Clim Dyn
DOI 10.1007/s00382-011-1203-1

Reliability of multi-model and structurally different single-model ensembles

Tokuta Yokohata · James D. Annan ·
Matthew Collins · Charles S. Jackson ·
Michael Tobis · Mark J. Webb · Julia C. Hargreaves

Received: 30 March 2011 / Accepted: 18 September 2011

Abstract The performance of several state-of-the-art climate model ensembles, including two multi-model ensembles (MMEs) and four structurally different (perturbed parameter) single model ensembles (SMEs), are investigated for the first time using the rank histogram approach. In this method, the reliability of a model ensemble is evaluated from the point of view of whether the observations can be regarded as being sampled from the ensemble. Our analysis reveals that, in the MMEs, the climate variables we investigated are broadly reliable on the global scale, with a tendency towards overdispersion. On the other hand, in the SMEs, the reliability differs depending on the ensemble and variable field considered. In general, the mean state and historical trend of surface air temperature, and mean state of precipitation are reliable in the SMEs. However, variables such as sea level pressure or top-of-atmosphere clear-sky shortwave radiation do not cover a sufficiently wide range in some. It is not possible to assess whether this is a fundamental feature of SMEs generated with particular model, or a consequence of the algorithm used to select and perturb the values of the parameters. As under-dispersion is a potentially more serious issue when using ensembles to make projections, we recommend the application of rank histograms to assess reliability when designing and running perturbed physics SMEs.

Keywords Climate model · Future climate prediction · Multi-model ensemble · Perturbed physics ensemble · Reliability · Rank histogram

T. Yokohata (✉)
Center for Global Environmental Research, National Institute
for Environmental Studies, 16-2 Onogawa, Tsukuba,
Ibaraki, 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, 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 · M. Tobis
Institute of Geophysics, The University of Texas at Austin,
10100 Burnet Rd., ROC-196, Mail Code R2200,
Austin, TX 78758, USA

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

1 Introduction

In order for society to efficiently mitigate and adapt to climate change, it is necessary to have climate projections accompanied by assessments of the uncertainty in the projections. Ensembles of climate models, sampling uncertainties in model formulation, are commonly used as the basis for generation of probabilistic projections. It is, therefore, very important to evaluate the performance of these ensembles.

There are a large number of methods one could adopt to evaluate the performance of model ensembles and there are many examples in the literature. These methods generally use one of two paradigms. One paradigm, sometimes called the truth centred paradigm (Knutti et al. 2010b), assumes that the truth should be close to the centre of the ensemble members (i.e. close to the ensemble mean). Knutti et al. (2010a) investigated the behaviour of the state-of-the-art climate model ensemble created by the World Climate

Published online: 12 October 2011

Research Programme's Coupled Model Intercomparison Project Phase 3 (CMIP3, Meehl et al. 2007), and found that the truth centred paradigm is incompatible with the CMIP3 ensemble: the ensemble mean does not converge to observations as the number of ensemble members increases, and the pairwise correlation of model errors (the differences between model and observation) between two ensemble members does not average to zero (Knutti et al. 2010a; Annan and Hargreaves 2010; hereafter AH10).

An alternative paradigm is to consider the truth as being drawn from the distribution sampled by the ensemble. In this case, the model ensemble can be regarded as perfect if the ensemble members and the truth are "statistically indistinguishable". In this case, the truth is not necessarily at the centre of the ensemble. Predictions made with such a model ensemble are regarded as "reliable" in the technical sense that the natural probabilistic interpretation (based on counting ensemble members) matches the frequency of occurrence of predicted events evaluated over multiple verifications. This idea of a "statistically indistinguishable" ensemble is common in the field of weather forecasting and other ensemble prediction fields, and under this paradigm the reliability of model ensembles can be evaluated through the rank histogram approach (Anderson 1996) whereby the distribution of the observed occurrence of an event in the prediction ensembles is evaluated. Such an analysis can reveal if prediction ensembles are too narrow, too broad, or biased. In the present paper, we analyse the reliability of model ensembles in statistical terms. We discuss the concept of "reliability" in more detail in Sect. 2.3.

AH10 applied the rank histogram method to the evaluation of spatial fields of time-averaged present-day variables from climate models and concluded that the CMIP3 ensemble appears reasonably reliable on large scales. However, AH10 only investigated the CMIP3 ensemble, and the three most commonly investigated climate variables [surface air temperature (SAT), sea level pressure (SLP), and precipitation (PRCP)]. They did not consider those variables which play an important role in determining the range of climate responses to increasing greenhouse gases, such as radiation and/or cloud effects at the top of atmosphere (TOA). Here we extend the evaluation to those variables and analyse several ensembles; two multi-model ensembles (MMEs) from CMIP3 and four structurally different single model ensembles (SMEs, sometimes also referred to a perturbed physics or perturbed parameter ensembles) with different ranges of climate sensitivity. We investigate the relationship between climate sensitivity and the reliability of the present-day climate simulation. We also check the validity of the rank histogram approach by comparing the model-data difference with the ensemble spread through calculating the root mean square model-data difference (RMSE), and the standard deviation of the ensemble (SD).

In Sect. 2, we describe the model ensembles and the application of the rank histogram approach, including a description of the statistical method used to define the reliability of model ensembles from the rank histogram, and a method for handling uncertainties in the observations. In Sect. 3, the results from the rank histogram analyses are described. In this study, we primarily investigate the reliability of the climatology (long-term mean of model simulation) of large-scale features of climate model ensembles, but we also consider the trend for surface air temperature where transient simulations are available (that is, for the coupled ocean–atmosphere models). Our main result is to show that, under this analysis, the performance of the MME is qualitatively different from, and superior to, the SMEs. A conventional analysis of RMSE and SD is also presented in Sect. 3, which supports our results and analysis using rank histograms. Finally, in Sect. 4, we present our conclusions.

2 Model ensembles and methods of analysis

2.1 Methods for the generation of ensembles

There are two qualitatively distinct methods in widespread use which aim to sample uncertainties arising from model parameterisations (Murphy et al. 2007).

One approach is to use an MME, which consists of simulations contributed by different models of climate research institutes from around the world, often referred to as an "ensemble of opportunity". Each model may be considered a social construct which embodies the beliefs of those modellers who created it as how best to represent the climate system, within the computational and technological constraints at the time. Thus the whole ensemble may be interpreted (at least potentially) as sampling our collective beliefs and uncertainties regarding the climate system, although the ad-hoc and uncoordinated nature of the model-building process around the world may raise some doubts as to the plausibility of such an assumption. The current MME of state-of-the-art global climate models is the CMIP3 ensemble (Meehl et al. 2007). While this ensemble samples uncertainties in model structure, each model has one parameter set and a fixed model structure. Some members of the MME may be different resolution versions of the same model structure (albeit with resolution-dependent parameters adjusted) and some others may share common components.

The other commonly-used approach is to choose a range of different parameter values in a single model, to form an SME. While uncertainties within a single model can be more systematically investigated in SMEs, and parameter values may be set to rather extreme values (compared to those used in the MME) so as to generate a wide range of responses, major uncertainties in model structure cannot be sampled other than switching between existing alternative parameterisation subroutines. Furthermore, different SMEs may use different strategies for varying parameters values such that, even using the same single model structure, different ensembles can show quite different behaviour (Collins et al. 2010, hereafter C10, Yokohata et al. 2010, hereafter Y10). It is therefore not clear a priori to what extent either the ad-hoc multi-model "ensemble of opportunity", or the more explicitly designed (but structurally limited) single model ensembles, can be considered to provide a realistic probabilistic range of future climate change, and methods to evaluate the performance of these ensembles are not yet well developed.

Some comparisons of features between the CMIP3 MME and SMEs have already been performed. Webb et al. 2006 showed that the spread of cloud feedback of the CMIP3 MME overlaps with that of an SME of the Hadley Centre atmosphere-slab ocean coupled model HadSM3 (Pope et al. 2000). A recent study by C10 analysed a number of different climate variables in a set of SMEs of HadCM3 (Gordon et al. 2000, atmosphere–ocean coupled version of HadSM3) from the point of view of global-scale model errors and climate change forcings and feedbacks, and compared them with variables derived from the CMIP3 MME. Knutti et al. (2006) examined another SME based on the HadSM3 model, and found a strong relationship between the magnitude of the seasonal cycle and climate sensitivity, which was not reproduced in the CMIP3 ensemble.

However, comparisons of SMEs based on different underlying models have not yet been examined so extensively. This may be partly because outputs from SMEs have not, to date, been archived in open databases like the CMIP3 MME. Since a single model is used for constructing an SME, the results depend on the model used and on the parameter sampling strategy. For example, climate sensitivity (equilibrium surface air temperature change due to CO_2 doubling) of different model SMEs differs substantially. The range of climate sensitivity obtained by a HadSM3 SME is similar to that of the CMIP3 MME, at about 2–5 K (Webb et al. 2006, Randall et al. 2007), while that of another SME, using MIROC (K-1 Model Developers 2004) is relatively high (about 4–10 K, Annan et al. 2005a), and that of NCAR CAM (Collins et al. 2006b) is relatively low (about 2–3 K, Jackson et al. 2008; Sanderson 2011). Even ensembles produced with the same model but with different sampling strategies can produce different distributions of climate sensitivity (Murphy et al. 2004; Stainforth et al. 2005).

Recently Y10 investigated the physical processes involved in determining the climate sensitivity of the structurally different SMEs of HadSM3 and MIROC3, and found that while shortwave (SW) cloud feedback plays an important role for the difference in the ensemble mean and the spread of climate sensitivity in the two SMEs, the mechanisms which determine the spread in shortwave cloud feedback might be different between the two SMEs. However, Y10 and other studies so far have not directly evaluated the reliability of structurally different SMEs compared to the MME.

