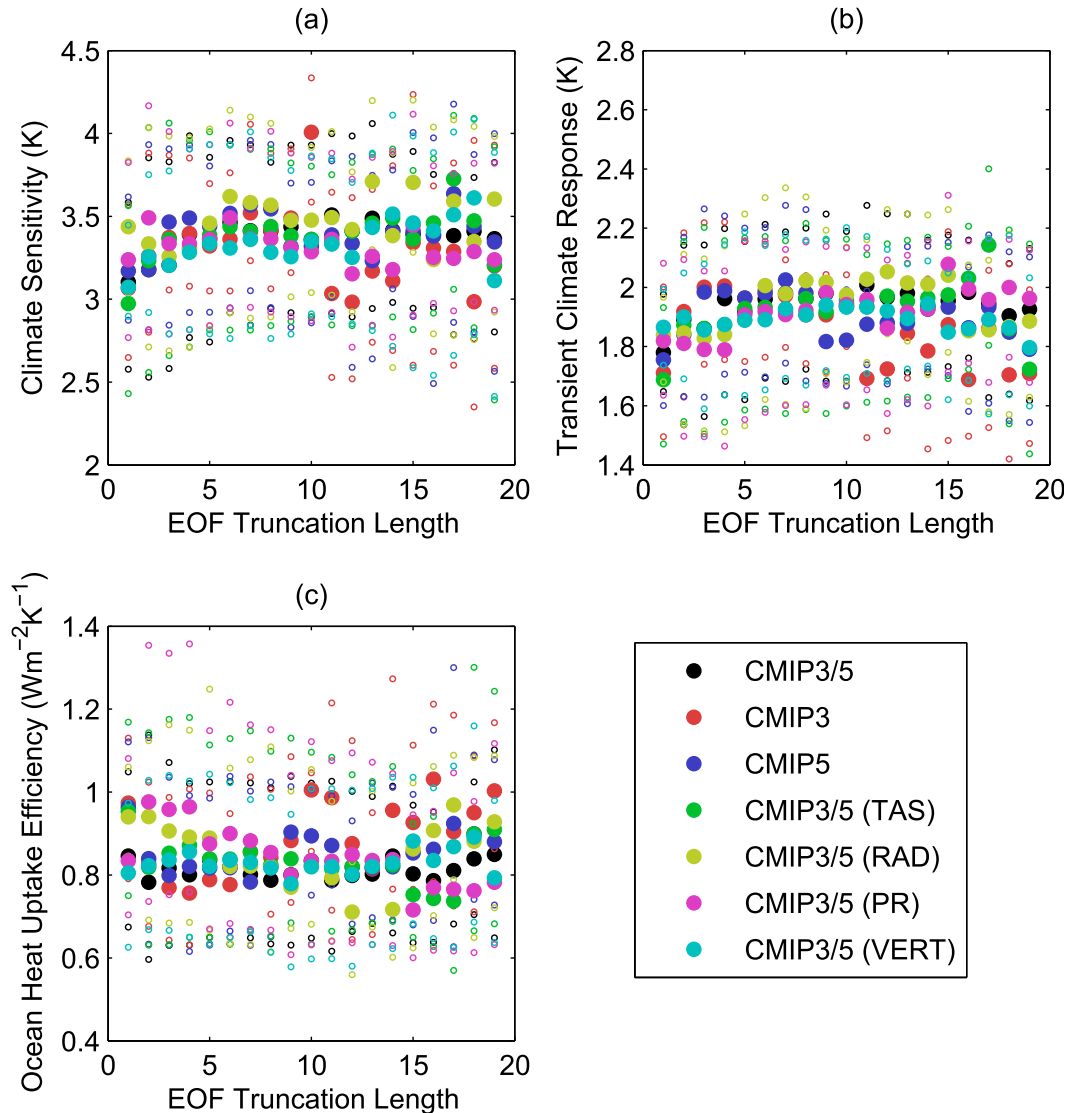


FIG. 7. Figure illustrating the sensitivity of the box-and-whisker plots shown in Fig. 4 to EOF truncation length. Filled circles show the median of the distribution for (a) climate sensitivity, (b) transient climate response, and (c) ocean heat uptake efficiency while the unfilled circles show the 10th and 90th percentiles of the distribution for EOF truncations from 1 to 20. Colors, as illustrated in the legend, show the distributions using different ensemble subsets (CMIP3, CMIP5, or the two combined) or different variable subsets (surface temperature, top of atmosphere radiative fluxes, or precipitation, all for the combined ensembles).

