

Clim Dyn (2013) 41:2745–2763
DOI 10.1007/s00382-013-1733-9

Reliability and importance of structural diversity of climate model ensembles

Tokuta Yokohata · James D. Annan · Matthew Collins · Charles S. Jackson · Hideo Shiogama · Masahiro Watanabe · Seita Emori · Masakazu Yoshimori · Manabu Abe · Mark J. Webb · Julia C. Hargreaves

Received: 26 July 2012 / Accepted: 8 March 2013 / Published online: 4 April 2013
© The Author(s) 2013. This article is published with open access at Springerlink.com

Abstract We investigate the performance of the newest generation multi-model ensemble (MME) from the Coupled Model Intercomparison Project (CMIP5). We compare the ensemble to the previous generation models (CMIP3) as well as several single model ensembles (SMEs), which are constructed by varying components of single models. These SMEs range from ensembles where parameter uncertainties are sampled (perturbed physics ensembles) through to an ensemble where a number of the physical schemes are switched (multi-physics ensemble). We focus on assessing reliability against present-day climatology with rank histograms, but also investigate the

effective degrees of freedom (EDoF) of the fields of variables which makes the statistical test of reliability more rigorous, and consider the distances between the observation and ensemble members. We find that the features of the CMIP5 rank histograms, of general reliability on broad scales, are consistent with those of CMIP3, suggesting a similar level of performance for present-day climatology. The spread of MMEs tends towards being “over-dispersed” rather than “under-dispersed”. In general, the SMEs examined tend towards insufficient dispersion and the rank histogram analysis identifies them as being statistically distinguishable from many of the observations. The EDoFs of the MMEs are generally greater than those of SMEs, suggesting that structural changes lead to a characteristically richer range of model behaviours than is obtained with parametric/physical-scheme-switching ensembles. For distance measures, the observations and models ensemble members are similarly spaced from each other for MMEs, whereas for the SMEs, the observations are generally well outside the ensemble. We suggest that multi-model ensembles should represent an important component of uncertainty analysis.

T. Yokohata (✉) · H. Shiogama · S. Emori · M. Abe
National Institute for Environmental Studies, Center for Global Environmental Research, 16-2 Onogawa, Tsukuba, Ibaraki 305-8506, Japan
e-mail: yokohata@nies.go.jp

J. D. Annan · J. C. Hargreaves
Research Institute for Global Change, Japan Agency for Marine-Earth Science and Technology, 3173-25 Showamachi, Kanazawa-ku, Yokohama, Kanagawa 236-0001, Japan

M. Collins
College of Engineering, Mathematics and Physical Sciences, University of Exeter, Harrison Building, North Park Road, Exeter EX4 4QF, UK

C. S. Jackson
Institute of Geophysics, The University of Texas at Austin, 10100 Burnet Rd., ROC-196, Mail Code R2200, Austin, TX 78758, USA

M. Watanabe · M. Yoshimori
Atmosphere and Ocean Research Institute, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8568, Japan

M. J. Webb
Met Office, Hadley Centre, FitzRoy Road, Exeter EX1 3PB, UK

Keywords Climate model · Multi-model ensembles · Reliability · Rank histogram · Degree of freedom · Perturbed physics ensembles

1 Introduction

Due to our lack of understanding of the climate system and limitations of computational power, climate models are far from perfect. The different models do, however, span a considerable range of output which leads to the possibility of making probabilistic predictions of the future based on

the models (Collins et al. 2012). How best to integrate ensembles of models into a probabilistic calculation is still a matter of debate. For example, one approach to generating probabilistic future predictions is to implement a weighting procedure based on the performance of the present day climate simulation (e.g., Sexton et al. 2012). One of the prerequisites for implementation of such a method is that the ensemble employed should initially be broad enough to include the truth. Understanding the characteristics of the ensembles that have already been generated is an important step in this process. Here we build on earlier work investigating the reliability of climate model ensembles (e.g., Annan and Hargreaves 2010, hereafter AH10, Yokohata et al. 2012, hereafter Y12). The multi-model ensembles (MMEs) are made up of output from common experiments run by the world's modelling centres. These models vary in construction and contain different parameterisations of climate processes, and different methods for the numerical integration (different grids, numerical schemes etc.). No one model is better than all the others in all aspects (e.g., Gleckler et al. 2008). As such, we may consider the MME as sampling at least some of our uncertainties in how a climate model should be constructed. One such MME is the Climate Model Inter-comparison Project phase three (CMIP3, Meehl et al. 2007) which contributed to the fourth assessment report of Intergovernmental Panel on Climate Change. Subsequently a new phase of CMIP (CMIP5, Taylor et al. 2012) has been started. This MME contains more models and new, hopefully improved, model versions of the older models, some with increased resolution and complexity (i.e., with additional feedbacks being prognostically modelled). The number of structurally distinct ensemble members (i.e., excluding initial condition ensembles) is increased in CMIP5 (Taylor et al. 2012), which should enable more robust conclusions to be drawn about the ensemble characteristics.

In addition to the MMEs, some modelling centres have, over the last decade, developed ensembles based on a single model (single model ensembles, SMEs). One kind of SMEs is a “perturbed physics” ensemble (PPE) in which uncertainties in model parameters are sampled (Murphy et al. 2004; Stainforth et al. 2005; Collins et al. 2006a; Webb et al. 2006; Annan et al. (2005a, b); Jackson et al. 2008; Sanderson (2011); Yokohata et al. 2010). Some new PPEs based on the newly developed models contributing to the CMIP5 have recently been generated (Shiogama et al. 2012; Klocke et al. 2011). The first SMEs merely varied the values of parameters (which are just single numbers in the model code), but recently, researchers have started to create ensembles with larger differences by switching between different sets of the physical schemes. An ensemble created in this way has been termed a “multi-

physics” ensemble (MPE) (Watanabe et al. 2012; Gettelman et al. 2012).

Here we investigate the reliability of the new CMIP5 ensemble and compared it to previous ensembles, both MMEs and SMEs. We use the rank histogram approach (AH10, Y12) which is often used in the field of numerical weather prediction (Jolliffe and Primo 2008, hereafter JP08). In previous work using these statistical tests (AH10, Y12 and Hargreaves et al. 2011), we were unable to reject the hypothesis of reliability for the CMIP3 MME for either modern climate or the climate change of the Last Glacial Maximum. This gives us some confidence in the CMIP3 ensemble. Conversely it was found that the SMEs were generally less reliable (Y12, Hargreaves et al. 2011), although it should be noted that no MPEs were analysed in those studies.

The methods for assessing reliability used in these previous analyses have some limitations. First, the statistical test of reliability depends on the “independent number of observation” as discussed in JP08, but that number was assumed rather than calculated in the previous work. In AH10 and Y12, climatological mean fields of observation are compared with those of model ensemble members at each grid point. Since the neighboring grid points are not necessarily independent, it is not easy to know the independent number in the fields which corresponds to the “effective degree of freedom” (EDoF). If the EDoF increases, the statistical test for the reliability becomes stricter (JP08).

Second, in the rank histogram analysis presented in Y12, the number of bins in the rank histogram (which should naturally be the number of ensemble member plus one) was reduced to 11 throughout, for consistency with the number of ensemble members in the CMIP3 ensemble. This may reduce the power of the test if the rebinning smooths the histogram of the larger ensemble. In addition, the rank histogram of each climate variable is investigated separately in Y12, but the overall characteristics of climate model ensembles may be investigated if we create multivariate rank histograms.

Third, the rank histogram does not provide information on the magnitude of model errors. In terms of model error, Y12 investigated only the relationship between the errors of ensemble mean and standard deviation of model ensemble members.

In this work, we address these issues, calculating the EDoF (using the formulation by Bretherton et al. 1999 as in Annan and Hargreaves 2011), exploring the effect of increasing the number of bins in the rank histogram, and calculating multivariate rank histograms. In addition to the rank histogram we explore other ways of evaluating the ensemble, analysing the distances between models and observational data by calculating the minimum spanning trees (e.g., Wilks 2004) and the

average of the distances between the observation and the models for all the ensembles.

In Sect. 2, the model ensembles of MMEs and SMEs and the methods of analysis are presented. The analysis methods include the explanation of the calculation of rank histogram and the statistical test for the reliability (2–2), the formulation of EDoF (2–3), and the distances between observation and model ensemble members (2–4). Results and discussion are presented in Sect. 3 and summarised in Sect. 4.

2 Model ensembles and methods of analysis

2.1 Climate model ensembles

For the MMEs, both the CMIP5 (Taylor et al. 2012) and CMIP3 (Meehl et al. 2007) ensembles are used for the analysis. The CMIP5 dataset is obtained from the federated archives initiated under Earth System Grid project (<http://esg-pcmdi.llnl.gov/>) led by Program for Climate Model Diagnosis and Intercomparison (PCMDI) and being advanced through the Earth System Grid Federation (ESGF; http://esgf.org/wiki/ESGF_Overview; Williams et al. 2011), established under the Global Organization for Earth System Science Portals (GO-ESSP; <http://go-essp.gfdl.noaa.gov/>). We use the CMIP5 model output of the historical simulation of 28 atmosphere–ocean coupled models (CMIP5-AO) for which sufficient data was available in the archives. The models used in the analysis, are summarised in Table 1. We use only one run for each model listed in Table 1, so the number of ensemble members of CMIP5-AO is 28.

The CMIP3 dataset was obtained from the PCMDI archives (Meehl et al. 2007), and we use the output from the historical simulations by both the atmosphere–ocean coupled model (CMIP3-AO) and the atmosphere–slab ocean coupled model (CMIP3-AS). The CMIP3 models used for the analysis are the same as those in Y12, and the details are summarised therein. We use only one run for each model listed in Table 2, so the number of ensemble members in CMIP3-AO for which suitable outputs are available is 16, and that of CMIP3-AS is 10.

In the present study, we also create a CMIP5+CMIP3-AO ensemble, which simply combines CMIP5-AO and CMIP3-AO. The number of CMIP5+CMIP3-AO ensemble member is 44. In this combined ensemble, we make no adjustment or allowance for the possibility that some models may be particularly closely related to one another, for example consecutive generations from a single modelling centre. Such issues are of course a major topic, but this research focus is beyond the scope of this work (e.g., Masson and Knutti 2011).

We use six different SMEs based on structurally distinct models as summarised in Table 3. The PPEs created by

HadCM3 (Gordon et al. 2000), HadSM3 (Pope et al. 2000), CAM3.1 (Collins et al. 2006b), and MIROC3.2 (K-1 model developers 2004) are here called HadCM3-AO, HadSM3-AS, NCAR-A, MIROC3-AS, respectively. These four ensembles were also used in Y12. In addition, a new PPE from the MIROC5 atmosphere–ocean coupled model (Watanabe et al. 2010), and a new MPE created from a mixture of elements from the MIROC3.2 and MIROC5 atmosphere models are analysed. These new ensembles are hereafter called MIROC5-AO and MIROC-MPE-A.

HadCM3-AO and HadSM3-AS were created in the Quantifying Uncertainty in Modelling Predictions (QUMP) project. The atmospheric components of HadCM3 and HadSM3 are identical, and have resolution of 2.5 latitudinal degrees by 3.75 longitudinal degrees with 19 vertical levels. The ocean component of HadCM3 has a resolution of 1.25×1.25 degrees with 20 levels. In HadSM3, a motionless 50 m slab ocean is coupled to the atmospheric model and ocean heat transport is diagnosed for each member.