2.2 Model ensembles

For the MMEs, we use results from CMIP3 (Meehl et al. 2007), obtained from the Program for Climate Model Diagnosis and Intercomparison (PCMDI) data archive. In the present study, we analyse two MMEs, made from two subsets of the CMIP3 database. One is the CMIP3 MME of the twentieth century experiments using coupled atmosphere–ocean climate models (CMIP3-AO). The climate variables from CMIP3-AO are averaged over the period during which observational data are available (details of the observations are presented in Sect. 2.3). In CMIP3-AO we use only one ensemble member from each model, in order not to give special weight to particular models where more than one initial-condition ensemble member exists (Knutti et al. 2010b). The other MME consists of the control experiments (the boundary conditions of the model are held fixed at values comparable to the present climate state) from the atmosphere-slab ocean coupled climate models (CMIP3-AS). In an atmosphere-slab ocean coupled model, the ocean heat transport in the slab ocean (which is the representation of the mixed layer ocean with depths around 50 m) is diagnosed from a calibration phase with imposed observed sea surface temperature and sea ice distributions. We use the last 20 years average of the control simulation of the CMIP3-AS. The climate models used in the analysis are summarised in Table 1.

We use four SMEs constructed using structurally different climate models. The experimental settings and the number of parameters that were varied in the four SMEs are summarised in Table 2. Two of the SMEs used for analysis were generated by varying atmospheric parameters in the closely related models, HadCM3 (Gordon et al. 2000) and HadSM3 (Pope et al. 2000). We denote these HadCM3-AO and HadSM-AS, respectively. The atmospheric components of HadCM3 and HadSM3 are identical, and have a 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 as described above for each member. The HadCM3-AO and HadSM-AS SMEs were generated

Table 1 List of the CMIP3 multi-model ensemble (MME) used for analysis

Model	CMIP3-AO		CMIP3-AS		References
	A	R	A	R	
CCSM3	○	○	○	○	Collins et al. (2006b), Smith and Gent (2004)
CGCM3.1-T47	○	○	○	○	McFarlane et al. (1992), Flato (2005)
CGCM3.1-T63	○	○	○	○	
CNRM-CM3	○	○			Salas-Melia et al. (unpublished data)
CSIRO-Mk3.0	○				Gordon et al. (2002)
ECHAM5/MPI-OM	○	○	○	○	Roeckner et al. (2003), Marsland et al. (2003), Haak et al. (2003)
ECHO-G	○	○	○		Roeckner et al. (1996), Legutke and Maier-Reimer (1999), Min et al. (2005)
FGOALS-g1.0	○	○	○		Yu et al. (2002, 2004)
GFDL-CM2.0	○	○	○	○	Delworth et al. (2006), Gnanadesikan et al. (2006), Wittenberg et al. (2006), Stouffer et al. (2006)
GFDL-CM2.1	○	○	○	○	
GISS-EH	○				Schmidt et al. (2006), Hansen et al. (2007)
GISS-ER	○				
INM-CM3.0	○				Galin et al. (2003), Diansky et al. (2002)
IPSL-CM4	○	○			Marti et al. (2006)
MIROC3-HiRes	○	○	○	○	K-1 Model Developers (2004)
MIROC3-MedRes	○	○	○	○	
MRI-CGCM	○	○	○	○	Shibata et al. (1999), Yukimoto et al. (2001)
PCM	○	○			Washington et al. (2000)
UKMO-HadCM3	○	○			Gordon et al. (2000), Pope et al. (2000)
UKMO-HadGEM1	○	○	○	○	Martin et al. (2004), Roberts (2004)

"CMIP3-AO" denotes twentieth century experiment by atmosphere–ocean coupled general circulation model (GCM), and "CMIP3-AS" denotes control experiment by atmosphere-slab ocean coupled GCM. "R" denotes climate variables related to radiation, such as SW and LW full-sky radiation, cloud radiative forcing, and clear-sky radiation. The *circle* indicates that the climate variable of each model member is available. Number of models used for analysis is different between model ensembles and climate variables

by the Quantifying Uncertainty in Model Predictions (QUMP) project.

HadSM-AS is made up of the pre-industrial control experiments of the 128 ensemble members used in Webb et al. 2006 and Y10, which is the same as "S-PPE-M" in C10. In HadSM-AS, 31 atmospheric parameters are perturbed at the same time. Originally the number of ensembles was 129, but one ensemble member is excluded because of the unrealistic cooling drift, caused by the interaction between negative SST anomalies and low cloud cover which is known to sometimes occur in models of this type (e.g. Stainforth et al. 2005, supplementary information).

HadCM3-AO has 17 members and we utilise the results from the twentieth century experiments forced by both natural and anthropogenic factors. This is the same

Table 2 List of the single model ensembles (or perturbed physics ensembles) used for analysis

Ensemble	Experiment	Number of parameter perturbed	Number of ensemble members	References
HadCM3-AO	Twentieth century by AOGCM	31	128	Murphy et al. (2007), Collins et al. (2006a)
HadSM3-AS	Control by ASGCM	31	17	Webb et al. (2006), Yokohata et al. (2010)
MIROC3-AS	Control by ASGCM	13	32	Annan et al. (2005a), Yokohata et al. (2010),
NCAR-A	Control by AGCM	15	100	Jackson (2009)

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 ensemble as “AO-PPE-A” in C10 in which 31 atmospheric parameters in HadCM3 are simultaneously perturbed. In order to sample a range of climate sensitivity while ensuring that each ensemble member simulates the present climate state reasonably well, the parameter settings of 17 model ensemble members were chosen based on analysis of the HadSM-AS runs (see Webb et al. 2006 and C10). The parameters that were varied in the creation of these two SMEs are described in Murphy et al. 2004 and Y10.

The third SME analysed is the ensemble produced using MIROC3.2 (K-1 Model Developers 2004) which we refer to as MIROC3-AS. The atmospheric component of MIROC3.2 used for the ensembles has reduced resolution compared to the standard T42 resolution, of 5.6 longitudinal degree by 5.6 latitudinal degree (T21) with 20 vertical levels. This is coupled to a motionless 50 m depth slab ocean. Ocean heat transport is calculated with the slab-ocean calibration procedure described above.

The MIROC3-AS SME is generated using an Ensemble Kalman Filter (EnKF) method for parameter estimation (Annan et al. 2005b). The EnKF is used to assimilate seasonally averaged observational data into the model, thereby generating an ensemble of runs with a range of values for the uncertain parameters, all reasonably compatible with present-day climatology. The number of parameters perturbed is 13. See Annan et al. (2005a) and Y10 for further details of the application of the EnKF to MIROC. The number of ensemble members generated was 40, but some ensemble members that exhibit a persistent warming drift and do not reach steady states during their doubled CO_2 experiment are excluded from the analysis in the same way as Y10. We use 32 members without the warming drift for the analysis. The numerical experiments are performed within the Japan Uncertainty Modelling Project (JUMP).

The fourth SME is constructed from the atmospheric GCM, NCAR CAM3.1 (Collins et al. 2006b, NCAR-A hereafter). It was generated by varying 15 model parameters important to clouds, convection, and radiation. One hundred samples from a 2,276-member ensemble were selected to represent observational constraints on the model’s parametric uncertainties.

This implementation of CAM3.1 has a resolution of 2.8 degree longitude by 2.8 degree latitude (T42) with 26 vertical levels. The acceptable model parameter settings are chosen using Bayesian inference and Multiple Very Fast Simulated Annealing (Jackson et al. 2004). The experimental design follows Jackson et al. (2008) with some differences relating to shorter model experiments (4 years instead of 11 years), an expanded list of uncertain model parameters, and a revised cost function. The updated cost function is based on quantities, observations, and regions that are currently being used evaluate the development of CAM through a set of Taylor diagram diagnostics within the Atmosphere Model Diagnostic Package (http://www.cgd.ucar.edu/cms/diagnostics/). These metrics emphasise fields between 30S and 30N including 2 m air temperature (Willmott and Matsuura 2000), vertically averaged air temperature (ERA40, Uppala et al. 2005), latent heat fluxes of the ocean (Yu et al. 2008), zonal winds at 300 mb (ERA40, Uppala et al. 2005), longwave and shortwave cloud forcing (CERES2, Loeb et al. 2009), precipitation over land and ocean (GPCP, Adler et al. 2003), sea level pressure (ERA40, Uppala et al. 2005), vertically averaged relative humidity (ERA40, Uppala et al. 2005). Other quantities include Pacific Ocean wind stress between 5S and 5N (ERS-2, CERSAT 1996) and the global mean annual mean radiative balance.

Note that one of the key features of an SME is control over the experimental design of the ensemble via the algorithm used to perturb the parameters. For example, C10 show results from different experimental approaches involving perturbing model parameters one-at-a-time, incorporating information from observations to produce members which evaluate well against observations and exploring parameter space comprehensively in order to fit statistical emulators. Hence the behaviour of the SME can, to a certain extent, be controlled by experimental design. Different sections of code controlling the same process can even be switched in and out. One might envisage perturbing parameters in a way to maximise the “reliability” defined in this study. However, suffice to say that none of the SMEs analysed here have been designed in that way and all use different approaches to choose and perturb parameters with the general goal of generating reasonable climate states while representing parametric uncertainties.

2.3 Reliability and rank histogram analysis

The term “reliability” is used here in the technical sense analogous to how it is commonly used in numerical weather prediction (NWP). In principle, a probabilistic prediction is termed reliable when the frequency of occurrence (over a large set of predictions) matches the predicted probability (Toth et al. 2003). In NWP applications, forecasts are typically evaluated over a data set with both spatial and temporal dimensions. For example, Jolliffe and Primo (2008) (hereafter JP08) used a data set which they estimated, after adjusting for temporal and spatial correlations, to have approximately 17 temporal and 25 spatial degrees of freedom. In our current context of climate model evaluation, we are only using the climatological mean state (as is widespread in the evaluation of climate model ensembles, e.g. Knutti et al. 2010a) and the long-term trend pattern, and therefore there is no temporal dimension to our data set. Investigations using other epochs to look at climatic changes in response to external forcing will be reported elsewhere (Hargreaves et al. submitted to Climate of the Past). Furthermore, since the data are historical, the analysis here is essentially that of a hindcast, and it is debatable to what extent the data can be considered to provide truly independent validation of the models. The relationship between current performance and true forecasting (such as prediction of climate change over the twenty-first century) remains unclear. Model performance in comparison to historical or current data is often assumed, or asserted, to give some guide to future performance. Results robustly demonstrating this, however, even in the limited context of cross-validation within the multi-model ensemble, are relatively rare. Hall and Qu (2006) provide one example where there does appear to be a strong relationship between current performance and future climate change across the multi-model ensemble, but contrasting results exist. For example in the case of Knutti et al. (2006), a strong relationship between current behaviour and equilibrium climate sensitivity, that is found to hold across a single model ensemble, has no skill in predicting the climate sensitivity of the members of the CMIP3 ensemble. Thus, reliability over a hindcast interval is not necessarily a sufficient condition to demonstrate that the model forecasts are probabilistically valid. On the other hand, to the extent that such reliability can be demonstrated (even in a hindcast situation), it must be considered a positive indication in judging whether the range of uncertainty sampled by an ensemble provides a plausible range of depictions of the climate system. Conversely, where an ensemble is not reliable in this sense (and especially when it is strongly biased such that reality lies outside its range), it must raise some doubts as to how credible it is (at least in raw form) as a representation of these uncertainties.