in a single mode). However, this also means that higher, noisy modes will have a significant influence on our distance metric unless we truncate the basis set at the appropriate point. In Fig. 7, we present results for a range of possible truncation lengths (note that the basis set for the CMIP5 and CMIP3 only case is necessarily different from the combined CMIP3/CMIP5 case because only a single ensemble has been used).

The analysis produces very similar distributions for climate sensitivity, transient climate response, and ocean heat uptake efficiency for values of  $t$  between 5

and 12 (see Fig. 7). At these truncation values, the results are not highly sensitive to the choice of variables included in the analysis or to the choice of whether CMIP3, CMIP5, or both ensembles are considered.

For values of  $t$  of less than 5 and greater than 13, we see a larger dependency on the choice of ensemble and variable. When  $t < 5$ , 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). For values of  $t$  greater than 13, the intermodel distance matrix becomes increasingly less well correlated with the absolute distance matrix as the higher modes reflect only subtle differences between models, effectively adding noise to the intermodel distance matrix. We thus opt for a truncation to nine modes, which we defend primarily because this choice produces distributions that are robust to variable and model choices (Fig. 7).

## 2) CHOICE OF OBSERVATIONAL CONSTRAINT

In section 3e(1), we show how the intermodel distance matrix is influenced by choices of EOF truncation and weighting. Here, we continue the calculation to test the impacts of using different observations to constrain  $S$  and TCR. The box-and-whisker plots in Figs. 4b–d show the implications of using different combinations of observations to form the intermodel distance matrix for the combined CMIP3/CMIP5 ensemble. The ALL case uses a multivariate EOF constructed as before, TAS constructs the EOFs using gridded surface temperature fields only, RAD uses only gridded top-of-atmosphere shortwave and longwave radiation fields, PR uses gridded total precipitation, and QT uses only zonally averaged temperature and specific humidity on model levels.

In each case, the analysis is repeated as before, with a Gaussian prior for the interpolated models with a variance equal to that of the original ensemble. It is found that the upper bound for both  $S$  and TCR is relatively insensitive to the choice of variable used to constrain the models (the 85th percentile is between 4.0 and 4.1 K in all cases). However, the lower bound on  $S$  is more sensitive, with the highest lower bound occurring in the ALL case (2.9 K for ALL case, down to 2.6 K for the PR case). The implication of this is that some of the lower-sensitivity metamodels can be eliminated using some diagnostics, but not with others.

## 4. Discussion

Our method uses projections from nonindependent climate model simulations to form a continuous likelihood distribution by forming a space using observable diagnostics, and interpolating unknown information throughout this space. Model quality information can be incorporated by sampling the space more densely in the region close to the observational projection. The resulting projections (or distributions of projected unknown climate parameters) are largely insensitive to the addition or removal of similar or identical models, or to the addition of models with a strong climatological bias.

As in Masson and Knutti (2011) and Knutti et al. (2010b), we find a strong level of self-similarity between

models from the same institution, which in almost all cases lie significantly closer to each other than they are to other models in the ensemble. This remains true for models that have changed a large portion of code between releases. This observation raises the question of why particular model biases can outlive an almost complete change in codebase. Various plausible explanations can be proposed; first, a sociological component to the tuning process is likely to exist, as model developers tend to span multiple model versions and it is likely that they each choose a preferred set of metrics and datasets with which to tune or evaluate their model. Systematic studies have found that differently tuned, and yet equally plausible, versions of a GCM can be produced by varying these priorities (Mauritsen et al. 2012). This choice of metrics, together with the methodologies used to tune parameters and choice of parameters themselves, may imprint a “developer’s fingerprint” upon models from a certain institution that persists for multiple development cycles. Also, in many GCMs, the different components of the coupled system are not all updated at the same time. The staggered tuning of the coupled system could cause a “memory” of the biases of previous models into the newer models.