See Y12 and references therein for further details on the construction of HadSM-AS (Murphy et al. 2004; Webb et al. 2006), HadCM3-AO (e.g., Collins et al. 2010), NCAR-A (Jackson et al. 2004, 2008), and MIROC3-AS (Annan et al. 2005a, b; Yokohata et al. 2010) ensembles. Here we outline the main features of the construction of the two new SMEs, MIROC5-AO, and MIROC-MPE-A. These SMEs were constructed within the Japan Uncertainty Modelling Project (JUMP). For MIROC5-AO, Shiogama et al. (2012) devised a method to create an ensemble by atmosphere–ocean coupled model without flux correction. This ensemble is based on a new version of MIROC developed for the CMIP5 project, whose physical schemes are sophisticated and model performance are improved from the former version (Watanabe et al. 2010). The atmospheric component of MIROC5 used in this study has T42 (about 300 km grid) horizontal resolution, whereas the original version of MIROC5 has T85 (about 150 km grid) resolution, with 40 vertical levels. The ocean component model has approximately 1° horizontal resolution and 49 vertical levels with an additional bottom boundary layer. Using results from AGCM experiments, Shiogama et al. (2012) chose sets of parameter values for which the energy budget at the top of the atmosphere was predicted to be close to zero in order for these members not to have climate drift, and then ran AOGCM models with these parameter sets. The number of ensemble members in MIROC5-AO is 36.

Although the climate sensitivity of MIROC3.2 is relatively high compared to other CMIP3 models at 4.0 K (Yokohata et al. 2008), that of MIROC5 is substantially lower at 2.6 K (Watanabe et al. 2010). Since the differences in the response to CO₂ increase are caused by changing model physical schemes, Watanabe et al. (2012)

>Table 1 List of CMIP5 multi-model ensemble

Model Name	Institute	Reference
1. ACCESS1.0	Commonwealth Scientific and Industrial Research Organization (CSIRO) and Bureau of Meteorology (BOM), Australia	http://wiki.csiro.au/confluence/display/ACCESS/ACCESS+Publications
1. BCC-CSM1.1	Beijing Climate Center, China Meteorological Administration	http://www.cccma.ec.gc.ca/models
1. CanESM2	Canadian Centre for Climate Modelling and Analysis	Gent et al. (2011)
1. CCSM4	National Center for Atmospheric Research	http://www.cnr.meteo.fr/cmip5 —follow model description link
1. CNRM-CM5	Centre National de Recherches Meteorologiques/Centre Europeen de Recherche et Formation Avancees en Calcul Scientifique	
1. CSIRO-Mk3.6.0	Commonwealth Scientific and Industrial Research Organization in collaboration with Queensland Climate Change Centre of Excellence	Rotstayn et al. (2010), http://cmip-pcmdi.llnl.gov/cmip5/
1. FGOALS-s2	LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences	http://www.lasg.ac.cn/FGOALS/CMIP5
1. GFDL-CM3	NOAA Geophysical Fluid Dynamics Laboratory	http://nomads.gfdl.noaa.gov/
2. GFDL-ESM2G		
3. GFDL-ESM2M		
1. GISS-E2-H	NASA Goddard Institute for Space Studies	www.giss.nasa.gov/research/modeling
2. GISS-E2-R		
1. HadCM3	Met Office Hadley Centre	Collins et al. (2001)
2. HadGEM2-CC		Smith et al. (2007)
3. HadGEM2-ES		Smith et al. (2010)
		Jones et al. (2011)
		Martin et al. (2011)
		Collins et al. (2011)
		Bellouin et al. (2007)
		Collins et al. (2008)
		Johns et al. (2006)
		Martin et al. (2006)
		Ringer et al. (2006)
		Volodin et al. (2010)
		http://frcmc.ipsl.fr
1. INM-CM4	Institute for Numerical Mathematics	
1. IPSL-CM5A-LR	Institut Pierre-Simon Laplace	
2. IPSL-CM5A-MR		
3. IPSL-CM5B-LR		
1. MIROC4h	Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology	Sakamoto et al. (2012)
2. MIROC5		Tatebe et al. (2012)
3. MIROC-ESM		Watanabe et al. (2010)
4. MIROC-ESM-CHEM		Watanabe et al. (2011)
1. MPI-ESM-LR	Max Planck Institute for Meteorology	Raddatz et al. (2007)
2. MPI-ESM-P		Marsland et al. (2003)
1. MRI-CGCM3	Meteorological Research Institute	Yukimoro et al. (2011)
1. NorESM1-M	Norwegian Climate Centre	
2. NorESM1-ME		

Historical simulations by atmosphere–ocean coupled model are used. 28 models available on the federated ESG archive are used for analysis

Table 2 List of CMIP3 ensemble

Model	Institute	CMIP3-AO	CMIP3-AS	References
1. CCSM3	National Center for Atmospheric Research	○	○	Collins et al. (2004) Smith and Gent (2004)
1. CGCM3.1-T47 2. CGCM3.1-T63	Canadian Centre for Climate Modelling and Analysis	○	○	McFarlane et al. (1992) Flato (2005) Pacanowski et al. (1993) Salas-Mélia et al. (2005)
1. CNRM-CM3	Meteorologiques/Centre Europeen de Recherche et Formation Avancees en Calcul Scientifique	○		
1. ECHAM5/MPI-OM	Max Planck Institute for Meteorology	○	○	Roeckner et al. (2003) Marsland et al. (2003) Haak et al. (2003)
1. ECHO-G		○		Roeckner et al. (1996) Legutke and Maier-Reimer (1999) Min et al. (2004)
1. FGOALS-g1.0	LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences	○		Yu et al. (2002) Yu et al. (2004)
1. GFDL-CM2.0 2. GFDL-CM2.1	NOAA Geophysical Fluid Dynamics Laboratory	○	○	Delworth et al. (2006) Gnanadesikan et al. (2006) Wittenberg et al. (2006) Stouffer et al. (2006)
1. IPSL-CM4	Institut Pierre-Simon Laplace	○		Malti et al. (2006)
1. MIROC3-Hi 2. MIROC3-Med	Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology	○	○	K-1 model developers (2004)
1. MRI-CGCM	Meteorological Research Institute	○	○	Shibata et al. (1999) Yukimoto et al. (2001)
1. PCM		○		Washington et al. (2000)
1. UKMO-HadCM3	Met Office Hadley Centre	○		Gordon et al. (2000) Pope et al. (2000)
2. UKMO-HadGEM1		○	○	Martin et al. (2004) Roberts (2004)

Historical simulations by atmosphere–ocean coupled model (CMIP3-AO) and the control simulations by atmosphere-slab ocean coupled (CMIP3-AS) are used for analysis. Number of ensemble members in CMIP3-AO is 16, and that for CMIP3-AS is 10 which is available in the PCMDI data archive

created a “multi-physics” ensemble (MPE) by switching physical schemes of MIROC3.2 to those of MIROC5. Using a full factorial design, three schemes were changed in the MPE: vertical diffusion, cloud microphysics, and cumulus convection. Including the two control models, there are, therefore, 8 simulations in total.

2.2 Reliability and rank histogram of model ensembles

In the present study, we follow the same philosophy in the definition of reliability and interpretation of the rank histogram as Y12, which is analogous to how it is commonly

used in numerical weather prediction. The definition of the term “reliable” in this study is as follows: the ensemble is reliable if the observational data can be considered as having been drawn from the distribution defined by the model ensemble. That is, the null hypothesis of a uniform rank histogram is not rejected (JP08). Of course, in reality, creation of a perfect ensemble is impossible, so with enough data and ensemble members, all ensembles may be found to be unreliable at some level. What we are really testing here is whether the ensembles may be shown to be unreliable for the metrics of interest. Investigating the spatial scale at which the ensembles become unreliable is

Table 3 List of single-model ensembles

Ensemble	Experiment	Model	Number of parameter perturbed	Number of ensemble members	References of model and ensembles
HadCM3-AO	20th century by AOGCM	HadCM3	31	128	Gordon et al. (2000) Murphy et al. (2007) Collins et al. (2006a)
HadSM3-AS	Control by ASGCM	HadSM3	31	17	Pope et al. (2000) Webb et al. (2006) Yokohata et al. (2010)
NCAR-A	Control by AGCM	CAM3.1	15	100	Collins et al. (2006b) Jackson et al. (2004) Jackson et al. (2008)
MIROC5-AO	Control by AOGCM	MIROC5	10	36	Watanabe et al. (2010) Shiogama et al. (2012)
MIROC3-AS	Control by ASGCM	MIROC3.2	13	32	K-1 model developers (2004) Annan et al. (2005a, b) Yokohata et al. (2010)
MIROC-MPE-A	Control by AGCM	MIROC3.2 and MIROC5 ^a	Physical scheme changed ^a	8	K-1 model developers (2004) Watanabe et al. (2010) Watanabe et al. (2012)

HadCM3-AO, HadSM3-AS, MIROC3-AS, and NCAR-A used for the analysis are the same as those in Yokohata et al. (2012). Left column is the name of ensembles, and in the second left column, “AOGCM” denotes atmosphere–ocean coupled GCM, and “ASGCM” denotes atmosphere–slab ocean coupled GCM. Number of parameters perturbed and ensemble members are shown in the third and fourth column, details of which are described in the “Reference” column

^a Physical schemes of MIROC3.2 and MIROC5 are changed. Details are described in the main text and Watanabe et al. 2012

an interesting topic for future work, but is outside the scope of this paper (Sakaguchi et al. 2012).

Since the data are historical, the analysis here is essentially that of a hindcast, and since some of these data may have been used during model construction and tuning, it is debatable to what extent they can be considered to provide validation of the models. Furthermore, the relationship between current performance and prediction of future climate change remains unclear (e.g., Abe et al. 2009; Knutti 2010, Shiogama et al. 2011). Thus, reliability over a hindcast interval is not necessarily a sufficient condition to demonstrate that the model forecasts are good (Y12). On the other hand, it is clearly preferable that an ensemble should account for sufficient uncertainties to provide a reliable depiction of reality. Where an ensemble is not reliable in this sense, it must raise some doubts as to how credible it is as a representation of uncertainties in the climate system.

The method for calculating the rank histograms in this study is the same as that described in AH10 and Y12, and involves constructing rank histograms for the gridded mean climatic state of the model ensembles for the present-day climate with respect to various observational data sets. We use the 9 climate variables of surface air temperature (SAT), sea level pressure (SLP), precipitation (rain), the

top of atmosphere (TOA) shortwave (SW) and longwave (LW) full-sky radiation, clear-sky radiation (CLR, radiative flux where clouds do not exist), and cloud radiative forcing (CRF, radiative effect by clouds diagnosed from the difference between full-sky and clear-sky radiation, Cess et al. 1990).

We consider uncertainties in the observations by using two independent datasets, listed in Table 3 of Y12. As in Y12, we used the point-wise difference between each pair of data sets as an indication of observational uncertainty, although this is likely to be somewhat of an underestimate of the true error.

In addition to the mean climate states, we evaluated the long-term trend in the historical experiments by CMIP5-AO, CMIP3-AO, and HadCM3-AO. Due to its robust attribution to external forcing, we evaluate the long-term trend of SAT over the last 40 years (1960–1999). We do not investigate the twentieth century trend of PRCP, SLP, or TOA radiation because the interannual to decadal variability is generally large in these variables, and there are large uncertainties and sometimes an artificial trend in observations owing to the difficulty in measurement of these variables (Trenberth et al. 2007).

The methodology of the rank histogram calculation is described below. First, the model data and observational

data were interpolated onto a common grid (resolution of T42 in CMIP5-AO, CMIP3-AO, HadCM3-AO, MIROC5-AO, MIROC-MPE-A, and T21 for the other model ensembles). Second, we inflate the model ensemble to account for observational uncertainties by adding random Gaussian deviates to the model outputs as follows,

$$X'_{\text{model}} = X_{\text{model}} + \sigma_{\text{obs}}Z,$$

where X_{model} is the value of model ensembles, σ_{obs} is the standard deviation of the mean of two observations as listed in Table 3 of Y12, and Z is randomly sampled values from a normalised Gaussian distribution. Details are described in Sect. 2.4 of Y12. In this way, the sampling distributions of the observations and perturbed model data will be the same if the underlying sampling distributions of reality and models coincide. Due to the large number of data points, our results are robust to sampling variability in these random perturbations. Third, at each grid point, we compared the value of the observation with the ensemble of model values, evaluating the rank of the observation in the ordered set of ensemble values and observed value. Here a rank of one corresponds to the case where the value of observation is larger than all the ensemble members. We generate a global map of the rank of observation, $R(l,m)$, where l and m denote the index of latitudinal and longitudinal grid point, for each variable and each ensemble. Using the global map of rank of observation, $R(l,m)$, the rank histogram, $h(i)$ is the histogram of the ranks, weighted by the fractional area of each grid box over the whole grid.