The reliability of the gridded mean climatic state of the model ensembles was investigated for the modern climate with respect to the observational data sets of various variables by calculating rank histograms, using the same method described in AH10. The nine climate variables analysed were surface air temperature (SAT), sea level pressure (SLP), precipitation (PRCP), the TOA shortwave (SW) and longwave (LW) full-sky radiation, clear-sky radiation (CLR, radiative flux where clouds do not exists), and cloud radiative forcing (CRF, radiative effect by clouds diagnosed from the difference between full-sky and clear-sky radiation, Cess et al. 1990). In this study, we consider uncertainties in observation by using two independent dataset as shown in Table 3. As for the mean states of SAT, PRCP and SLP, we used 20-year climatology (1980–1999) obtained from the standard datasets such as HadCRU3 (Brohan et al. 2006), ERA40 (Uppala et al. 2005), GPCP (Adler et al. 2003), CMAP (Xie and Arkin 1997), and HadSLP2 (Allan and Ansell 2006). As for the TOA radiation, we used ERBE-S9 (Harrison et al. 1990) and ISCCP-FD (Zhang et al. 2004) dataset, which are the standard dataset often used for the validation of model radiative properties (Trenberth et al. 2007). Because of the availability of ERBE-S9, we used 5-year climatology (1985–1989) for the evaluation of TOA radiation.

In addition to the mean climate states, we evaluated the long-term trend in the twentieth century experiments by 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). In the present study, we do not investigate the twentieth century trend of PRCP, SLP, or TOA radiations. This is partly because the inter-annual to decadal variability is generally large in these variables, and partly because there are large uncertainties and sometimes an artificial trend in observations owing to the difficulty in measurement of these climate variables (Trenberth et al. 2007). As for the SAT trend, we also performed the same calculations after removing the natural variability using a method proposed by Thompson et al. 2008, but the difference between with and without removing natural variability is very small (not shown). Therefore, we believe that our results of SAT trend are robust.

The methodology of the rank histogram calculation was as follows. First, the model data and observational data were interpolated onto a common grid (resolution of T42 in CMIP3-AO and HadCM3-AO, and T21 for the other model ensembles). Second, we perturb the model ensemble to account for the observational uncertainties, as described below in Sect. 2.4. Then, at each grid point, we compared the value of the observation with the ensemble of model values at each grid point, 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 (the average weight will be $1/n_{\mathrm{grid}}$, where n_{grid} is number of grid point), over the whole grid. Note that in the present study we performed an univariate analysis where only one variable is used for the calculation of one rank histogram. A multivariate analysis where multiple variables are used for one rank histogram is an important future work.

The features of the rank histogram can be interpreted as follows. If a model ensemble is perfect, that is, if the true climatic variable can be regarded as indistinguishable from a sample of the model ensemble, then the rank of the

Table 3 List of observations used for analysis. In each variable, we use two independent observations (observation 1 and 2), and consider their uncertainties as explained in Sect. 2.4

Variables	Observation 1		Observation 2	
	Data	Reference	Data	Reference
Surface air temperature mean state, 40-years trend (1960–1999)	HadCRU3	Brohan et al. (2006)	ERA40	Uppala et al. (2005)
Precipitation	GPCP	Adler et al. (2003)	CMAP	Xie and Arkin (1997)
Sea level pressure	HadSLP2	Allan and Ansell (2006)	ERA40	Uppala et al. (2005)
SW and LW full-sky and clear-sky radiation, and cloud radiative forcing	ERBE-S9	Harrison et al. (1990)	ISCCP-FD	Zhang et al. (2004)

observation lies with equal probability anywhere in the model ensemble (after accounting for observational error), and thus the rank histogram should have a uniform distribution. On the other hand, if the distribution of a model ensemble is relatively narrow, 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 V- or even U-shaped distribution with large end bins depending on the severity of this error. 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.

While a uniform rank histogram is a necessary condition for an ensemble to be reliable, it is not in itself a sufficient one (Hamill 2001). The rank histogram approach, as applied here over spatial fields of nine time-averaged climate variables and one trend, represents a leading-order diagnostic of the behaviour of an ensemble. It does not replace detailed investigation of model errors in the mean and in the natural variability and their causes, nor (as discussed in Sect. 2.3) does it necessarily imply that future projections made using these ensembles will have the same reliability characteristics. We regard it simply as another tool in the armoury of those who develop and those who use climate models in research and in decision-making. The implications for future projections are discussed in the Conclusions. We also note that in focussing on reliability we are only considering one aspect of ensemble performance. For example, another property of ensembles that is generally of interest is sharpness, or in other words the narrowness of the ensemble spread. Subject to it being reliable, a sharper ensemble will be more informative than a broader one, but in practice there is often a tension between these two properties since narrowing an ensemble will generally increase the risk of reality falling outside its range.

2.4 Uncertainties in the observations

As mentioned above, here we incorporate consideration of the observational uncertainty into the rank histogram calculation. Uncertainty due to instrument error or analysis errors has often been ignored in ensemble evaluation, but has been identified as a potentially important factor (Knutti et al. 2010b). A simple technique to account for observational error is to add perturbations of equivalent size to the model outputs (e.g. Anderson 1996). 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. The lack of formal estimates of observational uncertainty is a hindrance, however. We estimate the observational errors by using two different observational data sets for each climatic variable in the rank histogram analyses (see Table 3). For each grid point, the observational value X_{obs}, which is compared to the model ensembles and is used for the calculation of rank of observation, is calculated as the mean of the two observations, X_{obs1} and X_{obs2},

$$X_{\mathrm{obs}} = (X_{\mathrm{obs1}} + X_{\mathrm{obs2}})/2 \qquad (1)$$

The standard deviation of the mean of two observations, σ_{obs}, is estimated as follows.

$$\sigma_{\mathrm{obs}} = \sqrt{\frac{1}{2}\left[(X_{\mathrm{obs1}} - X_{\mathrm{obs}})^2 + (X_{\mathrm{obs2}} - X_{\mathrm{obs}})^2\right]} \qquad (2)$$

Both X_{obs} and σ_{obs} are calculated at each grid point. Given the limited data with which they were estimated, it may in principle be better to spatially smooth the uncertainties in some way but we did not attempt this here. There are some strong spatial patterns in the uncertainties so a simple global average would probably not be appropriate. Using randomly sampled values from a normalised Gaussian distribution, Z, we add observational uncertainty to the model ensemble variables by

$$X'_{\mathrm{model}} = X_{\mathrm{model}} + \sigma_{\mathrm{obs}} Z \qquad (3)$$

where X_{model} is the values of each model ensembles at each grid point. Although, in principle, this approach introduces a degree of sampling error into our analysis, in practice the number of grid points (and therefore random deviates) is sufficiently large that the results are robust under replication. By considering uncertainties in observation, the spread of the model ensembles becomes somewhat wider compared to the case without considering them. However, the effect is small except where the observational uncertainty is comparable to the ensemble spread. Our approach to estimating observational uncertainty likely underestimates the true error, as some sources of error may be common to the different data sets used. However, we consider that this approach is certainly more defensible than the common practice of ignoring observational error entirely. In theory sampling the random deviates from a t-distribution would be preferable to the Gaussian which we used, but it is not clear how best to estimate the number of degrees of freedom of this distribution. Our results are robust even to the use of a t-distribution with a few as 5 degrees of freedom, which is surely an underestimate given the significant spatial coherence of observational errors. More credible and detailed statistical models of observational uncertainty would be valuable in undertaking more precise evaluations of climate models.

2.5 Statistical analysis by goodness-of-fit test

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. While the Chi-square test of goodness of fit is a standard technique for the test of uniformity, it is not sensitive to the order of the distribution and thus it is not well suited for our purposes, having low power in detecting typical failure modes (JP08). Therefore, we use the technique introduced by JP08 and decompose the Chi-square statistics 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, $h(i)$ defined in Sect. 2.3, the Chi-square statistics can be described as

$$T = \sum_{i=1}^{k} \frac{[n_{\mathrm{obs}} h(i) - e_i]^2}{e_i}, \tag{4}$$

where k is the maximum rank and i is the rank of the observation, n_{obs} is “the number of observation” in JP08, and $e_i = n_{\mathrm{obs}}/k$ corresponds to the expected bin value for a uniform distribution. The number n_{obs} is also referred to as the “effective degrees of freedom of the data” in AH10 and JP08. Since values of neighbouring grid points are highly correlated, as discussed in AH10, their ranks of observation cannot be considered as independent of each other. The effective degree of freedom of the data, n_{obs}, which also corresponds to the independent number of ranks of observation in the global map of $R(l, m)$, is not entirely clear. AH10 followed JP08 in using a value of 40, which corresponds to the effective degree of freedom of synoptic climate fields. However, in that work, calibrating the statistical test through permutation testing (i.e, taking each model in turn as the target to be predicted by the remaining ensemble members) suggested that a value of around 5 degrees of freedom might be more appropriate. In a recent study, Annan and Hargreaves 2011 also estimated via EOF analysis that n_{obs} ranges from 4 to 11 for SAT, SLP and PPT. Permutation testing and EOF analysis of the ensembles used here (not shown) support a similar range of values, so we use 10 here as an approximate (but perhaps slightly high) estimate. This number must be considered as somewhat uncertain, but our results are qualitatively insensitive to the exact value used. The appropriate value to use may differ across variables, but data for estimating this are limited. A higher value implies a test with more power, making the test for reliability more stringent, and meaning that more rank histograms would be detected as significantly non-uniform at a given threshold.