Our method constructs a surface using observable diagnostics, in which both the models and observations can be plotted. This surface allows for the interpolation of unknown quantities, such as climate sensitivity, so that with appropriate prior centered on observations, a distribution for unknown quantities can be formed by interpolating between known values from the multi-model archive. These distributions cannot be wider than the original ensemble distribution for  $S$  and posterior likelihoods are judged only on the relative quality of the models’ simulation for the present day, but other metrics of performance could be included. In effect, this process resamples the range of model behavior present in the original ensemble, but in a fashion that greatly reduces the bias arising from highly interdependent ensemble members and rewarding more plausibility to models with a better mean climatology.

We demonstrate the method by producing CDFs for a number of unknown climate variables. The method allows us to construct idealized resampled ensembles and to study the effect of that resampling on variables of interest. For example, we find that, based on mean climatology at least, that a value of climate sensitivity of less than 3 K becomes significantly less likely when a narrow prior centered on observations is chosen. However, the results presented in this study used various aspects of the mean climatology as the constraint on unknown climate variables. Space limitations prevent us from investigating what constraints might emerge from a

consideration of transient skill metrics [such as model expressions of spatial fingerprints corresponding to greenhouse gas forcing as in Hegerl et al. (2000)] in the framework presented here, so we leave a comprehensive analysis of different metrics and their potential to constrain unknown climate variables to a further study.

Knutti et al. (2010b) suggested that ensemble-wide correlations between observable quantities and climate sensitivity are rare in the CMIP3 ensemble, although the results of Fasullo and Trenberth (2012) seem to provide at least one counterexample. Caution must be used, however, in inferring significance in correlation alone, given that the effective number of degrees of freedom assessed from intermodel variability is significantly smaller than the number of models in the ensemble (Annun and Hargreaves 2011).

Our results are subject to a number of caveats. At present, all model errors are considered only in a relative sense without addressing common systematic model errors in the ensemble, missing processes or feedbacks and uncertainty estimates for the interpolation process itself. We thus interpret the resulting PDFs as a lower bound on the systematic and parametric uncertainty. Clearly, the method can provide no information on the response of models that are not sampled in the ensemble, but it can reduce the bias arising from model interdependency. One can consider the distributions as a resampling of  $S$  in the multimodel archive in such a way that models with large interdependencies are not over-represented and the region of the space corresponding to more plausible climatologies is more densely sampled. The resulting distributions are based on interpolation only, and thus have a value of zero outside of the original ensemble range of  $S$ . Also, if all models are biased low or high in their prediction due to a feedback that does not exist in the current climate, such a re-weighting cannot correct for that bias.

The cases where the interpolated models are sampled with a Gaussian prior centered on the observational projection are subject to a number of caveats. Our resampling method, as presented, is conditional on the original ensemble variance, which is itself related to model replication in the ensemble, so although the weight of replicated models is reduced through our method, replication can still potentially influence the resulting distribution indirectly through the means of the ensemble variance. This issue could be addressed in future study with an effort to define a prior independently of the ensemble itself.

Our use of a single observed point as the center of the Gaussian prior implies that we consider that the internal variability is not a significant factor in the context of model bias. We defend this decision by showing that the

model spread arising from an ensemble of initial condition simulations from a single model has a significantly lower variance in the MDS space than the CMIP ensemble. However, this justification is subject to some caveats, first that the single model (CCSM4) representation of variance is representative of the real world, critically that the model's variability is not significantly underdispersive. This is notably not the case when considering decadal trends, as shown in Deser et al. (2012) where an initial condition ensemble from the same model, CCSM4 produces a comparable spread to the CMIP5 archive for near term projections. But, when considering the mean state bias as we do here to construct our state vector, we show that model-generated internal variability is negligible and therefore remain confident in our decision to consider the observational projection (and each model projection) as individual points in the MDS space.