The features of the rank histogram can be interpreted as follows. If a model ensemble was perfect such that the true observed climatic variable can be regarded as indistinguishable from a sample of the model ensemble, then the rank of each observation lies with equal probability anywhere in the model ensemble, and thus the rank histogram should have a uniform distribution (subject to sampling noise). On the other hand, if the distribution of a model ensemble is relatively under-dispersed such that the ensemble spread does not capture reality, then the observed values will lie towards the edge or outside the range of the model ensemble, and then the rank histogram will form a L- or U-shaped distribution. An ensemble with a persistent bias, either too high or too low, may either have a trend across the bins, or a strong peak in one end bin if the bias is sufficiently large. If the histogram has a domed shape with highest values towards the centre, then this implies that the ensemble is overly broad compared to a statistically indistinguishable one.

Since a model ensemble can be regarded as unreliable if the rank histogram of observations is significantly non-uniform, we performed a statistical test for uniformity, whose details are described in Y12. We use the technique introduced by JP08 and decompose the Chi square statistics, T ,

into components relating to “bias” (the trend across the rank histogram), “V-shape” (peak or trough towards the centre), “ends” (both left and right end bins are high or low), and “left-ends” or “right-ends” (the left or right end bin is high or low). Using the rank histogram, $h(i)$ as defined above, the Chi square statistics can be described as

$$T = \sum_{i=1}^k \frac{[n_{\text{obs}}h(i) - e_i]^2}{e_i} \quad (1)$$

where k is the number of bins in rank histogram (corresponds to the maximum rank), and i is the index of rank of the observation. $e_i = n_{\text{obs}}/k$ corresponds to the expected bin value for a uniform distribution, and $n_{\text{obs}}h(i)$ is the “observed value of i th bin” in JP08. In the present study, $h(i)$ is calculated as a form of probability, which corresponds to the probability that the rank of observation comes to i th bin, and n_{obs} is the “number of observation” in JP08. Since values of neighbouring grid points are highly correlated, their ranks of observation cannot be considered as independent of each other, and thus n_{obs} is also referred to as the “effective degrees of freedom of the data” as discussed in AH10 and Y12.

In Y12, a value of 10 was used for n_{obs} , based on the estimate by Annan and Hargreaves 2011, in which SAT, SLP and rain of the CMIP3 ensemble ranges from 4 to 11. However, the effective degree of freedom may be different among model ensembles, and the statistical test for the uniformity also depends on n_{obs} . In the present study, therefore, we estimate the effective degree of freedom based on the method of Bretherton et al. (1999), which is described in the next section.

As described in JP08, under the null hypothesis of a uniform underlying distribution, the Chi square statistic for the full distribution is sampled from approximately a Chi square distribution with $(k - 1)$ degrees of freedom. Using a table of the Chi square distribution and the value of T in Eq. (1), we can calculate the p value and reject the hypothesis of uniform distribution if the p value is smaller than the level of significance. Similarly, each of the components such as bias, V-shape, ends, left-ends, and right-ends calculated by the formulation of JP08, should have an approximate Chi square distribution with one degree of freedom. We can also estimate the p value of these components and test the hypothesis of a uniform distribution.

2.3 Effective degrees of freedom of model ensembles

We use the formulation of EDoF by Bretherton et al. (1999). Using the spatial patterns of climatology of model ensemble members, EDoF can be described as

$$N_{\text{ef}}(n) = \left(\sum_{k=1}^n f_k^2 \right)^{-1} \quad (2)$$

where N_{ef} is the effective degree of freedom, n is the number of members in a model ensemble, f_k is the fractional contribution of EOF k to the total variance. f_k is calculated from the EOF across the climatology of model ensemble members. Equation (2) means that if the fractional contribution from the small k EOF is large, then the differences in special patterns among model ensemble members can be explained by the pattern of small k EOF, and thus the EDoF of model ensemble is small.

In Bretherton et al. (1999), it is shown that for any sampling distribution, the estimate of EDoF presented in Eq. (2) based on a finite sample will tend to underestimate the true EDoF which would be obtained by an infinite sample from the same distribution. $N_{\text{ef}}^{\text{true}}$, the value of EDoF if the number of model ensemble members is infinity, can be estimated as follows.

$$N_{\text{ef}}^{\text{true}} = \frac{N_{\text{ef}}(n)}{1 - N_{\text{ef}}(n)/n} \quad (3)$$

The EDoFs calculated as above are used for the statistical test for the reliability of rank histogram. We set $n_{\text{obs}} = N_{\text{ef}}^{\text{true}}$ in Eq. (1), and then perform the statistical test using the rank histogram described in Sect. 2.2.

2.4 Distances between observation and model ensemble members

The rank histogram analysis discussed in Sect. 2.2 only considers the rank ordering of models and observations, and thus information on the distances between observation and ensemble members is missing. It also takes an intrinsically univariate and scalar viewpoint of the data, considering each observation independently of the others. An alternative approach, based on minimum spanning trees (Wilks 2004), handles multidimensional data sets directly, and also considers the distance between ensemble members and the observations. Therefore, we also investigate our ensembles using this approach, which we now briefly describe. We consider a 2D data field, and the equivalent output field from each ensemble member, as points (“nodes”) in a high dimensional space, with the length of the “edge” or line segment between each pair of them defined as the area-weighted RMS difference. In graph theory, a tree is a set of $n - 1$ edges which collectively connect n nodes, and if each edge is assigned a length function, then a minimum spanning tree is a tree of minimum total distances (which will be unique, if all the pairwise distances differ).

Therefore, in order to calculate the minimum spanning tree (MST), we first evaluate the pair-wise distances between the climatology of an observational data field and equivalent output from model ensemble, D_{kl} , via the global area-averaged RMS difference as follows,

$$D_{kl} = \sqrt{\frac{1}{n_i n_j} \sum_{j=1}^{n_j} \sum_{i=1}^{n_i} [X_k(i,j) - X_l(i,j)]^2 A_{ij}} \quad (4)$$

where i and j denotes the index for the grid points, and n_i and n_j are the numbers of grid points for the latitude and longitude. k and l in Eq. (4) are the index of observation and model ensemble members used for the pair-wise distances. Here, we defined $k < l$, $k = 0$ for the observation, and k or $l =$ from 1 to n_{ens} for the model ensembles, where n_{ens} is the model ensemble members described in Table 3. $X_k(i,j)$ and $X_l(i,j)$ denote the values of the climate variables used for the above analysis. A_{ij} is the weight of each grid area fraction (ratio of each grid area to global area).

Once the pair-wise distances between observation and model ensemble members defined in Eq. (4) are obtained, the MST for any set of nodes, and its total length (i.e., the sum of the lengths of its edges) can be readily generated using a standard algorithm. Here, in order to understand the relationship of the distances between the ensemble members and those between observation and ensemble members, leave-one-out analysis as described in Wilks (2004) is performed. First, the MST for the nodes excluding the observations, namely the MST for the model ensemble members, defined as $M(0)$, is calculated. Then, the MSTs in which the observational data is used to replace each ensemble member in turn from 1 to n_{ens} , defined as $M(k)$ for $k = 1$ to n_{ens} is calculated. Finally the rank of the total length of $M(0)$ among those of $M(k)$ for $k = 1$ to n_{ens} is evaluated. Here, the rank is defined as one if the $M(0)$ has the smallest total length. If the observations were drawn from the same distribution as the ensemble, then the length of the $M(0)$ should be indistinguishable from the lengths of the other $M(k)$. If, however, the observations are relatively distant from the ensemble, then the $M(0)$ will be shorter than the $M(k)$. Given a sufficiently large number of observational data sets, the histogram of the ranks of the associated MSTs can be generated (the MST rank histogram) but, since we only have a small number of data fields, we prefer to examine the ranks on an individual basis in Sect. 3.1.

In order to focus more directly on the distances between observation and ensemble members, we also calculate the average of distances between the observation and the models. For the observations, and then for each ensemble member in turn, we calculate the average of distances from it, to all the other nodes:

$$\overline{D}_k = \frac{1}{n_{\text{ens}}} \sum_{l \neq k}^{n_{\text{ens}}} D_{kl} \quad (5)$$

If the distances from the observation to ensemble members are larger than those among ensemble members, \overline{D}_0 is larger than \overline{D}_k with $k \neq 0$.

3 Results and discussions

3.1 Rank histogram of model ensembles

A multi-variate analysis of rank histograms is shown in Fig. 1. Here, we create the nine maps of the rank of the observation among model ensembles using the nine variables described in Sect. 2, and create the (area-weighted) rank histogram. As described in Sect. 2.2, the histogram will be uniform if the model ensemble is ideal (the observational data is drawn from the ensemble). On the other hand, the histogram will have a dome-shaped distribution if the ensemble is over-dispersed, and U- or L- shaped if the ensemble is under-dispersed. In Fig. 1, the

number of bins of each rank histogram is one plus the number of model ensemble members.

As Fig. 1 shows, the difference between the MMEs (red) and SMEs (blue) is striking. The rank histograms for the MMEs are dome-shape in general, while those of SMEs are U-shaped (with large peaks at the highest and lowest rank). This means that in SMEs, there are large areas where either all of the ensemble members underestimate the observation (the peak at the lowest rank) or all the members overestimate it (the peak at the highest rank). This result is similar to that shown in Y12.

The features of the MMEs are very similar to each other and consistent with the results of Y12 (Fig. 1). This suggests that the indication of a dome shaped rank histogram