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 of with $(k - 1)$ degrees of freedom. Using a table of the Chi-square distribution and the value of T in Eq. 4, 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. The Chi-square approximation is accurate in the case of a large data set. Here we only have 10 degrees of freedom, and the bin contents are fractional. Therefore, these statistics are somewhat imprecise. However, the bootstrapping (leave-one-out) analysis of AH10, which we also performed on this data set, lends support to the $p < 0.05$ threshold used here independent of the Chi-square approximation. That is, this threshold used also leads to rejection of around 5% of the ensemble members themselves.

3 Results and discussions

3.1 Rank histogram

The rank histograms for all the ensembles are calculated (Fig. 1) and analysed using the goodness of fit tests described in Sect. 2.5. The number of ensemble members for which data are available varies between variables, particularly in the CMIP3 database, with the minimum ensemble size being 10. The shape of the rank histogram derived from a larger ensemble will be more clearly defined than that of a smaller one. Thus, here we treat all the ensembles as if they were the same size as the minimum CMIP3 ensemble, and re-bin the rank histograms to 11 bins.

Table 4 shows the minimum of the p-value among the total Chi-square statistics and its five components (bias, v-shape, ends, left-ends, and right-ends) according to JP08. Numbers with p-values less than 0.05 are shown in bold type, which means that this rank histogram is non-uniform at the 5% significance level from the null hypothesis. We also show the number of variables with p-value less than 0.05 for each ensemble in Table 4.

The rank histograms of the multi-model CMIP3 ensembles are shown in Fig. 1a (twentieth century experiments by CMIP3-AO) and Fig. 1b (control experiments by CMIP3-AS). The analysis of SAT, SLP, and PRCP in Fig. 1a is similar to that of AH10. The differences are that in the present study we include the uncertainty in the observations as described in Sect. 2.4 and there are also small differences due to the different number of model runs used. The results for these variables are consistent with those earlier results. Considering all the variables, in general, the rank histograms of the CMIP3 MMEs are not U-shaped or L-shaped, but close to flat or dome-shaped.

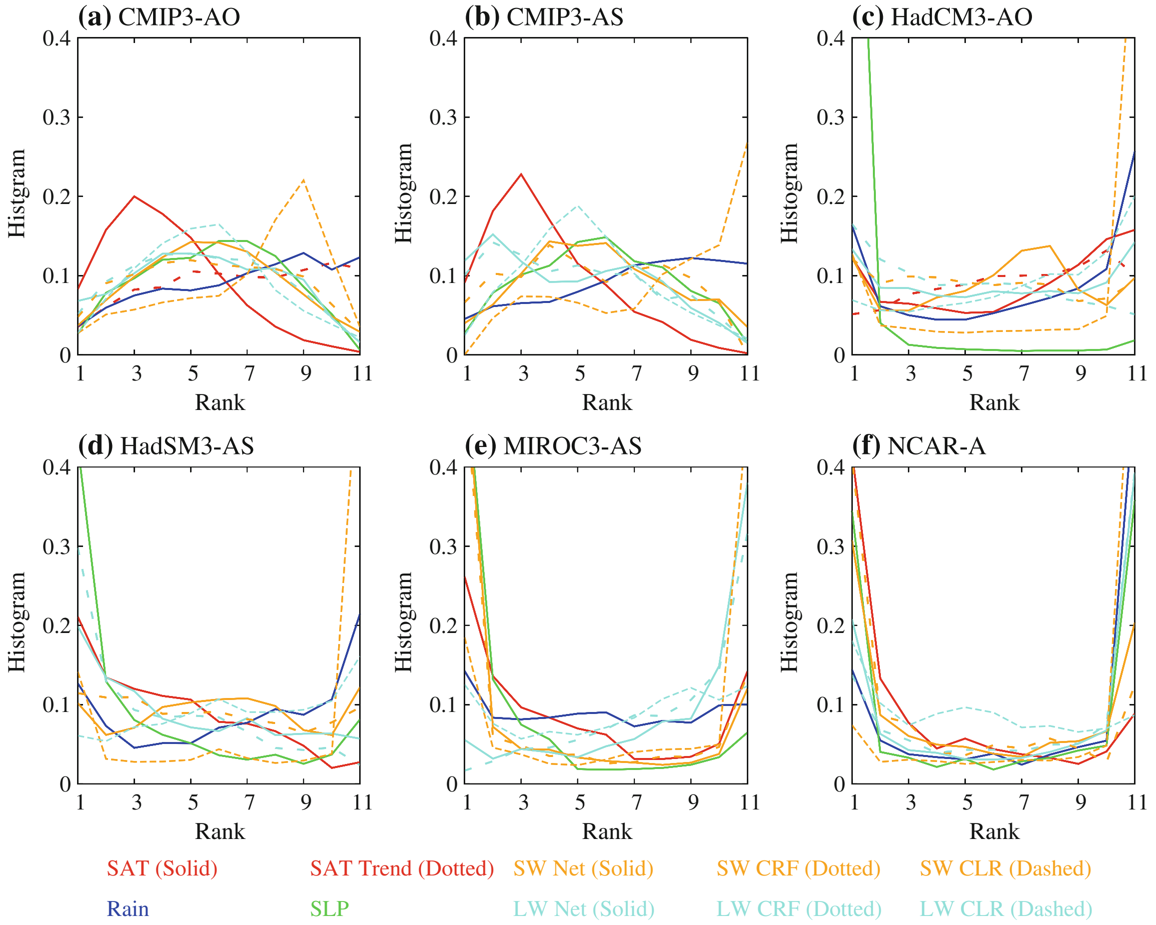


Fig. 1 Rank histogram of climate variables in the **a** twentieth century experiments by CMIP3-AO, **b** the control experiments by CMIP3-AS, **c** the twentieth century experiments by HadCM3-AO, **d** the control experiments by HadSM3-AS, **e** the control experiments by MIROC3-AS, **f** the control experiments by NCAR-AS. The horizontal axis indicates rank. Rank histogram of the climate variables of surface air temperature (*red solid*), surface temperature trend from 1960 to 1999 (*red dashed*), precipitation (*blue*), sea level pressure (*green*), the top of the atmosphere SW full-sky net downward radiation, cloud radiative forcing, and clear-sky radiation (*orange solid*, *dotted*, and *solid*), and their LW component (*aqua solid*, *dotted*, and *dashed*). In the figures of (**c**)–(**f**), the maximum values are more than 0.4 (0.9, 0.55, 0.6, 0.6, respectively), but the ranges are fixed to 0.4

Table 4 P values of Chi-square statistics calculated from rank histogram of the six climate model ensembles

Value	CMIP3-AO	CMIP3-AS	HadCM3-AO	HadSM3-AS	JUMP-AS	NCAR-A
T2	0.0576	**0.0375***	0.3367	0.0876	0.0609	**0.0003***
Rain	0.3830	0.3749	0.0506	0.1729	0.5696	**0.0000***
SLP	0.1440	0.1735	**0.0000***	**0.0002***	**0.0000***	**0.0000***
SW Net	0.1822	0.2424	0.7258	0.7281	**0.0000***	**0.0069***
LW Net	0.3110	0.2623	0.4401	0.2353	**0.0015***	**0.0006***
SW CRF	0.3796	0.3211	0.7349	0.7027	**0.0000***	**0.0004***
LW CRF	0.2823	0.3020	0.3455	**0.0227***	**0.0131***	**0.0008***
SW CLR	0.2890	0.0515	**0.0000***	**0.0000***	**0.0000***	**0.0000***
LW CLR	0.1083	0.1228	0.2276	0.4235	0.5806	0.3233
T2 trend	0.4961	NA	0.4875	NA	NA	NA
# of $p < 0.05$	0	1	2	3	6	8

First column indicate variables, and from second to seventh column indicates the minimum values of p values formulated by Jolliffe and Primo (2008). P values lower than 0.05 are shown in *bold* type with *. In the lowest row, numbers of variables with p values lower than 0.05 are shown

This suggests that the distributions of the climate variables in the CMIP3 MMEs are reliable or their model spread may be a little too broad. However, as shown in Table 4, for CMIP3-AO and CMIP3-AS, the rank histograms of all but one variable for one ensemble are not significantly different from the uniform distribution (Only in SAT of CMIP3-AS, the p value for bias is about 0.04, significant at the 5% level). This does not necessarily imply there is a problem with the ensemble since when analysing so many ensembles and variables it would be expected for some variables to fail the statistical test by chance.

In contrast, the shapes of the rank histograms of the SMEs differ depending on the ensemble and variable field considered. The results are show in Fig. 1c–f and Table 4. Apart from NCAR-A, all of the SAT, PRCP, and SAT trend results of the SMEs are reliable at the 5% level. Since predictions of these variables are very important for the future mitigation and adaptation to global warming, this result is encouraging. However, the climate variables SLP and SW-CLR are not reliable for any of the SMEs. This may be partly because the SMEs were constructed by perturbing uncertain physical parameters thought to affect climate sensitivity which are mainly related to clouds. As SLP and SW-CLR are determined by processes not related to cloud, such as the dynamical processes in the model, this may explain why the model ensembles do not cover a wide range of these variables. (SW-CLR is related to the distribution of atmospheric water vapour and aerosol which has a close link to the model dynamical processes). As shown in Fig. 1c–f, the rank histograms of SLP and SW-CLR in the SMEs are U-shaped (peaks in highest and lowest rank) or L-shape (peaks in highest or lowest rank) distribution, which means that, for much of the globe, the observations are outside the whole range of the ensembles.

