On the Effective Number of Climate Models

Christopher Pennell and Thomas Reichler

Department of Meteorology, University of Utah, Salt Lake City

Correspondence: *Thomas Reichler* (thomas.reichler@utah.edu)

Department of Meteorology, University of Utah
135 S 1460 E, Rm 819 (WBB)
Salt Lake City, UT 84112-0110
801-585- 0040 Fax: 801-581-4362

Abstract

Climate models are essential tools for assessing future climate change. In making predictions, it is beneficial to examine simulations from various models that are developed at centers around the world. The simple average over an ensemble of such models is often taken as the optimal prediction, which in studies of current climate is demonstrated as being more accurate than relying on any one individual model realization. However, this is only true to the extent that different models provide statistically independent information. Here, we examine the ability of current-generation models in simulating the observed present-day mean climate and show that similarities in model implementation play an important role in ensemble estimation. We demonstrate that the effective number of models is considerably smaller than the actual number comprising the ensemble. Our results suggest that the common practice of taking simple ensemble averages needs to be reconsidered.

1. Introduction

Over two dozen different climate models contribute to the ongoing mission of the Intergovernmental Panel on Climate Change (IPCC), whose aim is to provide reliable estimates of future climate change. Most findings of the IPCC's most recent 4th assessment report are based on simple averages over individual simulations produced by these models [*IPCC*, 2007]. This type of multi-model averaging improves a prediction only to the extent that different model outcomes are randomly distributed around the true future state, or in other words, independent from each other [*Abramowitz and Gupta*, 2008]. If this assumption is not met, resulting predictions are likely to be systematically biased and consequently, inaccurate.

There is reason to believe that the current generation of climate models considered by the IPCC violates the assumption of independence because these models have similar weaknesses [*Reichler and Kim*, 2007]. The probable reason is that models often share physical parameterization schemes and, at times, even large parts of the same code [*Pincus et al.*, 2008]. On the other hand, we may expect that effects from inter-model similarity could potentially be nullified over lengthy run-times. Even minor differences, for example in how small-scale processes are treated and sensible forcings are chosen [*Knutti*, 2008], could vastly amplify due to systemic non-linearities. Although this issue likely hampers the accuracy of individual simulations, its chaotic nature could lead to ideal simulation diversity within an ensemble.

With this in mind, the goal of this study is to explore the impact of model similarity in the context of a multi-model ensemble. We accomplish this by quantifying how well the current generation of models simulate present-day mean climate (section 2). Then, we examine the similarities in model deficiencies amongst the ensemble and clarify the source of these commonalities (section 3). Next, we discuss the potential impacts on current strategies for ensemble prediction (section 4). And finally, we discuss intricate details relating to our methodology (section 5).

2. The Effective Number of Models

We analyze deficiencies in 24 current-generation climate models from the 20th century experiment of the WCRP CMIP3 archive [*Meehl et al.*, 2007]. One model (BCC-CM1) is not included in our analysis since many of the atmospheric quantities used in this study are not provided for this particular model. We proceed by comparing mean climatologies for simulations as well as observations by calculating normalized RMS errors over the northern hemisphere for 37 physical and dynamical quantities (Table 1). These quantities are chosen based on the availability of suitable observations, as well as standard practices in climate model validation. Further data processing provides error distributions that are largely symmetric about their respective means, giving us confidence in the correct interpretation of our subsequent analysis of correlation coefficients. The results from these procedures provide for each model, quantity-specific scores relative to that model's mean performance. These scores collectively define a model's error structure (see Detailed Methodology).

<Table 1 about here>

These error structures together with our concept of an effective number of models are used to assess the amount of shared bias contained in the ensemble. The effective number of models (N_{eff}) is defined in the following way: N_{eff} equals to one if an ensemble consists of completely correlated error structures since the model members have identical deficiencies; alternatively, if all error structures are uncorrelated, then N_{eff} equals the actual number of models (N) in the ensemble. Although our concept may appear somewhat novel, it is similar to ideas such as the effective degrees of freedom or the effective sample size explored in the literature [*Wang and Shen*, 1999].

We estimate N_{eff} using two different methods: Method I incorporates an inverse technique based on the probability distribution of correlation coefficients [*van den Dool*, 2007] and Method II is based on an eigenanalysis of model correlations [*Bretherton*, 1999] (see Detailed Methodology). Testing our two methods on two artificial experiments produces reliable results.

We next apply these methods to the error structures determined from the 24 CMIP3 models. In order to quantify the similarity within the ensemble, we calculate N_{eff} for an increasing number of N models, ranging between 3 and 24. More specifically, we make robust estimates of N_{eff} by applying our two methods to 10,000 ensembles consisting of N randomly selected models (bootstrap without replacement).