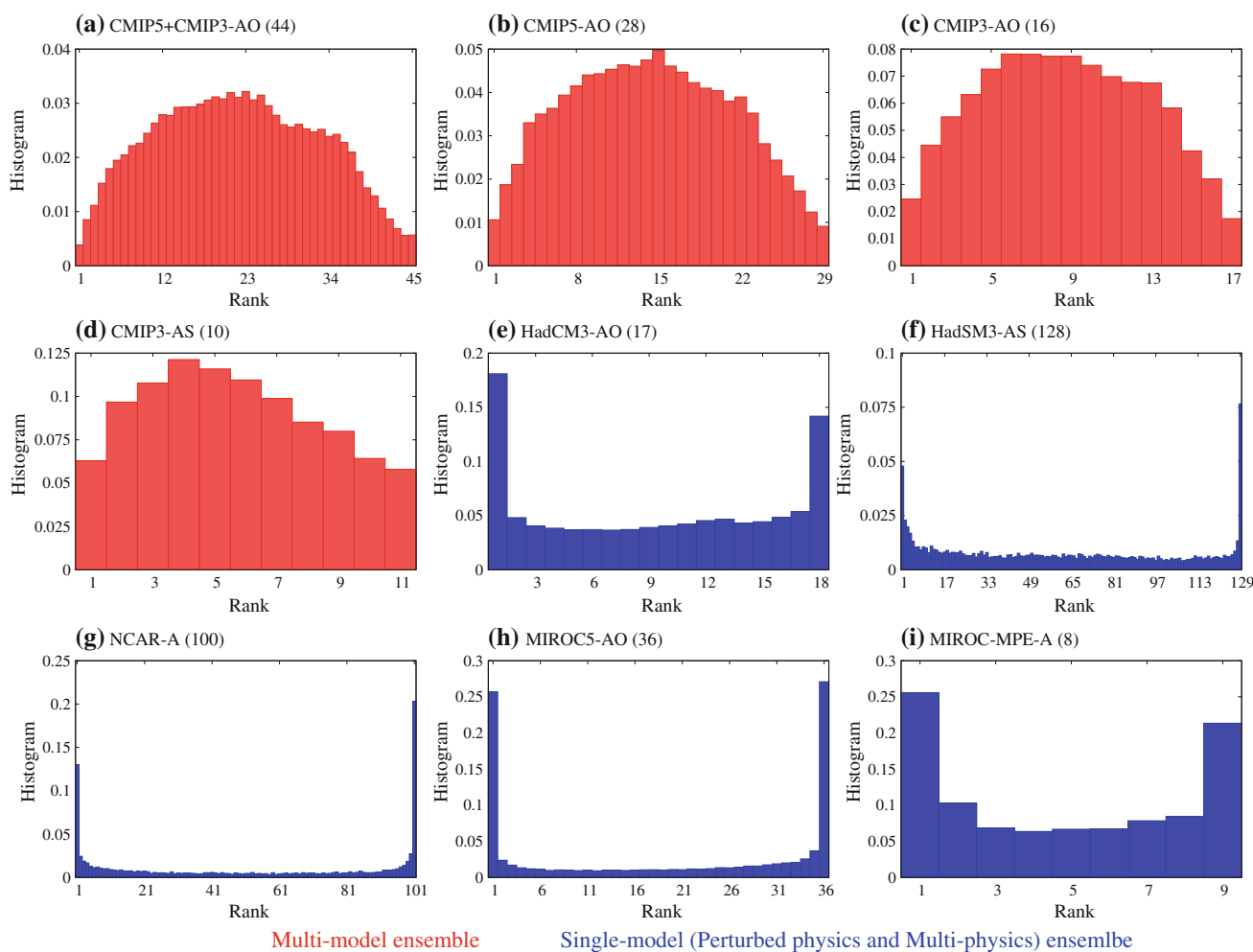


Fig. 1 Multi-variate rank histogram of multi-model and single-model ensembles. The multi-model ensembles are **a** CMIP5+CMIP3-AO, **b** CMIP5-AO, **c** CMIP3-AO, **d** CMIP3-AS, and single-model ensembles are **e** HadCM3-AO, **f** HadSM3-AS, **g** NCAR-A, **h** MIROC5-AO, **i** MIROC-MPE-A. Multi-model ensembles are shown in red, and the single model ensembles (perturbed physics and multi-physics ensembles) are shown in blue. In these ensembles, atmosphere–ocean coupled (AO), atmosphere-slab ocean coupled

(AS), and atmosphere-only (A) global climate models are used. Numbers of ensemble members are shown in *parenthesis*. Here we count the rank of observation among model ensemble members and create histogram, so the number of rank in horizontal axis is from one to the number of ensemble plus one. We use the nine climate variables such as surface air temperature, precipitation, sea level pressure, shortwave (SW) and longwave (LW) net flux, cloud radiative forcing, and clear-sky flux

first presented in Y12 may not have been due to chance, but rather may represent a persistent phenomenon (albeit of unknown source) in the generation of climate model ensembles. However, the histograms do not fail the significance tests described in the following section, so any intrinsic non-uniformity is relatively modest. Fig. 2 shows the rank histogram of each climate variable described in Sect. 2.2. In order to compare the features of the model ensembles, here the number of horizontal bins is set to the same value (the maximum rank = 9) as in Y12. As shown in Fig. 2, The characteristics of the rank histograms are also rather similar for the same variables for the two CMIP ensembles. The histograms of SAT, rain, and SLP are almost the same for CMIP5-AO and CMIP3-AO. The peaks of the histograms in the SW and LW radiation are slightly different between these ensembles, and this is the cause of the double-peak feature for CMIP5+CMIP3-AO

apparent in Fig. 1. While we do not investigate the issue of model similarity or near-duplication in this investigation, the presence of such models would not tend to bias the rank histograms in any particular direction, but adds some sampling noise and thus tend to increase the degree of non-uniformity.

As found in Y12, the histograms of SMEs tend to have the peaks at the highest and lowest rank, but the details of this varies between the model ensembles and variables. In general, the histogram of climate variables only related to dynamical process (SLP, SW clear-sky radiation) tend to be U-shape in SMEs, possibly because model parameters related to dynamical processes are not generally perturbed in the SMEs. It is interesting to note that the peaks at the highest and lowest end of MIROC-MPE-A are smaller than those of MIROC5-AO, MIROC3-AS. For example, the histograms of the LW radiation are not U-shape in

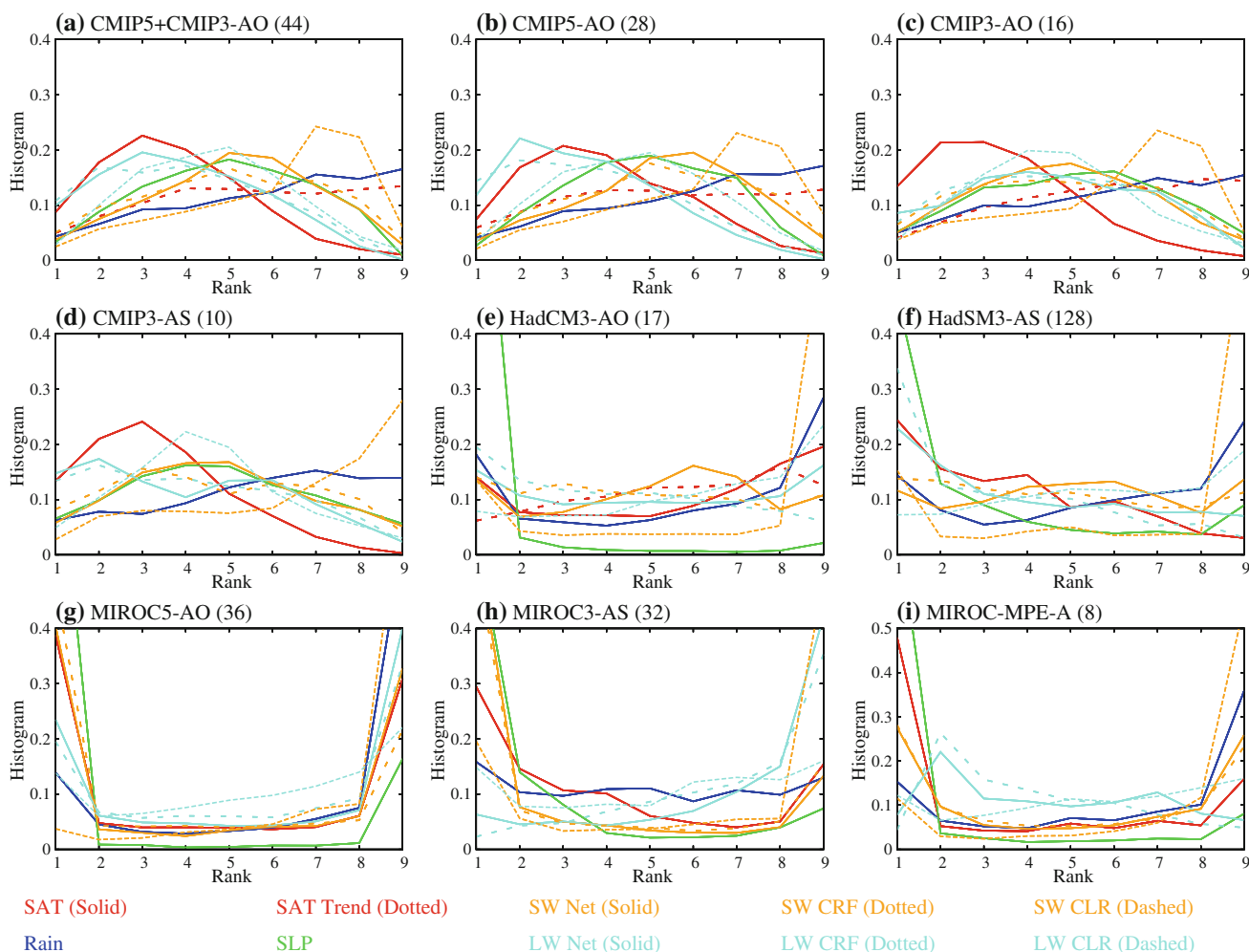


Fig. 2 Same as Fig 1 but for the rank histogram of the climate variables such as surface air temperature (SAT, red solid), SAT trend (red dotted), precipitation (blue), sea level pressure (green), SW net, cloud radiative forcing, clear-sky radiation (orange solid, dotted, and dashed), and LW net, cloud radiative forcing, and clear-sky radiation

(cyan solid, dotted, dashed) at the TOA. Model ensembles are **a** CMIP5+CMIP3-AO, **b** CMIP5-AO, **c** CMIP3-AO, **d** CMIP3-AS, **e** HadCM3-AO, **f** HadSM3-AS, **g** MIROC5-AO, **h** MIROC3-AS, **i** MIROC-MPE-A

MIROC-MPE-A. As described in Sect. 2.1, MIROC-MPE-A is constructed by replacing model schemes for cloud physics, vertical diffusion etc. (Watanabe et al. 2012). Therefore, it is reasonable to expect that MIROC-MPE-A would have more structural diversity than the ensembles of its original models, MIROC5-AO and MIROC3-AS, which would lead to the rank histograms for the MPE generally being closer to a flat distribution.

In Y12, the statistical test for reliability was performed by assuming the effective degree of freedom, n_{obs} in Eq. (1) is 10. However, in the present study, we estimate n_{obs} using the EOF analysis explained in the next section, and then perform the statistical tests in Sect. 3.3.

3.2 Effective degree of freedom of model ensembles

The EDOF of model ensembles formulated by Bretherton et al. (1999) is shown in Fig. 3. Here, all the nine variables used for the rank-histogram analysis are combined and EDOFs are calculated for the multivariate distribution. In order to calculate EOF consistently across different climate variables, each climate field is normalised by its global ensemble standard deviation. The dependency of EDOF on the ensemble size is investigated in Fig. 3. For example, at the point of $x = 10$ for the CMIP5-AO in Fig. 3, we chose

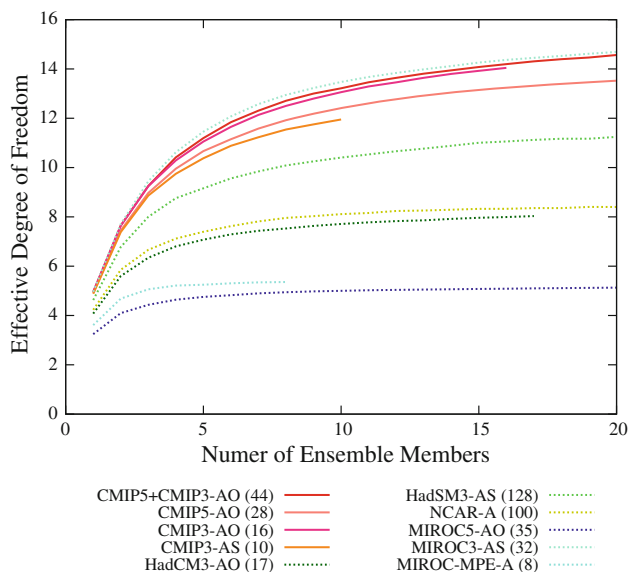


Fig. 3 Effective degree of freedom (EDoF) of climate model ensembles. Nine climate variables, such as SAT, rain, SLP, SW and LW net flux, cloud radiative forcing, and clear-sky radiation at the TOA are used for analysis. In order to calculate all the variables consistently, each field is normalised by the global ensemble standard deviation. Horizontal axis is the number of ensemble members used for the DoF calculation. We chose ensemble members by random sampling many times (1,000 at a maximum), calculate the DoF of each set of ensemble members, and plot their average. Numbers of ensemble members are shown in *parenthesis* in the legend

10 ensemble members (each member has 9 variables, so 90 variables in total are used for the calculation) by random sampling out of 28 ensemble members 1,000 times, and calculate the EDOF for each set of 10 ensemble members, then plot the average of the EDOFs.