In addition to this assumption, there is clearly also an issue of the degree to which the observational products represent reality. To thoroughly explore these assumptions would require a detailed assessment of each observational dataset used for surface temperature, precipitation, top-of-atmosphere fluxes, etc., and is clearly beyond the scope of this study. Hence, the results of the study carry the caveat that the bias is measured relative to the observational products used, and biases in these products would thus bias our results also.

By taking the model–observation distance matrix as a measure of model skill and interdependency, we (and most other studies on the matter) are also assuming that each of the models has been optimally tuned to match the observations. We effectively assume that the bias that each model exhibits is therefore irreducible by further tuning, and therefore is representative of the accuracy of the model's representation of climate processes. We also assume, as in Masson and Knutti (2011) and Bishop and Abramowitz (2013), that correlation of model errors represents a measure of model interdependency. Both of these assumptions may be debated by considering that different groups may use different observational targets (either by using different products entirely or by concentrating on skill in specific regions). To disentangle these systematic and parametric sources of model bias and similarity, one would require a coordinated superensemble of perturbed GCM simulations from a number of institutions, a currently nonexistent resource that we would strongly argue would be a great asset to the community.

Multivariate applications of the method are demonstrated with joint temperature/precipitation projections. Future work will examine in more detail the behavior of such distributions, especially the potential



for multiregion distributions. Before such an effort would be meaningful for any given regional projection, care must be taken to ensure that the metrics considered are relevant to the processes in question. The “one size fits all” global metrics considered in this study are used to illustrate the concept here but might not be applicable to regionally specific problems.

Placing this work in the context of established literature is difficult. Clearly, the use of observations to constrain more likely ensemble members draws parallels with other multimodel studies that have attempted to find global relationships in the ensemble (Hall and Qu 2006; Fasullo and Trenberth 2012) or with perturbed physics studies that constrain simulations by linear regression (Piani et al. 2005) or by more complex transfer functions (Knutti et al. 2006). However, in the light of recent works that highlight the complications of employing global correlations between observables and responses in the multimodel ensemble (Knutti et al. 2010b), we explicitly do not require them to exist in general, allowing the response surface to vary continuously across the observable space.

For the wider question of ensemble interpretation, our approach creates the potential for the researcher to create an interpolated ensemble whose emulated climate distribution is centered on observations by construction. Hence, the resulting distributions for future climate change are weighted toward interpolated models with the least mean-state bias for present-day simulations. However, those projections themselves have the potential to be quite diverse if they are not strongly constrained by the recent historical mean state bias. In the absence of additional observational (or physical) constraints on the future climate response, we believe this is an appropriate representation of the uncertainty in the projections themselves. In addition, the concept of “model democracy” in the original CMIP archives is a fallacy given the results of models that are highly replicated can potentially carry too much weight. Our interpolated ensemble can be seen as an attempt to restore real model democracy among models, at least in as much as they can be distinguished by their simulated climatological output.

## 5. Conclusions

We propose a novel method for combining results from an “ensemble of opportunity” such as CMIP3 or CMIP5, where the ensemble distribution allows for significant interdependencies between members and the potential for models of varying performance and complexity. The method employs a pairwise distance metric between ensemble members and observations, which

provides information both on model similarity and climatological bias. A multidimensional scaling approach allows the intermodel distances to be represented on a low-dimensional surface that can then be used to interpolate unknown climate parameters with a prior distribution of the researcher’s choice, effectively allowing the potential for the ensemble of opportunity to be resampled in a coherent fashion.