Both methods indicate that as the number of models increases within an ensemble, the amount of shared bias also increases (Fig. 1). When all 24 models are eventually collected, the decrease in N_{eff} suggests that effectively only 12 to 16 models actually exist in the ensemble. This corresponds to a 33 to 50% reduction, which demonstrates that the error structures of the ensemble’s members are correlated at a level beyond what would be expected purely by chance.

<Figure 1 about here>

3. Examining Model Commonalities

In order to explore the basis behind inter-model similarities in our ensemble, we group models at different levels according to the strength of the relationships between their error structures. Specifically, we use an agglomerative clustering method which operates on the pair-wise distances calculated among the 24 models used in our analysis. Here, the distance between two models is simply defined as ($1–r$), where r represents the correlation coefficient between the models' respective error structures. The outcome of our cluster analysis is depicted by the dendrogram shown in Fig. 2. Other distance metrics and methodologies (not shown) produce similar inter-model relationships.

<Figure 2 about here>

Since models from the same center tend to differ little in terms of their implementation [*Delworth et al.*, 2006; *Schmidt et al.*, 2005; *Hasumi and Emori*, 2004], it is reasonable to assume that the large amount of bias seen in the ensemble is due to similarities between these same center models. Fig. 2 demonstrates that error structures calculated from models developed at the same center do indeed tend to be quite similar. For instance, the two CGCM3.1 models developed at the Canadian Centre for Climate Modeling and Analysis have error structures that are highly correlated (r = 86%). Similarly close relationships are seen in GISS-ER and GISS-EH, MIROC3.2(medres) and MIROC3.2(hires), as well as GFDL-CM2.0 and GFDL-CM2.1. Also, it is worth noting that the two GFDL models appear the most distinct from the other remaining models. This is shown in Fig. 2 by how these two models collectively merge at a rather long distance with other models.

Similarities between same center models alone, however, can only partially explain the reduction of N_{eff} seen in Fig. 1. For example, if we remove seven specific models, leaving each center represented only once in the ensemble (N = 17 in this case), we still find that N_{eff} is between 24 and 35% smaller than the full ensemble. The actual values of N_{eff}, in this instance, are indicated by the symbols corresponding to the two methods in Fig. 1. Given the remaining disparity between N and N_{eff}, we conclude that there must be similarities inherent in models across different centers as well.

As outlined before, we arrived at the above results by examining model error structures for the northern hemisphere. Examining error structures for the tropics (30°S – 30°N) and southern hemisphere (90°S – 30°S) leads to very similar conclusions. Over these two regions, the decrease in N_{eff} even exceeds that seen over the northern hemisphere by about one model (not shown). As before, same-center models tend to exhibit strong commonalities in the two regions, except for the two GFDL models over the southern hemisphere. This somewhat surprising outcome may be related to the large differences between the two models in simulating temperature and salinity of the southern ocean [*Gnanadesikan et al.*, 2006], which in turn may feedback into the atmospheric simulations over that region.

4. Conclusion

To summarize, for each of 24 CMIP3 models, we calculate errors in simulating present-day climatological mean-fields for 37 different atmospheric quantities. We use two methods that quantify the amount of inter-model similarities in these errors as it relates to the number of models in a current-generation climate ensemble. As the number of models in an ensemble increase, we see that the disparity between the number of models and effective number of models increases as well. In a full 24 member ensemble, we find that there only effectively exist about

12 to 16 models. To explore the reasoning behind this reduction, we use clustering analysis to group models based on similarities in their error characteristics. We see that a portion of inter-model similarity can readily be explained by models developed at the same center being included in the CMIP3 archive. This may not be too surprising since models from the same center often share a considerable amount of code. Commonalities in model implementation, however, are also seen to exist across all models. This can partially be explained by the CMIP3 archive being an 'ensemble of opportunity' [*Tebaldi and Knutti*, 2007]. As opposed to utilizing sound sampling design for model selection, results are essentially accepted from any center willing to participate in the archive. The distribution of model simulations belonging to such an ensemble is unpredictable and likely includes shared biases.