As shown in Fig. 3, the EDOF of model ensembles increases with increasing number of ensemble members, appearing to asymptote to a relatively small value for some ensembles, but continuing to increase in other cases. The SMEs tend to exhibit systematically lower EDOF than the MMEs, with the exception of the MIROC3-AS SME. This analysis suggests that parametric variation is generally less effective than structural changes in spanning a diverse range of climatological behavior.

Each EDOF of the nine climate variables is shown in Fig. 4. Features of EDOFs are different among climate variables. In general, the EDOFs of the MMEs are large compared to those of the PPEs. This result is basically consistent with the result from the rank histogram analysis, as shown in Fig. 2. However, for SLP and LW-CLR, the EDOFs of the MMEs are generally small and some of the PPEs have larger EDOFs than those of the MMEs. This means that the spatial patterns of SLP and LW-CLR tend to be similar among the MME ensemble members.

3.3 Reliability of model ensembles from statistical tests of rank histogram

Statistical analyses for the test of uniformity of the rank histograms are performed using the EDOF calculated in Eq. (3) and described in Sect. 2.2. In Table 4, the p values calculated from the rank histograms are shown. As described in JP08, if the p value is smaller than the threshold, then the histogram can be considered to be significantly different from the uniform distribution. Note that the essential differences in the test between this study and Y12 are (1) the EDOF corresponding to n_{obs} in Eq. (1) is estimated in the form of Eq. (3) while it was assumed to be 10 in Y12, and (2) the number of bins of the rank histogram are equal to the number of ensemble members plus one in this study, while that is reduced to 11 in order to compare p values among model ensembles in Y12. Note that if the number of bins is larger, then (assuming the total ensemble spread does not increase and thus the end bins are unchanged), the tests for the U-shape and L-shape will become more powerful and the result is more likely to be significant.

The p values using the nine climate variables (denoted as “Overall” in Table 4) of MMEs are larger than the threshold (0.05, significant level = 5%), which means that, according to this analysis, these ensembles have not been shown to be unreliable. Although their rank histograms appear somewhat domed, they are acceptably close

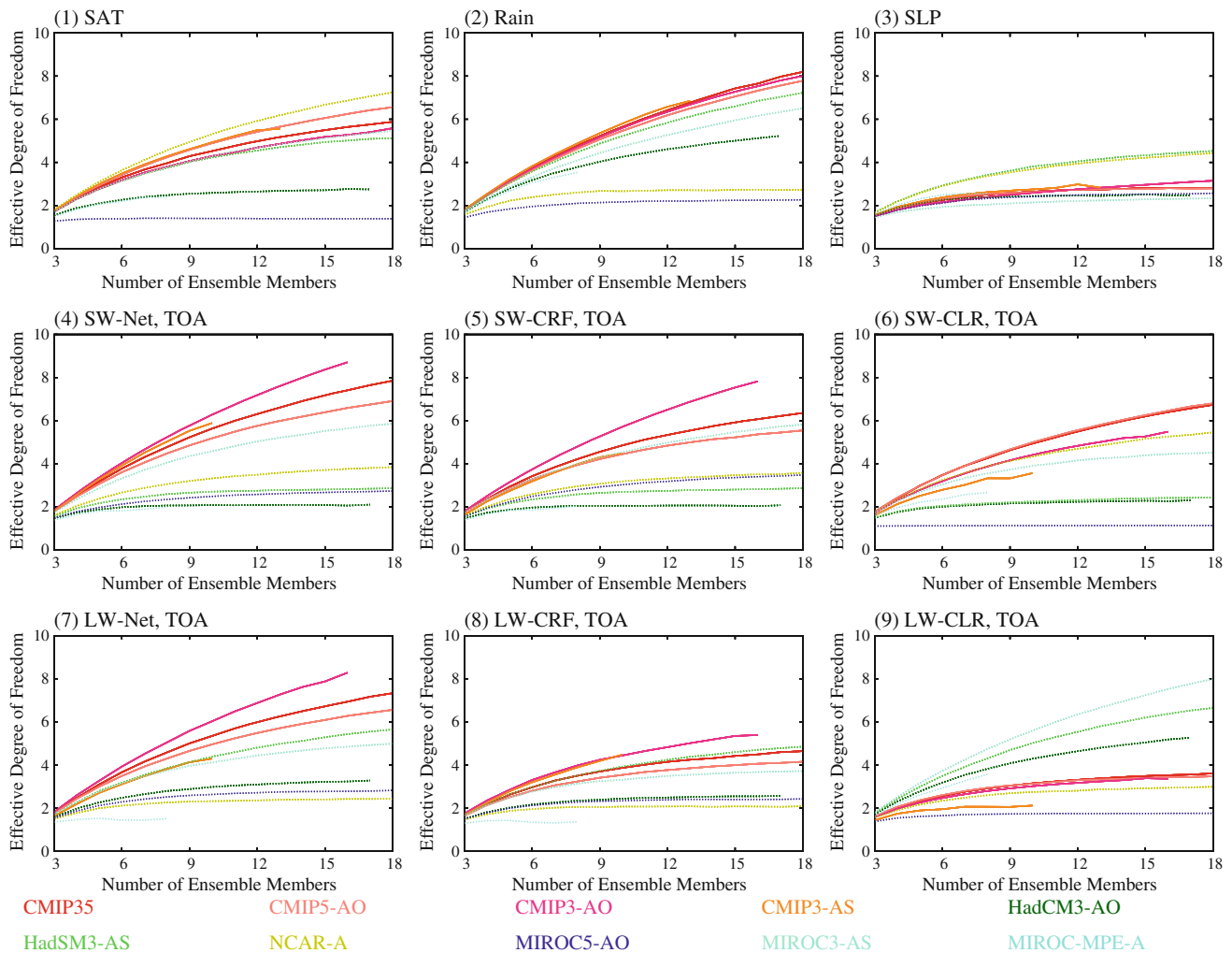


Fig. 4 Same as Fig. 3 but for the effective DoF of the nine climate variables which were averaged in Fig. 3. In order to identify the differences between model ensembles more clearly, the range of horizontal axis is chosen from 3 to 18

to the uniform distribution. On the other hand, the p values of many of the results from the SMEs are smaller than the threshold. These ensembles can then be said to be unreliable because their rank histograms are significantly different from the uniform distribution. The U-shaped characteristic of the SME histograms indicates that these ensembles are under-dispersed.

Among the SMEs, the p value of HadCM3-AO is almost on the threshold (0.05), and MIROC-MPE-A is larger than this threshold. One possible reason for the relatively good performance of these models in the statistical test is that the number of ensemble members (i.e., number of bins in Fig. 1) is small, as discussed above. Another possible reason for the reliability of MIROC-MPE-A compared to the other SMEs might be that the multi-physics ensemble has more structural diversity compared to the original MIROC5-AO or MIROC3-AS, and thus it is sufficiently diverse to span the observations.

In Table 4, p values of the nine climate variables (plus SAT trend for the ensembles performing the historical simulation) are also shown. In MMEs, the number of climate variables with p values smaller than the threshold is zero, which means these MMEs are reliable for all the variables investigated. On the other hand, in SMEs, the reliability varies between climate variables and model ensembles. HadCM3-AO and HadSM3-AS have relatively better performance compared to other ensembles (four of the p values are less than 0.05).

The statistical test of histogram uniformity also depends on the n_{obs} in Eq. (1), which corresponds to the EDoFs (JP08). As discussed in Bretherton et al. (1999), there are uncertainties in the EDoF in Eq. (3), so the true values of EDoF may be larger or smaller than those estimated in the present work. Therefore, we investigate the sensitivity of the statistical test to the EDoF. In Fig. 5, the relationship between the p value of the statistical test

Table 4 The minimum p values of Chi square statistics calculated from the rank histogram

Value	CMIP35-AO	CMIP5-AO	CMIP3-AO	CMIP3-AS	HadC3-AO	HadS3-AS	NCAR-A	MIRO5-AO	MIRO3-AS	MIRO-MPE-A
# of ens	44	28	16	10	17	128	100	36	32	8
Over-all	0.1666	0.2187	0.2444	0.5025	0.0499	0.0021	0.0000	0.0000	0.0000	0.1529
SAT	0.1397	0.1973	0.1126	0.0744	0.5760	0.2447	0.0000	0.0089	0.0715	0.0318
Rain	0.1543	0.1520	0.2712	0.3895	0.0421	0.0065	0.0000	0.0000	0.4034	0.0474
SLP	0.3705	0.3251	0.5283	0.4251	0.0000	0.0032	0.0000	0.0000	0.0011	0.0001
SW Net	0.0612	0.1220	0.0595	0.1735	0.7786	0.2361	0.0001	0.0000	0.0000	0.2237
LW Net	0.1350	0.0685	0.1739	0.3324	0.5814	0.2361	0.0003	0.0022	0.0096	0.7967
SW CRF	0.2648	0.2985	0.2486	0.4832	0.8193	0.7722	0.0000	0.0000	0.0000	0.2468
LW CRF	0.3430	0.2498	0.3248	0.3590	0.5973	0.0920	0.0011	0.0379	0.0829	0.6906
SW CLR	0.1845	0.1672	0.3413	0.1861	0.0005	0.0000	0.0000	0.0005	0.0000	0.0052
LW CLR	0.2338	0.2738	0.2795	0.4113	0.2557	0.4781	0.3169	0.4899	0.3190	0.6425
SAT trend	0.4565	0.5968	0.3194	NA	0.3356	NA	NA	NA	NA	NA
# of $p < 0.05$	0	0	0	0	3	3	8	8	5	4

The third row indicates the value calculated from multi-variate (“overall”) histogram using the nine climate variables used in Fig. 1. Below that, the minimum p value for the ten climate variables are shown. The SAT trend can be calculated only for the model ensembles with historical simulations. At the last row, the number of variables with p value less than 0.05 out of ten (ensembles with the historical simulations) or nine (ensembles with the control simulations) is indicated. p values less than 0.05 indicated by the bold font

and EDoF are shown. For the MMEs, the p -values calculated from the Chi square statistics of the “V-shape” component (metric of dome-shape, JP08), namely the test of ensemble being “over-dispersed”, are shown in the form of $(1-p)$ value. If this value is close to one, the ensemble can be considered “over-dispersed”. On the other hand, for the SMEs, the p -values calculated from the Chi square statistics of “ends” components (metric of U-shape, JP08), namely the test of ensemble to be “under-dispersed”, are shown. We plot these values because, as discussed above, it seems that the histograms of MMEs are tending towards dome-shaped and those of the SMEs are U-shaped, so that these tests are the most critical. The EDOFs of model ensembles estimated from Eq. (3) are shown as the circles on the lines. The p -values of CMIP5+CMIP3-AO, CMIP5-AO, CMIP3-AO are closer to the threshold of being “over-dispersed” compared to that of CMIP3-AS. The EDOFs would have to be about a factor two larger than estimated for the ensembles to fall above the threshold of being “over-dispersed”. These results are consistent with a previous study investigating the “dissimilarity” of model ensembles. Masson and Knutti (2011) also found that the HadCM3-AO ensemble members are more similar to each other than the CMIP model ensemble members are.

Conversely, the p -values of all the SMEs apart from MIROC-MPE-A are smaller than the threshold of “under-dispersed”. Only the p -value of HadCM3-AO is sensitive to small changes in the EDOF, as a slight decrease would put it above the threshold.