Among the SMEs examined here, the SMEs of HadCM3-AO and HadSM-AS perform better. For those ensembles only two (SLP and SW-CLR in HadCM3-AO) or three (SLP, SW-CLR, and LW-CRF in HadSM3-AS) are unreliable as shown in Table 4. On the other hand, the SMEs by MIROC3-AS and NCAR-A fail the reliability test of a number of variables (six and eight, respectively): in addition to SLP and SW-CLR, SW, SW-CRF, LW, and LW-CRF are not reliable in both SMEs and PRCP is additionally not reliable in NCAR-A. It is of interest to note that the more reliable climate model ensembles (CMIP3-AO, CMIP3-AS, HadCM3-AO, and HadSM-AS) have a wide range of climate sensitivity centred on the canonical range (CMIP3: 2.5–4.5 K for 5–95% range, Randall et al. 2007, HadSM3: 1.3–5.2 K for 2-sigma, Y10). On the other hand, the SMEs with relatively narrow distribution, such as the MIROC3-AS and NCAR-A tend to have climate sensitivity that is either relatively high (MIROC3.2: 3.7–9.8 K for 2-sigma, Y10), or relatively low (NCAR CAM3: 2.2–3.2 K, Sanderson 2011).

In addition, both SW- and LW-CRF, and LW-CLR in CMIP3-AO, CMIP3-AS, HadSM3-AO, and HadCM3-AS are reliable in general (Table 4, LW-CRF in HadSM3 is not reliable), and their climate sensitivities with relatively wide range are determined by the feedback due to the changes in SW- and LW-CRF, and LW-CLR (i.e. SW and LW cloud feedback and water vapour feedback, Soden and Held 2006, Yokohata et al. 2008). On the contrary, SW-CRF in the MIROC3-AS and NCAR-A are not reliable (Table 4), and the changes in SW-CRF is responsible for relatively high climate sensitivity in the MIROC3.2 SME (Y10) and relatively low climate sensitivity in the standard NCAR CAM3 (e.g. Yokohata et al. 2008). Although it is reported that there are some relationships between the present states of cloud (e.g. Williams and Webb 2008, Yokohata et al. 2010) or water vapour (Sherwood et al. 2010) and climate feedback processes, it is not straightforward to relate the reliability of the present behavior and the climate sensitivity as discussed in Sect. 2.3. In the future, therefore, it is very important to investigate the reliability of present climate states and that of the climate sensitivity.

Another major difference between the SMEs considered here, however, is the number of parameters perturbed. While 31 parameters are varied in the HadCM3-AO and HadSM3-AS ensembles, 13 and 15 parameters are varied in the MIROC3-AS and NCAR-A ensembles respectively. The relatively reliable climate state, and wide range of CS in the HadSM3-AS may also, therefore, be as a result of a larger number of parameters having been varied in these ensembles, or other factors relating to the design of the ensembles.

Another issue to consider is that all of these ensembles have already been tuned to some extent to match observational data. In the case of the SMEs this is explicit in their construction. In each case, a "prior" ensemble (with parameters selected widely from prior distributions) has been narrowed down to "posterior" ensembles through comparison with observations, although the details of this process differ for each ensemble and at least in the case of the Hadley Centre ensembles, there was also an explicit goal of sampling widely in parameter space subject to observational constraints. In the case of the MME, this tuning process was probably more ad-hoc and subjective. Where ensembles have been tuned to data, it is reasonable to expect that these data will be closer to the 50th percentile in the resulting posterior ensemble than they were in the prior. We illustrate this principle for the idealised case of a single observation and Gaussian uncertainties. Given an ensemble of models from which an observable variable takes the mean value $m_1 = 0$ (without loss of generality) and standard deviation s_1, and an observation of this variable which takes the value m_2 with associated uncertainty s_2, the observation is initially at a normalised distance m_2/s_1 from the ensemble mean. When the ensemble is optimally updated in the light of this observation (i.e. tuned to the data), a direct application of Bayes Theorem gives the well-known result that the ensemble will have mean $m_2 \times s_1^2/(s_1^2 + s_2^2)$ and standard deviation $\sqrt{s_1^2 \times s_2^2/(s_1^2 + s_2^2)}$ for this observable. Thus the observation is now at a normalised distance of $m_2/s_1 \times \sqrt{s_2^2(s_1^2 + s_2^2)} < m_2/s_1$, and so has moved closer to the ensemble mean.

Thus if the ensembles were reliable prior to any tuning to observations, we may expect that the rank histograms of the ensembles to be somewhat domed if they have been carefully tuned, although it is unlikely that the ensembles have been optimally tuned (and certainly not to all the data considered here) given the impracticality of this operation.

3.2 Root mean square error and standard deviation of model ensembles

In order to validate our approach and also investigate the distance between the observation and model ensemble mean, we calculate the root mean square error (RMSE) between the model ensemble and the data, and the standard deviation (SD) of the ensemble, as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{m}\left[X_{\mathrm{em}}(i) - X_{\mathrm{obs}}(i)\right]^2} \tag{5}$$

where $i = 1, 2, \ldots, m$ is the index of grid point, $X_{\mathrm{em}}(i)$ is ensemble means of model value, $X_{\mathrm{obs}}(i)$ is observed values at the ith grid point;

$$\mathrm{SD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m-1}\sum_{j=1}^{m}\left[X_{\mathrm{model}}(i,j) - X_{\mathrm{em}}(i,j)\right]^2} \tag{6}$$

where $j = 1, 2, \ldots, n$ is the index of number of model ensembles, $X_{\mathrm{mdl}}(i, j)$ is model value of the jth members at the ith grid point.

As shown in Fig. 2, in the relatively reliable model ensembles such as CMIP3-AO, CMIP3-AS, HadCM3-AO, and HadSM3-AS, the value of RMSE is comparable to that of the SD. This means that the distance between model ensemble mean and observation (RMSE) and the spread of model ensembles (SD) is close, or the former is smaller than the latter in some cases. This is a necessary (although not sufficient) condition for an ensemble to be reliable, enabling the model ensemble to reasonably cover the observation (truth). On the other hand, for the model ensembles of MIROC-AS and NCAR-A, the value of the RMSE is larger than the SD in general. This means that model ensemble means are far away from the observation compared to the ensemble spread, and therefore the model ensembles cannot cover the observations. This may be either due to the relatively small spread of the ensemble, or a relatively large error in the mean. These results are consistent with the calculation of rank histogram shown in the previous sections.

We note in particular for SLP and SW-CLR, the SMEs of HadSM3-AS, HadCM3-AO, MIROC3-AS and NCAR-A, have narrower distributions than CMIP3. The narrowness supports the suggestion in Sect. 3.1 that the right parameters were not varied in the SMEs (or perhaps, that more substantial structural changes are required to generate a greater range of results). In addition, for the SW- and LW- Net and CRF, which are important variables determining climate sensitivity, MIROC3-AS and NCAR-A are

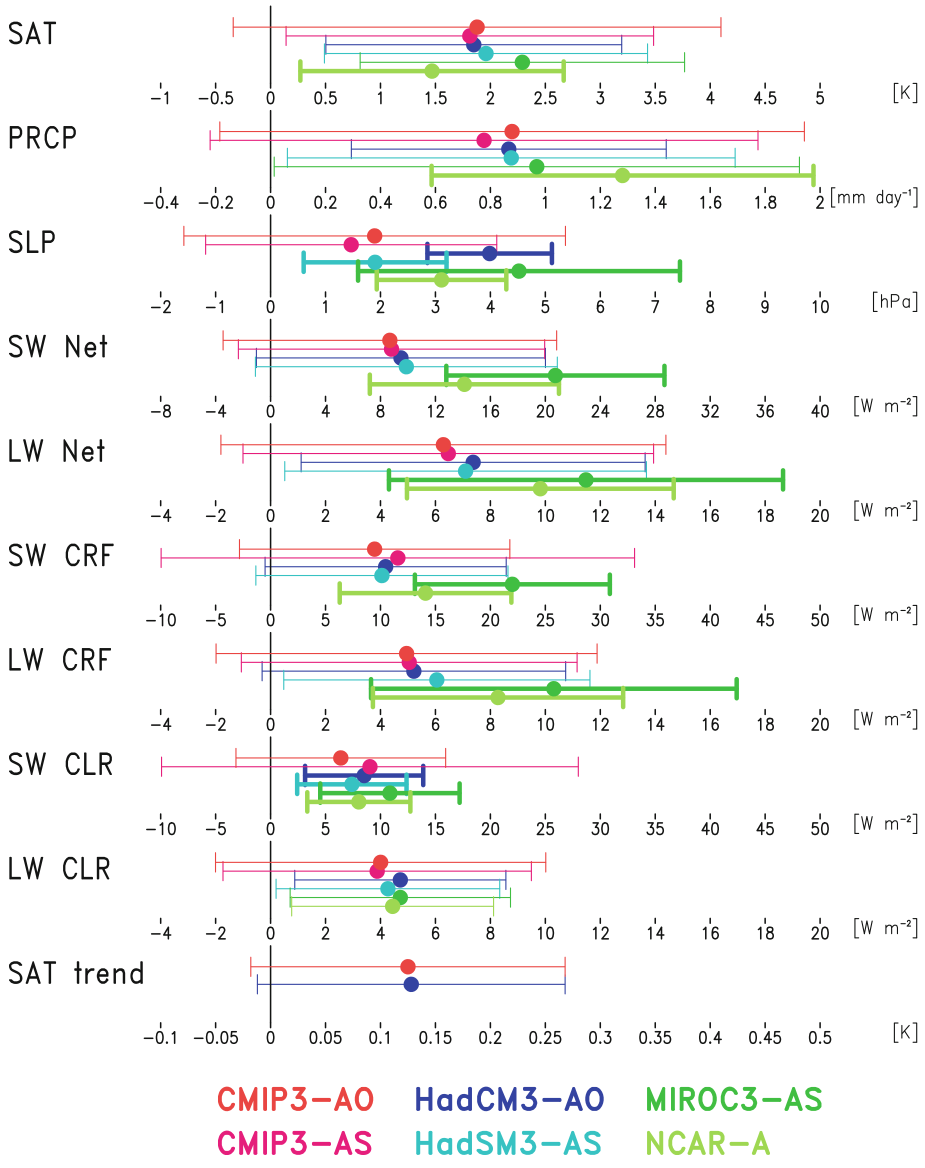


Fig. 2 Root mean square error (RMSE, *circle*) and standard deviation (SD, *half of error bar*) of climate variables of the six model ensembles, CMIP3-AO (*red*), CMIP3-AS (*magenta*), HadCM3-AO (*blue*), HadSM3-AS (*light blue*), MIROC3-AS (*green*), and NCAR-A (*light green*), respectively. Model ensembles with *p* value less than 0.005 are shown with *thick error bars*. As for SAT trend, only results of CMIP3-AO and HadCM3-AO are shown because the SAT trend can be calculated in the twentieth century experiment by atmosphere–ocean coupled general circulation models (AOGCMs)

heavily biased and have slightly narrow distributions compared to the results from CMIP3. These results support our suggestion that the climate sensitivities of these ensembles may also be biased.