That the number of effective climate models is considerably less than the actual size of the CMIP3 ensemble suggests that simple arithmetic averages over different models simulations can give spurious confidence in a prediction. In order to produce more reliable estimates of future climate change it may be necessary to refine strategies for selecting and weighting the members of multi-model ensembles. Concerns about the effectiveness of simple multi-model averaging have led to some recent alternatives. Perturbed physics ensembles, for instance, sample a broad range of parametric uncertainty usually not explored by modeling centers and weight individual "model versions" based on their skill [*Murphy et al.*, 2004]. This approach has currently been attempted using only individual models, however, as incorporating multiple models is computationally prohibitive. Even modern probabilistic approaches, which consider different models simultaneously, typically require an assumption of model independence in order to produce tractable results [*Furrer et al.*, 2006; *Tebaldi et al.*, 2005]. Still, recent evidence suggests that weighted averages based on model skill show promise in improving ensemble prediction [*Min and Hense*, 2006; *Murphy et al*., 2004].

Simply constructing unbiased estimates does not, of course, guarantee predictive accuracy. And given the relatively modest number of models included in the CMIP3 ensemble, defining reasonable sampling strategies seems difficult at best. In light of these findings, it is apparent that the underlying processes which give rise to multi-model bias should be better understood. Quantifying the amount of inter-model similarities, in terms of an effective number of models, is a step toward intelligently weighting redundant ensemble members and may benefit future work in multi-model climate prediction.

5. Detailed Methodology

We evaluate a model's performance in simulating present-day climate by first calculating normalized RMS errors for each climate quantity as

$$E^2 = \frac{1}{K}\sum_{n=1}^{K} w_n (o_n - s_n)^2 / \sigma_n^2 \qquad (1).$$

Here, $(o_n - s_n)$ represents the difference between an observational and model simulated field at grid-point n, with K total grid-points pertaining to specific large regions. w_n provides proper spatial and vertical mass weighting, while σ_n^2 denotes the interannual variance taken from observations at n. Identical methodology has recently been applied in the literature [*Reichler and Kim*, 2007].

By conducting a logarithmic transformation of these errors, we ensure symmetric numerical distributions. For each model, we subtract its mean error so that errors are relative only to a model's overall performance. Examining statistical moments via testing of the null multivariate normal hypothesis [*Wilks*, 2006] provides acceptable evidence that errors are now essentially normally distributed (p-values of 0.995 and 0.527 for skew and kurtosis respectively).

Method I, for calculating the number of effective models, employs an inverse procedure based on analytical properties of the correlation coefficient distribution [*van den Dool*, 2007]. If two independent variables are Gaussian distributed, then their correlation coefficient r is Gaussian distributed with zero mean and variance $1/(N - 1)$ [*Bain and Engelhardt*, 1992]. N_{eff} is then estimated by equating the sample variance of quantity correlations S_r^2 with the expected population variance as

$$S_r^2 = \frac{1}{N_{eff} - 1} \qquad (2).$$

Given correlation coefficients are symmetric about zero, larger similarities amongst models will subsequently result in larger sample variability thereby reducing N_{eff}.

For Method II, we consider the eigenvalues that result from an eigenanalysis of the model error structure correlation matrix [*Bretherton et al.*, 1999]. N_{eff} can then be calculated as

$$N_{eff} = \frac{(\sum_{i=1}^{N} \lambda_i)^2}{\sum_{i=1}^{N} \lambda_i^2} \quad (3).$$

Here, λ_i is the i^{th} eigenvalue and N is the actual number of models. If error structures are independent, then all eigenvalues will have the same value and $N_{eff} = N$. However, if all error structures are identical, then there will exist only one non-zero eigenvalue and $N_{eff} = 1$. Here, N_{eff} is bounded inclusively between one and the number of models N.

Acknowledgments. We acknowledge Huug van den Dool for useful discussions, Junsu Kim for providing data and code, the modelling groups for providing the CMIP-3 data for analysis, the Program for Climate Model Diagnosis and Intercomparison for collecting and archiving the model output, and the JSC/CLIVAR Working Group on Coupled Modelling for organizing the model data analysis activity. The multi-model data archive is supported by the Office of Science, U.S. Department of Energy. This work was supported by NSF grant ATM0532280.

References

Abramowitz, G., Gupta, H. Towards a model space and independence metric. *Geophys. Res. Lett.* 35, L05705 (2008)

Bain, L. J., Engelhardt, M. *Introduction to Probability and Mathematical Statistics* 2nd edn (Brooks/Cole, 1992)

Bretherton, C. S. et al. The effective number of spatial degrees of freedom of a time-varying field. *J. Clim.* 12, 1990 (1999)

Delworth, T.L. et al. GFDL's CM2 global coupled climate models -- Part 1: Formulation and simulation characteristics. J. Climate. 19, 643 (2006)

Furrer, R. et al. Spatial patterns of probabilistic temperature change projections from a multivariate Bayesian analysis. *Geophys. Res. Lett.* 34, L06711 (2007)