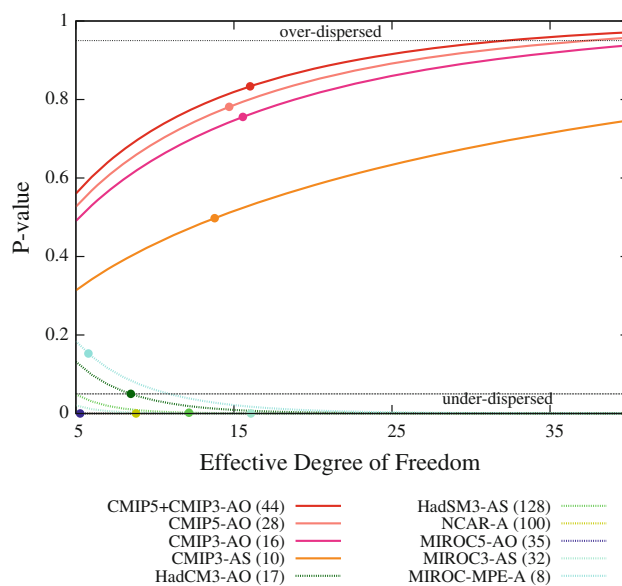


Fig. 5 Dependencies of p value of Chi square statistics of ranks histogram on the effective degree of freedom. For the multi-model ensembles (CMIP5+CMIP3-AO, CMIP5-AO, CMIP3-AO, CMIP3-AS), “1- p value” is shown and p -value is calculated from the Chi square statistics of the “V-shape” component (metric of dome-shape, Jolliffe and Primo (2008)). If the values of horizontal axis is larger than 0.95, the model ensemble can be regarded as “over-dispersed”. For the single model ensembles (HadCM3-AO, HadSM3-AS, NCAR-A, MIROC5-AO, MIROC3-AS, MIROC-MPE-A), p value calculated from the Chi square statistics of the “ends” component (metric of U-shape) are shown. If the values of horizontal axis is larger than 0.05, the model ensemble can be regarded as “under-dispersed”. Colors and line types are the same as those in Fig. 4, and the circles on the curves lies on the values of effective degree of freedom calculated in Fig. 4 and Eq. (4)

The reason for the tendency towards a dome shape in the MME is unclear. Y12 describes how tuning an ensemble to observations will tend to centralise it on them (meaning that the distance from ensemble mean to observations, normalised by ensemble spread, will shrink). Thus, tuning to modern observations might tend to result in a domed rank histogram if the untuned ensemble had a flat distribution. However, several of the SMEs have certainly been tuned to observations, without this phenomenon occurring and being under-dispersed, and there seems no direct way to measure to what extent this tuning has been explicitly or implicitly performed for MMEs, and for which climatic variables.

We should note that the rank histogram technique is often used in the field of numerical weather prediction where a larger number of observations and simulations are available (and thus the effective degrees of freedom are greater) compared to the present work. Therefore, the rank histogram results shown here may be less convincing. For this reason, in the next section we investigate the relationship between the observations and model ensembles based on their distances.

3.4 Distances between observation and model ensembles

Since the rank histogram discussed above evaluates only the rank ordering of observations amongst model ensemble members, we also investigate the distances of the model ensembles to the observations in various ways. First, we calculate the minimum spanning tree (MST) by removing observation and each ensemble members one by one as

described in Sect. 2.3. Using this procedure we obtain total ensemble number plus one MSTs. If the ensemble members are collectively far away from the observation (compared to their distances from each other), then the MST omitting the observation is smaller than the MSTs removing model ensemble members. With only a small number of data sets, we do not explicitly form the rank histogram and test for non-uniformity, but instead examine the rank of the MST for each variable in turn and consider whether it lies at the extreme end of the set of MST lengths.

Table 5 shows the rank of MST omitting the observation among the set of all MSTs obtained by removing the observation and each ensemble member. Here, a rank of one corresponds to the smallest MST. For calculation of the overall MST, we use the 9 climate variables. In order to calculate the distances consistently across different climate variables, each climate field is normalised by its global ensemble standard deviation. As shown in Table 5, the rank of MSTs without observation in CMIP5-AO, CMIP3-AO, and CMIP3-AS MMEs appears to vary widely across the possible range. The ranks of some variables of CMIP3-AO are large, which means that the distance between the model ensemble members is larger than that between the observation and the ensemble members. On the other hand, in SMEs and also the MIROC MPE, the ranks of MSTs without observation are often one or very small (e.g. within the lowest 5 % of all MSTs), which suggest that the MST without observation is very small and the distance between model ensemble members and observation is large compared to the distance among ensemble members.

Table 5 Rank of minimum spanning tree (MST) without observation among MSTs of observation plus model ensemble members removing each ensemble members

Value	CMIP35-AO	CMIP5-AO	CMIP3-AO	CMIP3-AS	HadC3-AO	HadS3-AS	NCAR-A	MIROC3-AS	MIROC5-AO	MIROC-MPE-A
# of ens	44	28	16	10	17	128	100	32	36	8
Over-all	16	12	10	5	1*	2*	1*	1*	1*	1*
SAT	10	8	6	6	1*	2*	1*	1*	1*	1*
Rain	17	7	14	8	1*	4*	1*	5	1*	1*
SLP	21	14	15	8	1*	2*	2*	1*	1*	1*
SW Net	21	14	15	7	3	5*	1*	1*	1*	1*
LW Net	21	12	13	4	1*	4*	1*	2	1*	2
SW CRF	14	11	11	5	3	5*	1*	1*	1*	1*
LW CRF	17	12	12	4	2	3*	1*	2	1*	2
SW CLR	20	12	14	4	1*	1*	1*	1*	1*	1*
LW CLR	11	7	11	6	1*	2*	1*	1*	1*	1*
SAT trend	10	8	6	NA	5	NA	NA	NA	NA	NA

The first row indicates the climate variables used for the analysis which are the same as those in Fig. 1, and the left column shows the climate model ensembles with number of ensemble members in parenthesis. Ranks within the lowest 5 % are indicated as bold font with *

Features of the distances between observation and model ensemble members are further investigated in Fig. 6. For each ensemble we calculate the pair-wise distances between observations and model ensemble members, and between the model ensemble members. The results are shown in Fig. 6. For each variable, the averages of the pair-wise distances between each model ensemble member and the observations are shown as circles. The distribution obtained by calculating, for the whole ensemble, the average pair-wise distances between one ensemble member and all the other ensemble members plus the observational data are shown by the box and error bar icons. Consistent with the MST analysis, the average of distances from observation to model ensemble members (circle) in MMEs does not appear inconsistent with the range of average distances from a particular ensemble member to other members plus the

observation (error bars). On the other hand, in SMEs, the average of distances from observation to ensemble members (circle) is larger compared to those from ensemble members (error bars) as shown in Fig. 6. It is also noticeable that the distances between the MME members are generally rather larger than for the SMEs.

In Fig. 6, the values of circle indicate the average of error of model ensemble members. Especially for the climate variables such as SAT, SLP, SW and LW clear-sky radiation as shown in Fig 6(1), (3), (6), and (9), the average of errors in MMEs and SMEs are similar, but the distances between model ensembles in MMEs are larger than SMEs. As discussed in the analysis of rank histogram, the inability for the SMEs to have sufficient diversity may be related to the fact that parameters in dynamical processes were not perturbed in SMEs.

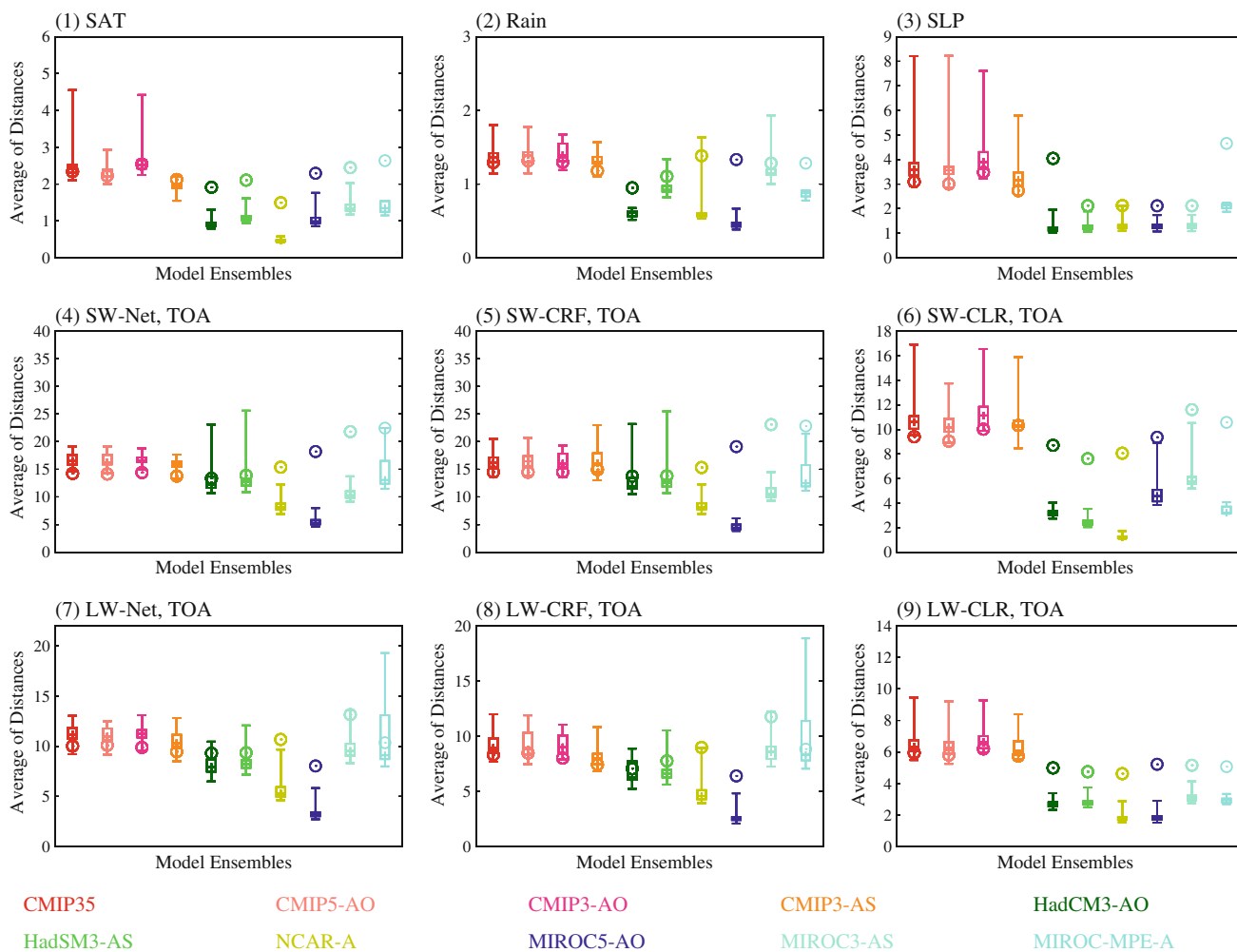


Fig. 6 Average of distances between ensemble members and observation. Error bars (2.5–97.5 %) and boxes (33–67 %) and central lines (median) represent the range of mean distances between a specific ensemble member and all other members plus observations. Circles represent the mean distance from observation to all ensemble

members. Colors for identifying ensemble members are the same as those of Fig. 5, and each panel shows 1 SAT, 2 rain, 3 SLP, 4 SW net radiation 5 SW cloud radiative forcing, 6 SW clear-sky radiation, 7 LW net radiation, 8 LW cloud radiative forcing, and 9 LW clear-sky radiation at the TOA

4 Summary

In the present study, the reliability of the state-of-the-art MME of CMIP5 as well as CMIP3, and a number of SMEs (summarised in Tables 1, 2, 3) are investigated with rank histograms calculated from the simulations of present-day climatology. The climate variables of surface air temperature, precipitation, sea-level pressure, and shortwave and longwave radiation at the top of the atmosphere are used for the analysis. The overall features of the ensembles are investigated through multi-variate analysis using all these climate variables. The reliability of model ensembles is evaluated in a more thorough and consistent way than in AH10 and Y12: the “effective degree of freedom” (EDoF) in Chi square statistics, n_{obs} in Eq. (1), is estimated by Eq. (3) formulated in Bretherton et al. (1999). Then, the statistical tests for the reliability of model ensembles are performed based on the rank histogram using estimated n_{obs} , and the numbers of bins in the histogram are not reduced. In addition to the rank histogram, the distances between the observation and model ensemble members are also investigated in various ways. Our results are summarised as follows.