3.3 Global map of observed rank among model ensembles

In order to illustrate the typical features and patterns of the biases in the different model ensembles and variable, the spatial maps of the rank of each ensemble and a selection of the variables are shown in Figs. 3, 4, 5, 6, and 7.

Figure 3 shows the rank of SAT observation in the present climate for each ensemble. The spatial patterns of rank highlight the areas where the ensembles are biased; the size of the patterns appears consistent with there being of order 10 degrees of freedom across the globe. Apart from HadCM3-AO, the model ensembles tend to underestimate the SAT over the ocean.

The rank of observed PRCP in the present climate among model ensembles is shown in Fig. 4. Some features are common across model ensembles. In all the ensembles, model ensembles overestimate the precipitation over the Central Pacific (north and south of the inter-tropical convergence zone, ITCZ), North America, North Asia, and some part in the Southern Ocean. These are the regions with less precipitation, and thus model ensembles may not have good performance in the dry regions. As discussed in Sect. 3.1, PRCP in the MMEs and SMEs are reliable apart from NCAR-A.

Figures 5 and 6 show the rank for SW and LW CRF at the TOA, respectively. As for the other variables, there are some similarities between the ensembles in the patterns of rank shown for each variable. There also appears to be a

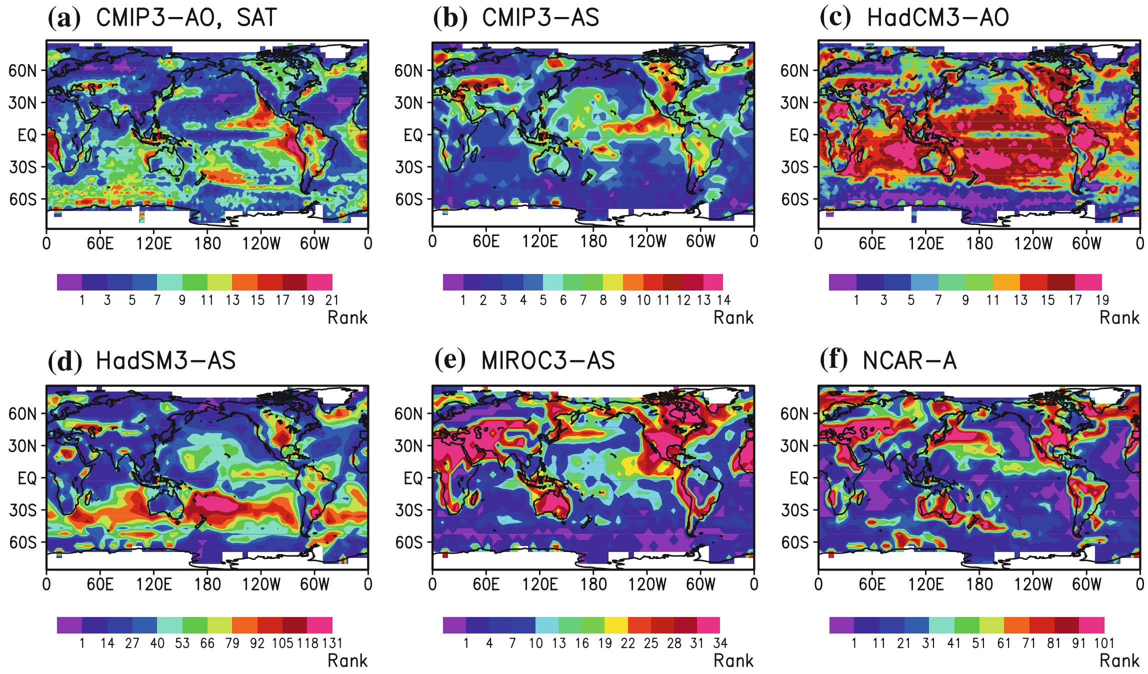


Fig. 3 Global map of rank of surface air temperature (SAT) observation among the MMEs and SMEs. *Blue color* indicates model underestimation and *red color* indicates model overestimation. Term of observation used for average is 1990–1999

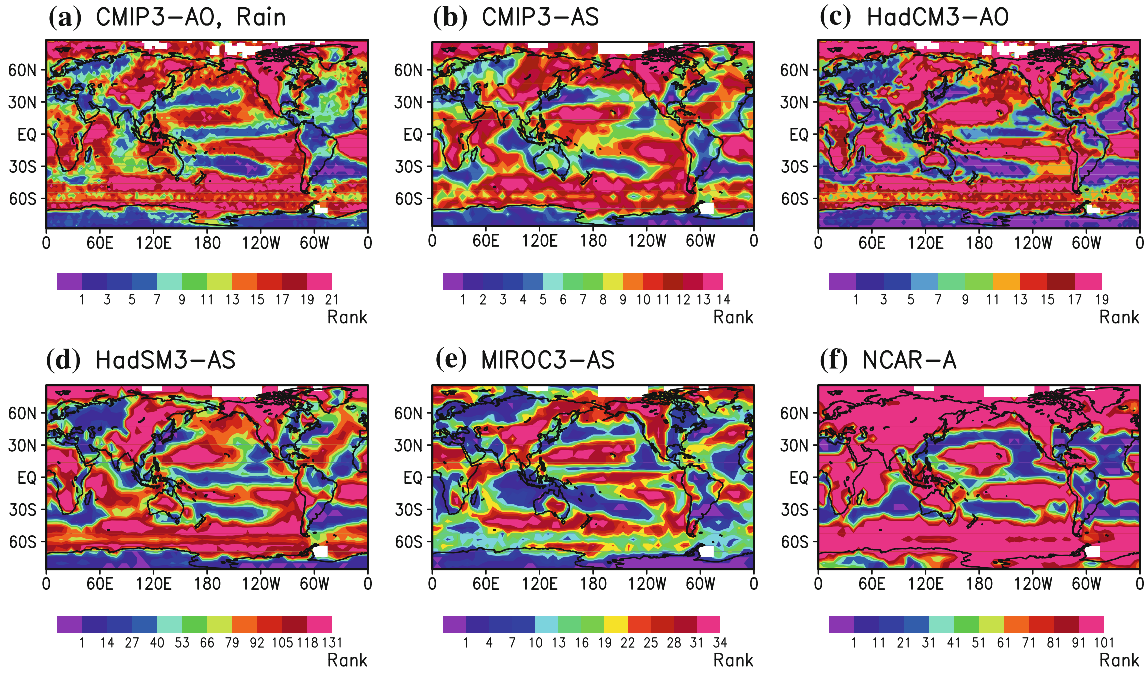


Fig. 4 Same as Fig. 3 but for rank of observation for precipitation (1990–1999) among the climate model ensembles

roughly inverse relationship between the ranks of two variables. These effects are particularly obvious for the first four ensembles, which are the ones found to be statistically reliable in general (the CMIP3-AO/AS and HadCM3-AO/HadSM3-AS). As shown in Fig. 3, the ensembles overestimate the magnitude of SW CRF (too much SW reflection by clouds) over the Central Pacific, north and south of the ITCZ. On the other hand, the model ensembles underestimate the magnitude of SW CRF, especially over the eastern coast of Pacific where thick stratocumulus is available (Williams and Webb 2008). The other two SMEs, MIROC3-AS and NCAR-A, have large areas where all the

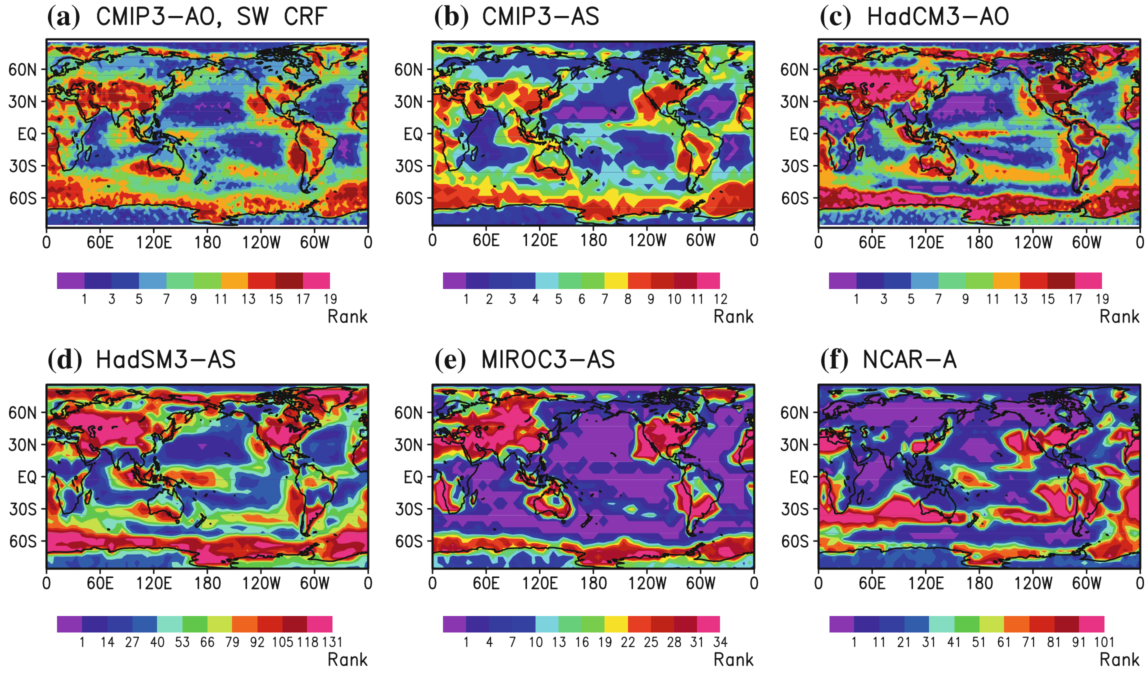


Fig. 5 Same as Fig. 3 but for rank of observation for SW cloud radiative forcing (CRF) at the top of the atmosphere (1986–1990) among the climate model ensembles