The technique has been demonstrated on the CMIP3 and CMIP5 ensembles, where we use a multivariate climatological metric to evaluate intermodel distances, and produce resampled distributions for a number of univariate and multivariate model outputs. The sampling distributions used here are presented as sensitivity studies; a uniformly resampled distribution reduces the weight allocated to highly replicated models (as compared to the original ensemble) and a resampled distribution centered on observed climatology but with variance equal to that of the original ensemble results in a down-weighting of CMIP3 models with very low climate sensitivity and poor climatological simulations, which subsequently is reflected in an increase in the minimum expected value of climate sensitivity for the metrics considered.

Although we find that the CMIP5 and CMIP3 ensembles are not heavily biased toward a particular result, this is largely fortuitous. We propose that with the increasing availability of multiple model versions, perturbed physics ensembles, and sharing of model code, the use of only “model democracy” becomes increasingly hard to justify. New methods that account for model dependence and model performance such as the approach presented here are required. However, defining appropriate metrics to evaluate models appropriateness for future climate projections remains a formidable challenge.

*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 Intercomparison 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), U.S. Department of Energy, Cooperative Agreement DE-FC02-97ER62402. AIRS data 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).

## 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**, 133–136, doi:[10.2151/sola.2009-034](https://doi.org/10.2151/sola.2009-034).
- Adler, R., and Coauthors, 2003: The version 2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–present). *J. Hydrometeorol.*, **4**, 1147–1167, doi:[10.1175/1525-7541\(2003\)004<1147:TVGPCP>2.0.CO;2](https://doi.org/10.1175/1525-7541(2003)004<1147:TVGPCP>2.0.CO;2).
- Annan, J., and J. Hargreaves, 2010: Reliability of the CMIP3 ensemble. *Geophys. Res. Lett.*, **37**, L02703, doi:[10.1029/2009GL041994](https://doi.org/10.1029/2009GL041994).
- , and —, 2011: Understanding the CMIP3 multimodel ensemble. *J. Climate*, **24**, 4529–4538, doi:[10.1175/2011JCLI3873.1](https://doi.org/10.1175/2011JCLI3873.1).
- Aumann, H. H., and Coauthors, 2003: AIRS/AMSU/HSB on the Aqua mission: Design, science objectives, data products, and processing systems. *IEEE Trans. Geosci. Remote Sens.*, **41**, 253–264, doi:[10.1109/TGRS.2002.808356](https://doi.org/10.1109/TGRS.2002.808356).
- Bishop, C. H., and G. Abramowitz, 2013: Climate model dependence and the replicate earth paradigm. *Climate Dyn.*, **41**, 885–900, doi:[10.1007/s00382-012-1610-y](https://doi.org/10.1007/s00382-012-1610-y).
- Borg, I., and P. J. F. Groenen, 1997: *Modern Multidimensional Scaling: Theory and Applications*. Springer, 471 pp.
- 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**, D12106, doi:[10.1029/2005JD006548](https://doi.org/10.1029/2005JD006548).
- 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. *Geophys. Res. Lett.*, **41**, 1803–1808, doi:[10.1002/2014GL059205](https://doi.org/10.1002/2014GL059205).
- Delaunay, B., 1934: Sur la sphere vide. *Izv. Akad. Nauk SSSR, Otd. Mat. Estestv. Nauk*, **7**, 793–800.
- de Leeuw, J. D., 1977: Applications of convex analysis to multidimensional scaling. *Recent Developments in Statistics*, J. R. Barra et al., Eds., North Holland Publishing Company, 133–146.
- DelSole, T., and J. Shukla, 2009: Artificial skill due to predictor screening. *J. Climate*, **22**, 331–345, doi:[10.1175/2008JCLI2414.1](https://doi.org/10.1175/2008JCLI2414.1).
- Deser, C., A. Phillips, V. Bourdette, and H. Teng, 2012: Uncertainty in climate change projections: The role of internal variability. *Climate Dyn.*, **38**, 527–546, doi:[10.1007/s00382-010-0977-x](https://doi.org/10.1007/s00382-010-0977-x).
- Fasullo, J. T., and K. E. Trenberth, 2012: A less cloudy future: The role of subtropical subsidence in climate sensitivity. *Science*, **338**, 792–794, doi:[10.1126/science.1227465](https://doi.org/10.1126/science.1227465).
- Furrer, R., S. Sain, D. Nychka, and G. Meehl, 2007: Multivariate Bayesian analysis of atmosphere–ocean general circulation models. *Environ. Ecol. Stat.*, **14**, 249–266, doi:[10.1007/s10651-007-0018-z](https://doi.org/10.1007/s10651-007-0018-z).
- Gleckler, P., K. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.*, **113**, D06104, doi:[10.1029/2007JD008972](https://doi.org/10.1029/2007JD008972).