1. The rank histograms using all the climate variables of MMEs have a tendency towards being dome-shaped with a peak around the middle rank, while those of SMEs are U-shape with strong peaks at the highest and lowest ranks (Fig. 1). This indicates that the spread of MMEs tend towards being “over-dispersed” in that the rank of observations generally stays close to the middle of the range, while that of SMEs tend to be “under-dispersed” in which all the ensemble members often overestimate or underestimate the observation. Even though the over-dispersion of the MMEs does not reach the level of statistical significance, the similarity of CMIP5 to CMIP3 (Fig. 1 and 2), suggests that this has arisen as a consequence of the way in which the diverse range of models has been constructed, rather than merely occurring by chance.
2. The EDoF of model ensembles are calculated by changing the ensemble sizes (Figs. 3, 4), and it is found that the MMEs generally have large EDoF compared to the SMEs. One of the SMEs, MIROC3-AS has similar EDoF to the MMEs. The method used to sample the parameters might effect the resultant EDoF in the PPEs.
3. Using the EDoF formulated in Eq. (3), a statistical test for the reliability of model ensembles is performed (Table 4). Multi-variate histograms using all the climate variables (“Overall” in Table 4) indicate that the rank histograms of MMEs are not significantly different from the uniform distribution, and thus, with

respect to this analysis, the MMEs, may be considered to be reliable. On the other hand, the rank histograms of the SMEs, except the histogram of MIROC-MPE-A, are U-shaped and significantly different from the uniform distribution indicating that they are under-dispersed (see Fig. 1). These results suggest that the structural diversity is important in order to include the observation among the spread of model ensembles. Large EDoF in MMEs should contribute to their reliability.

4. The dependencies of reliability on the EDoF are also investigated (Fig. 5). The MMEs, which tend towards being over-dispersed, remain reliable within an increase of EDoF of about a factor of two. Most of the SMEs are also robustly under-dispersed, but HadCM3-AO could be considered reliable if the EDoF has been slightly overestimated. The rank histogram of MIROC-MPE-A is not statistically different from the uniform distribution. This may be because the number of ensemble members is small, which causes the statistical test to be less powerful, and also because the “multi-physics” ensembles can sample the structural uncertainties to some extent by changing the physical schemes (Watanabe et al. 2012).
5. MSTs (minimum sum of distances between ensemble members, Table 5) and the averages of the distances between the observations and model ensemble members (Fig. 6) are calculated. In the MMEs, the distances between ensemble members are not different from those between the observation and ensemble members. On the other hand, the distances between ensemble members in the SMEs are smaller than those between the observation and ensemble members. These results are consistent with the analysis of rank histograms in which the spread of MMEs include the observation, but that of SMEs do not.

It should be noted that the SMEs examined here were not explicitly designed to be reliable according to the rank histogram metric, although they were designed with some expectation that each member of the ensemble would verify well against a basket of observations. It would be an interesting endeavor to set out to produce a reliable PPE or MPE and to design a perturbation algorithm accordingly. As shown in Collins et al. (2010), the algorithm for parameter perturbations in a PPE does influence the diversity of the mean climates and trends seen in each member, suggesting that such an endeavor might be possible, perhaps using some iterative algorithm. Such challenges remain a subject of future research.

On the other hand, since our analysis reveals that the MMEs are reliable when compared to the subset of observational fields examined, or their spread tends to be

“over-dispersed” rather than “under-dispersed”, it may be useful to apply unequal weights to generate improved simulations of future predictions (e.g., Collins et al. 2012). For example, if we chose a subset of ensemble members from the CMIP5 ensemble, the rank histogram approach should be useful. We can choose a subset of members whose reliability become higher, i.e., with a rank histogram close to uniform. However, present-day reliability does not necessarily imply reliability for future projections, hence additional work is required to investigate the relationships between simulation errors and uncertainties in projections (e.g., Collins et al. 2012). Further cause for caution arises from the only test of reliability performed to date for a climate change, that of the Last Glacial Maximum (Hargreaves et al. 2011), which does not find any evidence of the ensemble being over-dispersed. In addition, a domed rank histogram may be also a consequence of tuning towards observations, in which case such weighting would amount to double-counting the data. These issues require further investigation so, at present, the most robust strategy may be to use the whole MME when using climate model ensembles for probabilistic prediction.

Acknowledgments We acknowledge the World Climate Research Programme’s Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Table 1 of this paper) for producing and making available their model output. For CMIP the US 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. M.C. was partially supported by funding from NERC grants NE/I006524/1 and NE/I022841/1. MW is supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101). T.Y., J.D.A., H.S., S.E., M.Y., J.C.H. were supported by the Global Environment Research Fund of the Ministry of the Environment of Japan (S-10, Integrated Climate Assessment – Risks, Uncertainties and Society, ICA-RUS). We thank anonymous reviewers for their constructive comments. We gratefully acknowledge Tamaki Yasuda and Osamu Arakawa for helping to get CMIP data.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

References

- Abe M, Shiogama H, Hargreaves JC, Annan JD, Nozawa T, Emori S (2009) Correlation between Inter-model similarities in spatial pattern for present and projected future mean. *Clim SOLA* 5:133–136. doi:10.2151/sola.2009-034
- Roeckner E et al (1996) The atmospheric general circulation model ECHAM4, MPI Report No. 218
- Roeckner E et al (2003) The atmospheric general circulation model ECHAM5 Report No. 349
- Marti O et al (2006) The new IPSL climate system model: IPSL-CM4. Scientific Note IPSL Pole Modeling, No. 26
- Bellouin N et al (2007) Improved representation of aerosols for HadGEM2, Meteorological Office Hadley Centre, Technical Note 73
- Trenberth KE et al (2007) Observations: surface and atmospheric climate change. In: Solomon et al (eds) *Climate change 2007: the physical science basis. Contribution of Working Group I to the fourth assessment report of the intergovernmental panel on climate change*. Cambridge University Press, Cambridge
- Collins WJ et al (2008) Evaluation of the HadGEM2 model, Meteorological Office Hadley Centre, Technical Note 74
- Williams DN et al (2011) The earth system grid federation: software framework supporting CMIP5 data analysis and dissemination. CLIVAR Exchanges, No. 56, International CLIVAR Project Office, Southampton, United Kingdom, 40–42
- Yukimoro S et al (2011) Technical report of the Meteorological Research Institute, 64, 83
- Annan JD, Hargreaves JC (2010) Reliability of the CMIP3 ensemble. *Geophys Res Lett* 37:L02703. doi:10.1029/2009GL041994
- Annan JD, Hargreaves JC (2011) Understanding the CMIP3 multi-model ensemble. *J Clim* 24:4529–4538. doi:10.1175/2011JCLI3873.1
- Annan JD, Hargreaves JC, Ohgaito R, Abe-Ouchi A, Emori S (2005a) Efficiently constraining climate sensitivity with ensembles of Paleoclimate simulations. *Sci On-line Lett Atmos* 1:181–184
- Annan JD, Hargreaves JC, Edwards NR, Marsh R (2005b) Parameter estimation in an intermediate complexity Earth system model using an ensemble Kalman filter. *Ocean Model* 8(1–2):135–154
- Bretherton CS, Windmann M, Dymnikov VP, Wallace JM, Blade I (1999) The effective number of spatial degrees of freedom of a time-varying field. *J Clim* 12:1990–2009
- Cess RD et al (1990) Intercomparison and interpretation of climate feedback processes in 19 atmospheric general circulation models. *J Geophys Res* 95:16601–16615
- Collins M, Tett SFB, Cooper C (2001) The internal climate variability of HadCM3, a version of the Hadley Centre coupled model without flux adjustments. *Clim Dyn* 17(1):61–81
- Collins WD, Rasch PJ, Boville BA, Hack JJ, McCaa JR, Williamson DL, Kiehl JT, Briegleb B, Bitz C, Lin S (2004) Description of the NCAR Community Atmosphere Model (CAM3.0), Technical Note TN-464 + STR, National Center for Atmospheric Research, Boulder, 214 pp
- Collins M, Booth BBB, Harris GR, Murphy JM, Sexton DMH, Webb MJ (2006a) Towards quantifying uncertainty in transient climate change. *Clim Dyn* 27:127–147
- Collins WD, Rasch PJ, Boville BA, Hack JJ, McCaa JR, Williamson DL, Briegleb BP, Bitz CM, Lin SJ, Zhang M (2006b) The formulation and atmospheric simulation of the Community Atmosphere Model version 3 (CAM3), *J. Climate* 19:2144–2161
- Collins M, Booth BBB, Harris GR, Murphy JM, Sexton DMH, Webb MJ (2010) Climate model errors, feedbacks and forcings: a comparison of perturbed physics and multi-model ensembles. *Clim Dyn*. doi:10.1007/s00382-010-0808-0
- Collins WJ et al (2011) Development and evaluation of an Earth-system model—HadGEM2. *Geosci Model Dev Discuss* 4:997–1062. doi:10.5194/gmdd-4-997-2011
- Collins M, Chandler RE, Cox PM, Huthnance JM, Rougier J, Stephenson DB (2012) Quantifying future climate change. *Nat Clim Change* 2:403–409. doi:10.1038/nclimate1414
- Delworth TL et al (2006) GFDL’s CM2 global coupled climate models—Part 1: formulation and simulation characteristics. *J Clim* 19:643–674
- K-1 Model Developers (2004) K-1 coupled GCM (MIROC) description. K-1 Tech. Rep. 1, University of Tokyo, 1–34