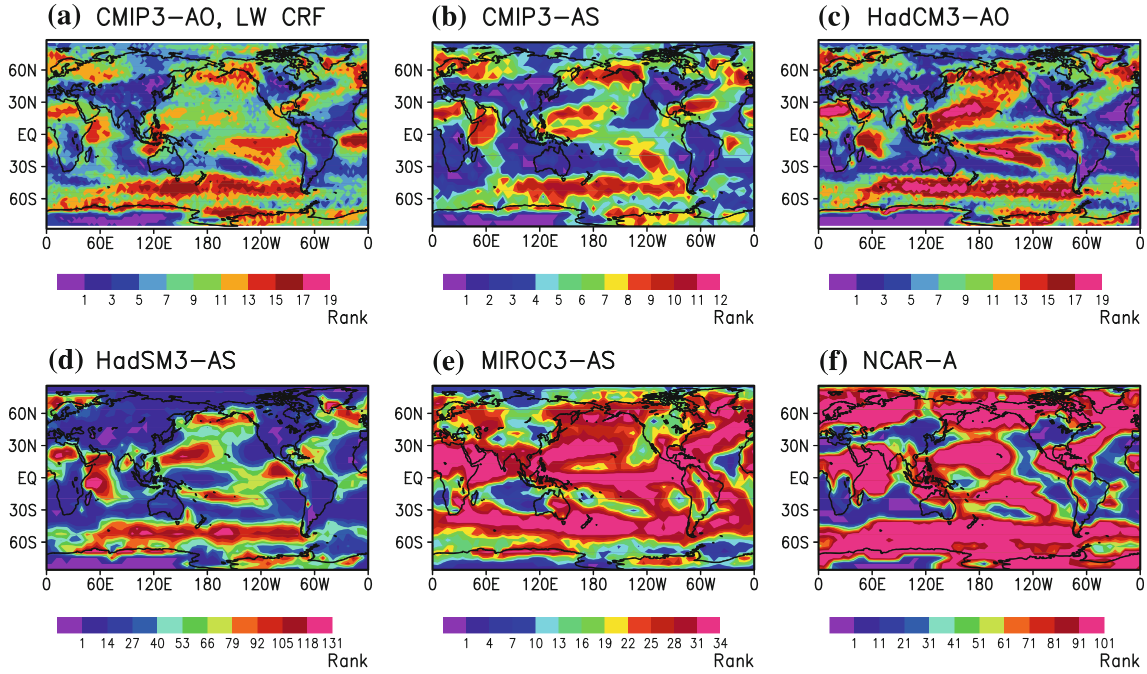


Fig. 6 Same as Fig. 3 but for rank of observation for LW CRF at the top of the atmosphere (1986–1990) among the climate model ensembles

model ensembles overestimate the magnitude of the SW CRF, and large areas where all the model ensembles overestimate the LW CRF.

Figure 7 shows the SW CLR at the TOA among the climate model ensembles. As discussed in the previous section, only the two CMIP3 MMEs are reliable for SW CLR (Table 4). This may be because it has been mostly parameters related to cloud processes that have been varied in the creation of SMEs, and SW CLR is not determined by these processes. The patterns of the rank do, however, look rather similar between all the ensembles, being low over land, and the Southern Ocean, and high over much of the rest of the ocean.

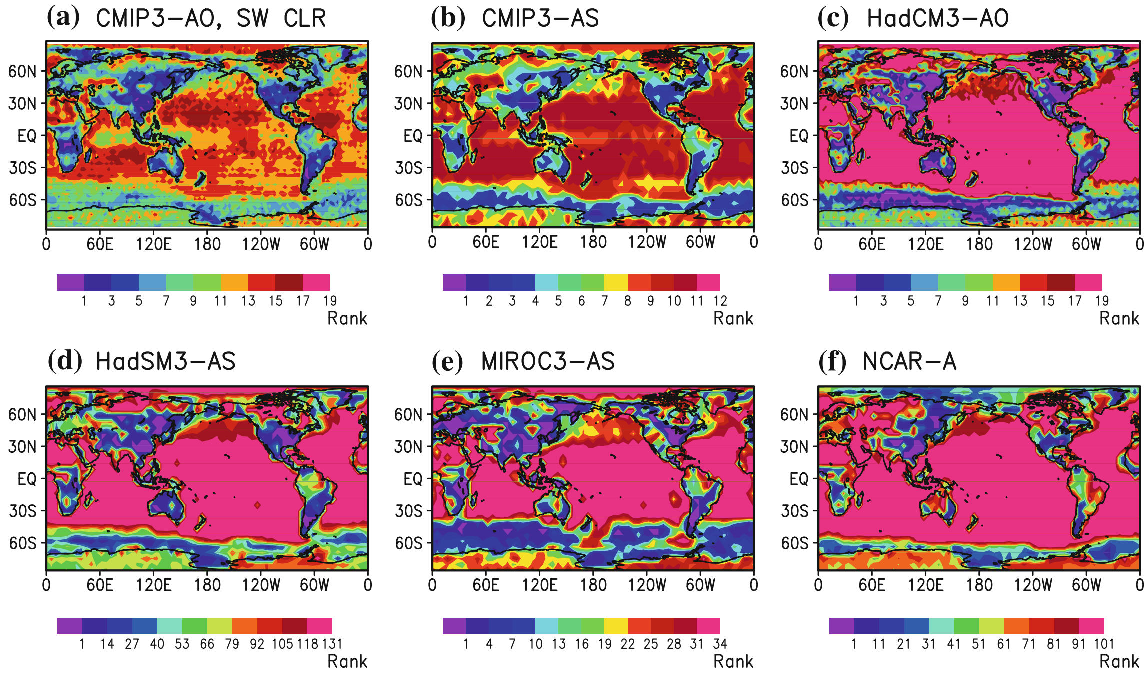


Fig. 7 Same as Fig. 3 but for rank of observation for clear-sky SW radiation at the top of the atmosphere (1986–1990) among the climate model ensembles

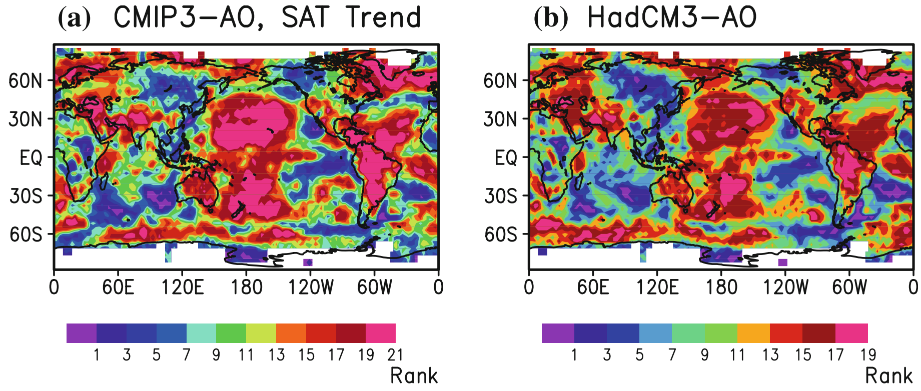
Figure 8 shows the rank of observed SAT trend over the last 40 years. Since SAT trend should be calculated from twentieth century experiment, only the results of CMIP3-AO and HadCM3-AO are shown. Interestingly, global map of rank of observation is quite similar between the two ensembles. Both the CMIP3-AO and HadCM3-AO overestimate the twentieth century SAT trend over the Pacific Ocean and South America. These features must be a common bias in the current climate models.

4 Conclusions

In the present study, simulations of the present-day climate by two kinds of climate model ensembles, multi-model ensembles (MMEs) of CMIP3 and single model ensembles (SMEs) of structurally different climate models, HadSM3/CM3, MIROC3.2, and NCAR CAM3.1, are investigated through the rank histogram approach. The reliability of various climate variables of these model ensembles are assessed by performing a goodness-of-fit test for the uniformity of the rank histogram.

Our analysis reveals that in the CMIP3 MMEs (both ensembles by AOGCM and ASGCM), all the climate variables we investigated (SAT, PRCP, SLP, TOA SW and LW radiation, cloud radiative forcing, clear-sky radiation) are reliable, with one marginally significant exception found out of the large number of statistical tests (SAT for the ASGCM). On the other hand, in the SMEs, the reliability varies between climate variables and model ensembles. For the mean state of SAT and PRCP, and the last 40-years trend of SAT, SMEs are mostly reliable.

Fig. 8 Same as Fig. 3 but for rank of observation for SAT trend (1986–1990) among the climate model ensembles. Only the results of CMIP3-AO and HadCM3-AO are shown because data are not available in other ensembles



Since these variables are very important for future climate prediction, these results are quite encouraging. Overall, however, the SMEs are less reliable than the MMEs. The two climate variables which are mainly determined by dynamical process, SLP and SW-CLR (determined by the distribution of water vapour), do not cover sufficiently wide ranges in any of the SMEs. We hypothesise that part of the reason for this may be because these SMEs were originally designed to investigate climate sensitivity, and so the focus was not on varying parameters which affect dynamical processes.

As well as the rank histogram, we also inspected the distribution of the rank of observation. Global map of rank of observation can reveal the typical features of biases in climate model ensembles. All the MMEs and SMEs tend to underestimate precipitation over the dry region, and to overestimate the cloud reflection over the Pacific Ocean. Analyses such as these should be useful for future climate model development as they indicate the robust biases found in the state-of-the art climate model ensembles.

We also find an interesting relationship between the reliability of present climate states and spread of climate sensitivity. In general, spread in climate sensitivity is mainly determined through the SW and LW cloud feedback, which is the change of SW and LW cloud radiative forcing under global warming. Our analysis reveals that in the CMIP3 MMEs and SMEs which have sufficient spread in climate sensitivity, or a spread which is consistent with studies published in the literature (about 2–5 K), both SW and LW cloud radiative forcing are reliable. On the other hand, in the SMEs with relatively high climate sensitivity (about 4–10 K), or the SMEs with relatively low climate sensitivity (about 2–3 K) compared to the studies in the literature, SW and LW radiation and cloud radiative forcing are not reliable.