- Greene, A., L. Goddard, and U. Lall, 2006: Probabilistic multimodel regional temperature change projections. *J. Climate*, **19**, 4326–4343, doi:[10.1175/JCLI3864.1](https://doi.org/10.1175/JCLI3864.1).
- Hall, A., and X. Qu, 2006: Using the current seasonal cycle to constrain snow albedo feedback in future climate change. *Geophys. Res. Lett.*, **33**, L03502, doi:[10.1029/2005GL025127](https://doi.org/10.1029/2005GL025127).
- Haughton, N., G. Abramowitz, A. Pitman, and S. J. Phipps, 2014: On the generation of climate model ensembles. *Climate Dyn.*, **43**, 2297–2308, doi:[10.1007/s00382-014-2054-3](https://doi.org/10.1007/s00382-014-2054-3).
- Hegerl, G., P. Stott, M. Allen, J. Mitchell, S. Tett, and U. Cubasch, 2000: Optimal detection and attribution of climate change: Sensitivity of results to climate model differences. *Climate Dyn.*, **16**, 737–754, doi:[10.1007/s003820000071](https://doi.org/10.1007/s003820000071).
- Huber, M., I. Mahlstein, M. Wild, J. Fasullo, and R. Knutti, 2011: Constraints on climate sensitivity from radiation patterns in climate models. *J. Climate*, **24**, 1034–1052, doi:[10.1175/2010JCLI3403.1](https://doi.org/10.1175/2010JCLI3403.1).
- Jun, M., R. Knutti, and D. W. Nychka, 2008: Local eigenvalue analysis of CMIP3 climate model errors. *Tellus*, **60A**, 992–1000, doi:[10.1111/j.1600-0870.2008.00356.x](https://doi.org/10.1111/j.1600-0870.2008.00356.x).
- Knutti, R., 2008: Should we believe model predictions of future climate change? *Philos. Trans. Roy. Soc.*, **366A**, 4647–4664, doi:[10.1098/rsta.2008.0169](https://doi.org/10.1098/rsta.2008.0169).
- , 2010: The end of model democracy? *Climatic Change*, **102**, 395–404, doi:[10.1007/s10584-010-9800-2](https://doi.org/10.1007/s10584-010-9800-2).
- , G. A. Meehl, M. R. Allen, and D. A. Stainforth, 2006: Constraining climate sensitivity from the seasonal cycle in surface temperature. *J. Climate*, **19**, 4224–4233, doi:[10.1175/JCLI3865.1](https://doi.org/10.1175/JCLI3865.1).
- , G. Abramowitz, M. Collins, V. Eyring, P. J. Gleckler, B. Hewitson, and L. Mearns, 2010a: Good practice guidance paper on assessing and combining multi model climate projections. *Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections*, T. F. Stocker, et al., Eds., 1–13. [Available online at <http://www.proclim.ch/4dcgi/proclim/en/IPCC?1495>.]
- , R. Furrer, C. Tebaldi, J. Cermak, and G. Meehl, 2010b: Challenges in combining projections from multiple climate models. *J. Climate*, **23**, 2739–2758, doi:[10.1175/2009JCLI3361.1](https://doi.org/10.1175/2009JCLI3361.1).
- , D. Masson, and A. Gettelman, 2013: Climate model genealogy: Generation CMIP5 and how we got there. *Geophys. Res. Lett.*, **40**, 1194–1199, doi:[10.1002/grl.50256](https://doi.org/10.1002/grl.50256).
- Masson, D., and R. Knutti, 2011: Climate model genealogy. *Geophys. Res. Lett.*, **38**, L08703, doi:[10.1029/2011GL046864](https://doi.org/10.1029/2011GL046864).
- , and —, 2013: Predictor screening, calibration, and observational constraints in climate model ensembles: An illustration using climate sensitivity. *J. Climate*, **26**, 887–898, doi:[10.1175/JCLI-D-11-00540.1](https://doi.org/10.1175/JCLI-D-11-00540.1).
- Mauritsen, T., and Coauthors, 2012: Tuning the climate of a global model. *J. Adv. Model. Earth Syst.*, **4**, M00A01, doi:[10.1029/2012MS000154](https://doi.org/10.1029/2012MS000154).
- NASA, 2011: CERES EBAF datasets, Langley Research Center. [Available online at <http://ceres.larc.nasa.gov/products.php?product=EBAF-TOA>.]
- Pennell, C., and T. Reichler, 2011: On the effective number of climate models. *J. Climate*, **24**, 2358–2367, doi:[10.1175/2010JCLI3814.1](https://doi.org/10.1175/2010JCLI3814.1).
- Piani, C., D. J. Frame, D. A. Stainforth, and M. R. Allen, 2005: Constraints on climate change from a multi-thousand member ensemble of simulations. *Geophys. Res. Lett.*, **32**, L23825, doi:[10.1029/2005GL024452](https://doi.org/10.1029/2005GL024452).
- Reichler, T., and J. Kim, 2008: How well do coupled models simulate today's climate? *Bull. Amer. Meteor. Soc.*, **89**, 303–311, doi:[10.1175/BAMS-89-3-303](https://doi.org/10.1175/BAMS-89-3-303).
- Rogelj, J., M. Meinshausen, and R. Knutti, 2012: Global warming under old and new scenarios using IPCC climate sensitivity range estimates. *Nat. Climate Change*, **2**, 248–253, doi:[10.1038/nclimate1385](https://doi.org/10.1038/nclimate1385).
- Rougier, J., M. Goldstein, and L. House, 2013: Second-order exchangeability analysis for multimodel ensembles. *J. Amer. Stat. Assoc.*, **108**, 852–863, doi:[10.1080/01621459.2013.802963](https://doi.org/10.1080/01621459.2013.802963).