- Flato GM (2005) The third generation coupled global climate model (CGCM3), <http://www.ec.gc.ca/ccmac-ccema/default.asp?lang=En&n=1299529F-1>
- Gent PR et al (2011) The community climate system model version 4. *J Clim* 24:4973–4991. doi:10.1175/2011JCLI4083.1
- Gottelman A, Kay JE, Shell KM (2012) The evolution of climate sensitivity and climate feedbacks in the community atmosphere model. *J Clim* 25(5):1453–1469
- Gleckler PJ, Taylor KE, Doutriaux C (2008) Performance metrics for climate models. *J Geophys Res* 113:D06104. doi:10.1029/2007JD008972
- Gnanadesikan A et al (2006) GFDL's CM2 global coupled climate models—Part 2: the baseline ocean simulation. *J Clim* 19:675–697
- Gordon CC et al (2000) The simulation of SST, sea ice extents and ocean heat transport in a version of the Hadley Centre coupled model without flux adjustments. *Clim Dyn* 16:147–168
- Haak H et al (2003) Formation and propagation of great salinity anomalies. *Geophys Res Lett* 30:1473. doi:10.1029/2003GL17065
- Hargreaves JC, Paul A, Ohgaito R, Abe-Ouchi A, Annan JD (2011) Are paleoclimate model ensembles consistent with the MARGO data synthesis? *Clim Past* 7:917–933. doi:10.5194/cp-7-917-2011
- Jackson CS, Sen MK, Stoffa PL (2004) An efficient stochastic Bayesian approach to optimal parameter and uncertainty estimation for climate model predictions. *J Clim* 17:2828–2841
- Jackson CS, Sen MK, Huerta G, Deng Y, Bowman KP (2008) Error reduction and convergence in climate prediction. *J Clim* 21:6698–6709
- Johns TC et al (2006) The new Hadley Centre climate model HadGEM1: evaluation of coupled simulations. *J Clim* 19(7):1327–1353. doi:10.1175/JCLI3712.1
- Jolliffe I, Primo C (2008) Evaluating rank histograms using decompositions of the Chi square test statistic. *Mon Weath Rev* 136:2133–2139. doi:10.1175/2007MWR2219.1
- Jones CD et al (2011) The HadGEM2-ES implementation of CMIP5 centennial simulations. *Geosci Model Dev* 4:543–570. doi:10.5194/gmd-4-543-2011
- Klocke D, Pincus R, Quaas J (2011) On constraining estimates of climate sensitivity with present-day observations through model weighting. *J Clim* 24(23):6092–6099. doi:10.1175/2011JCLI4193.1
- Knutti R (2010) The end of model democracy? *Clim Change* 102:395–404. doi:10.1007/s10584-010-9800-2
- Legutke S, Maier-Reimer E (1999) Climatology of the HOPE-G Global Ocean General Circulation Model, DKRZ Techn. Report 21
- Marsland SJ, Haak H, Jungclaus JH, Latif M, Röske F (2003) The Max-Planck-Institute global ocean/sea ice model with orthogonal curvilinear coordinates. *Ocean Model* 5(2):91–127. doi:10.1016/S1463-5003(02)00015-X
- Martin GM, Dearden C, Greeves C, Hinton T, Inness P et al (2004) Evaluation of the atmospheric performance of HadGAM/GEM1, Hadley Centre Technical Note No. 54, Hadley Centre for Climate Prediction and Research/Met Office, Exeter
- Martin GM et al (2006) The physical properties of the atmosphere in the new Hadley Centre Global Environmental Model, HadGEM1—Part 1: model description and global climatology. *J Clim* 19(7):1274–1301. doi:10.1175/JCLI3636.1
- Martin GM et al (2011) The HadGEM2 family of Met Office unified model climate configurations. *Geosci Model Dev* 4:723–757. doi:10.5194/gmd-4-723-2011
- Masson D, Knutti R (2011) Climate model genealogy. *Geophys Res Lett* 38:L08703. doi:10.1029/2011GL046864
- McFarlane NA, Boer GJ, Blanchet J-P, Lazare M (1992) The Canadian climate centre second-generation general circulation model and its equilibrium climate. *J Clim* 5:1013–1044
- Meehl GA, Stocker T et al (2007) Global climate projections. I. Climate change 2007: the physical science basis. In: Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) Contribution of working Group I to the fourth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge
- Min et al (2004) Climatology and internal variability in a 1000-year control simulation with the coupled climate model ECHO-G, *Tellus A*
- Murphy JM, Sexton DMH, Barnett DN, Jones GS, Webb MJ, Collins M, Stainforth DA (2004) Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature* 430:768–772
- Murphy JM, Booth BBB, Collins M, Harris GR, Sexton D, Webb MJ (2007) A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. *Philos Trans R Soc Lond A* 365:1993–2028
- Pacanowski RC, Dixon K, Rosati A (1993) The GFDL modular ocean model users guide, Version 1.0. GFDL Ocean Group technical report No. 2, Geophysical Fluid Dynamics Laboratory, Princeton
- Pope VD, Gallani ML, Rowntree PR, Stratton RA (2000) The impact of new physical parametrizations in the Hadley Centre climate model-HadAM3. *Clim Dyn* 16:123–146
- Raddatz TJ et al (2007) Will the tropical land biosphere dominate the climate-carbon cycle feedback during the twenty first century? *Clim Dyn* 29:565–574. doi:10.1007/s00382-007-0247-8
- Ringer MA et al (2006) The physical properties of the atmosphere in the new Hadley Centre Global Environmental Model, HadGEM1—Part 2: aspects of variability and regional climate. *J Clim Am Meteorol Soc* 19(7):1302–1326. doi:10.1175/JCLI3713.1
- Roberts MJ (2004) The Ocean Component of HadGEM1. GMR Report Annex IV.D.3, Hadley Centre for Climate Prediction and Research/Met Office, Exeter
- Rotstayn LD, Collier MA, Dix MR, Feng Y, Gordon HB, Farrell SPO, Smith IN, Syktus J (2010) Improved simulation of Australian climate and ENSO-related climate variability in a GCM with an interactive aerosol treatment. *Int J Climatol* 30(7):1067–1088. doi:10.1002/joc.1952
- Sakaguchi K, Zeng X, Brunke MA (2012) The hindcast skill of the CMIP ensembles for the surface air temperature trend. *J Geophys Res* 117:D16113. doi:10.1029/2012JD017765
- Sakamoto TT, Komuro Y, Nishimura T, Ishii M, Tatebe H, Shiogama H, Hasegawa A, Toyoda T, Mori M, Suzuki T, Imada Y, Nozawa T, Takata K, Mochizuki T, Ogochi K, Emori S, Hasumi H, Kimoto M (2012) MIROC4h—a new high-resolution atmosphere-ocean coupled general circulation model. *J Met Soc Jpn* 90:325–359. doi:10.2151/jmsj.2012-301
- Salas-Méllia D, Chauvin F, Déqué M, Douville H, Guérémy JF, Marquet P, Planton S, Royer JF, Tyteca S (2005) Description and validation of the CNRM-CM3 global coupled model. CNRM technical report 103. Available from http://www.cnrn.meteo.fr/scenario2004/paper_cm3.pdf
- Sanderson BM (2011) A multi-model study of parametric uncertainty in predictions of climate response to rising greenhouse gas concentrations. *J Clim*. doi:10.1175/2010JCLI3498.1
- Sexton DMH, Murphy JM, Collins M, Webb MJ (2012) Multivariate probabilistic projections using imperfect climate models Part I: outline of methodology. *Clim Dyn* 38:2513–2542. doi:10.1007/s00382-011-1208-9
- Shibata K, Yoshimura H, Oizumi M, Hosaka M, Sugi M (1999) A simulation of troposphere, stratosphere and mesosphere with an MRI/JMA98 GCM. *Pap Meteorol Geophys* 50:15–53
- Shiogama H, Emori S, Hanasaki N, Abe M, Masutomi Y, Takahashi K, Nozawa T (2011) Observational constraints indicate risk of drying in the Amazon basin. *Nat* 684 Commun, 2, Article No 253

- Shiogama H, Watanabe M, Yoshimori M, Yokohata T, Ogura T, Annan JD, Hargreaves JC, Abe M, Kamae Y, O'ishi R, Nobui R, Emori S, Nozawa T, Abe-Ouchi A, Kimoto M (2012) Perturbed physics ensemble using the MIROC5 coupled atmosphere-ocean GCM without flux corrections: experimental design and results. *Clim Dyn* 39:3041–3056. doi:[10.1007/s00382-012-1441-x](https://doi.org/10.1007/s00382-012-1441-x)
- Smith RD, Gent PR (2004) Reference manual for the Parallel Ocean Program (POP), Ocean component of the Community Climate System Model (CCSM2.0 and 3.0). Technical Report LA-UR-02-2484, Los Alamos National Laboratory, Los Alamos
- Smith DM, Cusack S, Colman AW, Folland CK, Harris GR, Murphy JM (2007) Improved surface temperature prediction for the coming decade from a global climate model. *Science* 317:796–799. doi:[10.1126/science.1139540](https://doi.org/10.1126/science.1139540)
- Smith DM, Eade R, Dunstone NJ, Fereday D, Murphy JM, Pohlman H, Scaife AA (2010) Skilful multi-year predictions of Atlantic hurricane frequency. *Nat Geosci* 3:846–849. doi:[10.1038/ngeo1004](https://doi.org/10.1038/ngeo1004)
- Stainforth DA et al (2005) Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature* 433:403–406
- Stouffer RJ et al (2006) GFDL's CM2 global coupled climate models—Part 4: idealized climate response. *J Clim* 19:723–740
- Tatebe H, Ishii M, Mochizuki T, Chikamoto Y, Sakamoto TT, Komuro Y, Mori M, Yasunaka S, Watanabe M, Ogochi K, Suzuki T, Nishimura T, Kimoto M (2012) The Initialization of the MIROC climate models with hydrographic data assimilation for decadal prediction. *J Met Soc Jpn* 90:275–294. doi:[10.2151/jmsj.2012-A14](https://doi.org/10.2151/jmsj.2012-A14)
- Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. *Bull Am Meteor Soc* 93:485–498. doi:[10.1175/BAMS-D-11-00094.1](https://doi.org/10.1175/BAMS-D-11-00094.1)
- Volodin E, Dianskii NA, Gusev AV (2010) Simulating presentday climate with the INMCM4.0 coupled model of the atmospheric and oceanic general circulations. *Izv Atmos Ocean Phys* 46(4):414–431. doi:[10.1134/S000143381004002X](https://doi.org/10.1134/S000143381004002X)
- Washington WM, Weatherly JM, Meehl GA, Semtner AJJ, Bettge TW, Craig AP, Strand WG, Arblaster J, Wayland VB, James R (2000) Parallel climate model (PCM) control and transient simulations. *Clim Dyn* 16:755–774
- Watanabe M et al (2010) Improved climate simulation by MIROC5: mean states, variability, and climate sensitivity. *J Clim* 23:6312–6335. doi:[10.1175/2010JCLI3679.1](https://doi.org/10.1175/2010JCLI3679.1)
- Watanabe S et al (2011) MIROC-ESM: model description and basic results of CMIP5-20c3m experiments. *Geosci Model Dev Discuss* 4:1063–1128. doi:[10.5194/gmdd-4-1063-2011](https://doi.org/10.5194/gmdd-4-1063-2011)
- Watanabe M, Shiogama H, Yokohata T, Kamae Y, Yoshimori M, Ogura T, Annan JD, Hargreaves JC, Emori S, Kimoto M (2012) Using a multi-physics ensemble for exploring diversity in cloud shortwave feedback in GCMs. *J Clim* 25:5416–5431. doi:[10.1175/JCLI-D-11-00564.1](https://doi.org/10.1175/JCLI-D-11-00564.1)
- Webb MJ et al (2006) On the contribution of local feedback mechanisms to the range of climate sensitivity in two GCM ensembles. *Clim Dyn* 27:17–38
- Wilks DS (2004) The minimum spanning tree histogram as a verification tool for multidimensional ensemble forecasts. *Mon Wea Rev* 132:1329–1340
- Wittenberg AT et al (2006) GFDL's CM2 global coupled climate models—Part 3: tropical Pacific climate and ENSO. *J Clim* 19:698–722
- Yokohata T et al (2008) Comparison of equilibrium and transient responses to CO₂ increase in eight state-of-the-art climate models. *Tellus* 60:946–961
- Yokohata T, Webb MJ, Collins M, Williams KD, Yoshimori M, Hargreaves JC, Annan JD (2010) Structural similarities and differences in climate responses to CO₂ increase between two perturbed physics ensembles. *J Clim* 23(6):1392–1410
- Yokohata T, Annan JD, Collins M, Jackson CS, Tobis M, Hargreaves JC (2012) Reliability of multi-model and structurally different single-model ensembles. *Clim Dyn*. doi:[10.1007/s00382-011-1203-1](https://doi.org/10.1007/s00382-011-1203-1)
- Yu Y, Yu R, Zhang X, Liu H (2002) A flexible global coupled climate model. *Adv Atmos Sci* 19:169–190
- Yu Y, Zhang X, Guo Y (2004) Global coupled ocean-atmosphere general circulation models in LASG/IAP. *Adv Atmos Sci* 21:444–455
- Yukimoto S, Noda A, Kitoh A, Sugi M, Kitamura Y et al (2001) The new Meteorological Research Institute global ocean-atmosphere coupled GCM (MRI-CGCM2)-Model climate and variability. *Pap Meteorol Geophys* 51:47–88