The relationship between reliability of the present climate simulation and uncertainty in future climate prediction is very important because one of our goals of assessing the ability of climate model ensembles is to utilise its information for constraining uncertainties in climate prediction. The type of analysis presented here cannot show that the projections by reliable model ensembles will continue to form a reliable prediction into the future, but the results are at least encouraging in that they reveal no strong evidence of unreliability or other major biases or limitations in the CMIP3 ensemble, contrary to analyses based on the paradigm of a truth-centred ensemble. While it would appear to be a challenge to create an SME that is as reliable even for the present-day climate, the evidence here is that those ensembles (HadCM3-AO and HadSM3-AS) in which a large number of parameters are varied come closer to fulfilling the criteria. Thus careful experimental design and large computational resource may make this possible. In addition, it should be remembered that SMEs are of great value as a tool for understanding uncertainty in the model space. Thus, using various kinds of climate model ensembles including both MMEs and SMEs, we may expect to reduce uncertainties in climate prediction in the future.

Acknowledgments TY, JDA, and JCH were supported by Innovative Program of Climate Change Projection for twenty-first Century by MEXT, Japan. JDA were supported by the Global Environmental Research Fund, S-5 of the MoE, Japan. The simulations of the MIROC ensembles were performed on the Earth Simulator. We would like to thank David Sexton for helpful comments. We wish to thank JUMP members, Masakazu Yoshimori, Hideo Shiogama, Seita Emori, Manabu Abe, Kaoru Tachiiri, Michio Kawamiya for valuable discussions and comments, and Yoichi Kamae for technical support for drawing figures.

References

Adler RF et al (2003) The version 2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979–present). J Hydrometeorol 4:1147–1167

Allan RJ, Ansell TJ (2006) A new globally complete monthly historical mean sea level pressure data set (HadSLP2): 1850–2004. J Clim 19:5816–5842

Anderson JL (1996) A method for producing and evaluating probabilistic forecasts from ensemble model integrations. J Clim 9:1518–1530

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 multimodel ensemble. J Clim (in press)

Annan JD, Hargreaves JC, Ohgaito R, Abe-Ouchi A, Emori S (2005a) Efficiently constraining climate sensitivity with ensembles of Paleoclimate simulations. Sci Online 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

Brohan P, Kennedy JJ, Harris I, Tett SFB, Jones PD (2006) Uncertainty estimates in regional and global observed temperature changes: a new dataset from 1850. J Geophys Res 111:D12106. doi:10.1029/2005JD006548

CERSAT (1996) Altimeter and microwave radiometer ERS products user's manual, user's manual, v. 2.2, Oct. 21, C2-MUT-A-01-IF, CERSAT IFREMER BP 70, 29280 Plouzan'e, France

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, 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 Clim 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

Delworth TL et al (2006) GFDL's CM2 global coupled climate models—Part 1: formulation and simulation characteristics. J Clim 19:643–674

Diansky NA, Bagno AV, Zalensny VB (2002) Sigma model of global ocean circulation and its sensitivity to variations in wind stress. Izv Atmos Ocean Phys 38:477–494

Flato GM (2005) The third generation coupled global climate model (CGCM3). http://www.ec.gc.ca/ccmac-cccma/default.asp?lang=En&n=1299529F-1

Galin VY, Volodin EM, Smyshliaev SP (2003) Atmospheric general circulation model of INM RAS with ozone dynamics. Russ Meteorol Hydrol 5:13–22

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

Gordon HB, Rotstayn LD, McGregor JL, Dix MR, Kowalczyk EA, O'Farrell SP, Waterman LJ, Hirst AC, Wilson SG, Collier MA et al (2002) The CSIRO Mk3 climate system model (CSIRO Atmospheric Research, Aspendale, Australia). CSIRO Atmospheric Research Tech. Rep. No. 60

Haak H et al (2003) Formation and propagation of great salinity anomalies. Geophys Res Lett 30:1473. doi:10.1029/2003GL17065

Hall A, Qu X (2006) Using the current seasonal cycle to constrain snow albedo feedback in future climate change. Geophys Res Lett 33:L03502

Hamill TM (2001) Interpretation of rank histogram for verifying ensemble forecasts. Mon Weather Rev 129:550–560

Hansen J et al (2007) Climate simulations for 1880–2003 with GISS modelE. Clim Dyn 29:661–696. doi:10.1007/s00382-007-0255-8

Harrison EF, Minnis P, Barkstrom BR, Ramanathan V, Cess RD, Gibson GG (1990) Seasonal variation of cloud radiative forcing derived from the earth radiation budget experiment. J Geophys Res 95(D11):18687–18703

Jackson CS (2009) Use of Bayesian inference and data to improve simulations of multi-physics climate phenomena. J Phys Conf Ser 180. doi:10.1088/1742-6596/180/1/012029

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

Jolliffe I, Primo C (2008) Evaluating rank histograms using decompositions of the Chi-square test statistic. Mon Weather Rev 136:2133–2139. doi:10.1175/2007MWR2219.1

K-1 Model Developers (2004) K-1 coupled GCM (MIROC) description. K-1 Tech. Rep. 1. University of Tokyo, pp 1–34

Knutti R et al (2006) Constraining climate sensitivity from the seasonal cycle in surface temperature. J Clim 19:4224–4233

Knutti R, Furrer R, Tebaldi C, Cernak J, Meehl GA (2010a) Challenges in combining projections from multiple climate models. J Clim 23:2739–2758. doi:10.1175/2009JCLI3361.1

Knutti R et al (2010b) Good practice guidance paper on assessing and combining multi model climate predictions. In: Meeting report of the intergovernmental panel on climate change expert meeting on assessing and combining multi model climate predictions

Legutke S, Maier-Reimer E (1999) Climatology of the HOPE-G global ocean general circulation model. DKRZ Techn. Report 21

Loeb NG, Wielicki BA, Doelling DR, Smith GL, Keyes DF, Kato S, Manalo-Smith N, Wong T (2009) Toward optimal closure of the earth's top-of-atmosphere radiation budget. J Clim 22:748–766. doi:10.1175/2008JCLI2637.1

Marsland SJ et al (2003) The Max–Planck-Institute global ocean/sea ice model with orthogonal curvilinear coordinates. Ocean Modell 5:91–127

Marti O et al (2006) The new IPSL climate system model: IPSL-CM4. Scientific Note IPSL Pole Modeling, No. 26

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

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 et al (2007) The WCRP CMIP3 multimodel dataset: a new era in climate change research. Bull Am Meteorol Soc 88:1383–1394

Min S-K, Legutke S, Hense A, Kwon W-T (2005) Climatology and internal variability in a 1000-year control simulation with the coupled climate model ECHO-G—I. Near-surface temperature, precipitation and mean sea level pressure. Tellus 57A:605–621

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. Phil Trans R Soc Lond A 365:1993–2028

Pope VD, Gallani ML, Rowntree PR, Stratton RA (2000) The impact of new physical parameterizations in the Hadley Centre climate model-HadAM3. Clim Dyn 16:123–146

Randall DA et al (2007) Climate models and their evaluation. Climate change 2007: the physical science basis. In: Solomon S et al (eds) Cambridge University Press, pp 589–662

Roberts MJ (2004) The ocean component of HadGEM1. GMR Report Annex IV.D.3, Hadley Centre for Climate Prediction and Research/Met Office, Exeter

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

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

Schmidt GA et al (2006) Present day atmospheric simulations using GISS ModelE: Comparison to in situ, satellite and reanalysis data. J Clim 19:153–192

Sherwood SC, Ingrem W, Tsushima Y, Satoh M, Roberts M, Vidale PL, O'Goman PA (2010) Relative humidity changes in a warmer climate. J Geophys Res 115:D09104. doi:10.1029/2009JD012585

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

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

Soden BJ, Held IM (2006) An assessment of climate feedbacks in coupled ocean–atmosphere models. J Clim 19:3354–3360

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

Thompson DWJ, Kennedy JJ, Wallace JM, Jones PD (2008) A large discontinuity in the mid-twentieth century in observed global mean temperature. Nature 453:646–650

Toth ZO, Talagrand G, Candille G, Zhu Y (2003) Probability and ensemble forecasts. In: Jollife IT, Stephenson DB (eds) Forecast verification: a practitioner's guide in atmospheric science. Wiley, Chichester, pp 137–163

Trenberth KE et al (2007) Observations: surface and atmospheric climate change. In: Solomon S 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

Uppala SM et al (2005) The ERA-40 re-analysis. Q J R Meteorol Soc 131:2961–3012

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

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

Williams KD, Webb MJ (2008) A quantitative performance assessment of cloud regimes in climate models. Clim Dyn 33:141–157. doi:10.1007/s00382-008-0443-1

Willmott CJ, Matsuura K (2000) Terrestrial air temperature and precipitation: monthly and annual climatologies (version 3.01). http://www.climate.geog.udel.edu/~climate/html_pages/README.ghcn_clim.html

Wittenberg AT et al (2006) GFDL's CM2 global coupled climate models—Part 3: tropical Pacific climate and ENSO. J Clim 19:698–722

Xie P, Arkin PA (1997) Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. Bull Am Meteorol Soc 78:2539–2558

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

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

Yu L, Jin X, Weller RA (2008) Multidecade global flux datasets from the objectively analyzed air-sea fluxes (OAFlux) project: latent and sensible heat fluxes, ocean evaporation, and related surface meteorological variables. Woods Hole Oceanographic Institution, OAFlux Project Technical Report. OA-2008-01, 64 pp. Woods Hole. Massachusetts

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

Zhang Y, Rossow WB, Lacis AA, Oinas V, Mishchenko MI (2004) Calculation of radiative fluxes from the surface to top of atmosphere based on ISCCP and other global data sets: refinements of the radiative transfer model and the input data. J Geophys Res 109:D19105. doi:10.1029/2003JD004457