- Sanderson, B. M., and R. Knutti, 2012: On the interpretation of constrained climate model ensembles. *Geophys. Res. Lett.*, **39**, L16708, doi:[10.1029/2012GL052665](https://doi.org/10.1029/2012GL052665).
- , —, and P. Caldwell, 2015: A representative democracy to address interdependency in a multi-model ensemble. *J. Climate*, doi:[10.1175/JCLI-D-14-00362.1](https://doi.org/10.1175/JCLI-D-14-00362.1), in press.
- Sherwood, S. C., S. Bony, and J.-L. Dufresne, 2014: Spread in model climate sensitivity traced to atmospheric convective mixing. *Nature*, **505**, 37–42, doi:[10.1038/nature12829](https://doi.org/10.1038/nature12829).
- Sibson, R., 1981: A brief description of natural neighbour interpolation. *Interpreting Multivariate Data*, V. Barnett, Ed., John Wiley, 21–36.
- Stainforth, D. A., and Coauthors, 2005: Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, **433**, 403–406, doi:[10.1038/nature03301](https://doi.org/10.1038/nature03301).
- Tebaldi, C., and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Philos. Trans. Roy. Soc.*, **365A**, 2053–2075, doi:[10.1098/rsta.2007.2076](https://doi.org/10.1098/rsta.2007.2076).
- , and B. Sansó, 2009: Joint projections of temperature and precipitation change from multiple climate models: A hierarchical Bayesian approach. *J. Roy. Stat. Soc.*, **172A**, 83–106, doi:[10.1111/j.1467-985X.2008.00545.x](https://doi.org/10.1111/j.1467-985X.2008.00545.x).