Gnanadesikan, A. et al. GFDL's CM2 Global Coupled Climate Models. Part II: The Baseline Ocean Simulation. *J. Clim.* 19, 675 (2006)

Hasumi, H., Emori, S. *K-1 coupled GCM (MIROC) description, K-1 Technical Report No. 1* (Center for Climate System Research, University of Tokyo, 2004) [Available online at: http://www.ccsr.u-tokyo.ac.jp/kyosei/hasumi/MIROC/tech-repo.pdf as of October 17, 2008]

Knutti, R. Why are climate models reproducing the observed global surface warming so well? *Geophys. Res. Lett.* 35, L18704 (2008)

Meehl, G. A. et al. *Global Climate Projections. In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge Univ. Press, Cambridge, New York, 2007)

Meehl, G. A. et al. The WCRP CMIP3 multimodel dataset: A new era in climate change research. *Bull. Amer. Meteor. Soc.* 88, 1383 (2007)

Min, S.-K., Hense, A. A Bayesian approach to climate model evaluation and multi-model averaging with an application to global mean surface temperatures from IPCC AR4 coupled climate models. *Geophys. Res. Lett.* 33, L08708 (2006)

Murphy, J. et al. Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature* 429, 768 (2004)

Pincus, R. et al. Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models. *J. Geophys. Res.* 113, D14209 (2008)

Reichler, T., Kim, J. How well do coupled models simulate today's climate? *Bull. Am. Meteorol. Soc.* 89, 303 (2007)

Schmidt, G. A. et al. Present day atmospheric simulations using GISS ModelE: Comparison to in-situ, satellite and reanalysis data. *J. Climate* 19, 153 (2005)

Tebaldi, C. et al. Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the analysis of multimodel ensembles. *J. Climate* 18(10), 1524 (2005)

Tebaldi, C., Knutti, R. The Use of the Multi-Model Ensemble in Probabilistic Climate Projections. *Phil. Trans. R. Soc. A.* 2053, 365 (2007)

van den Dool, H. *Emperical Methods in Short-Term Climate Prediction* (Oxford Univ. Press, New York, 2007)

Wang, X., Shen, S. S. Estimation of spatial degrees of freedom of a climate field. *J. Clim.* 12, 1280 (1999)

Wilks, D. S. *Statistical Methods in the Atmospheric Sciences* 2nd edn (Elsevier, 2006)

Figure Captions

Figure 1. Effective number of models (N_{eff}) and model similarities for the northern hemisphere (30°-90°N). N_{eff} as a function of N for Method I (solid) and Method II (dashed) (see text). Grey shading indicates 95% confidence intervals. + and x symbol denote N_{eff} after excluding seven specific companion models [GISS-ER, GISS-EH, UKMO-HadCM3, MIROC3.2(medres), GFDL-CM2.0, CGCM3.1(T47), and CSIRO-Mk3.0].

Figure 2. Hierarchical clustering scheme based on model correlations. Similar (dissimilar) models merge closer to the right (left).

Table 1. Climate quantities used in this study. Acronyms listed under 'Validating observations' (5th column) are commonly used in the literature to denote specific observational data-sets. The average is taken of validating observation sets where more than one is given for a particular quantity.

	Quantity	Domain	Acronym	Units	Validating observations
physics	surface air temperature	global	TAS	K	CRU, ICOADS, NOAA
	surface skin temperature	land	TS	K	ISCCP
	zonal/meridional surface wind stress	ocean	TAUU, TAUV	10^{-2} Nm^{-2}	GSSTF2, ICOADS
	sea level pressure	ocean	PSL	hPa	ERSLP, HADSLP, ICOADS
	surface sensible/latent heat fluxes	ocean	HFSS, HFLS	Wm^{-2}	GSSTF2, HOAPS2, ICOADS, JOFURO, OAFLUX
	total cloudiness	global	CLT	%	CERES, ISCCP
	surface radiation (up/down, short-/longwave)	global	RSDS, RSUS, RLDS, RLUS	Wm^{-2}	BSRN, CERES, GEBA, ISCCP
	TOA outgoing shortwave radiation	global	RSUT	Wm^{-2}	CERES, ERBE, ISCCP
	TOA outgoing longwave radiation	global	RLUT	Wm^{-2}	CERES, ERBE, ISCCP, NOAA
	TOA cloud radiative forcing	global	CFLT, CFST	Wm^{-2}	CERES, ERBE, ISCCP
	precipitation	global	PR	mm/day	CMAP, GPCP
	precipitable water	global	PRW	mm	HOAPS2, NVAP
	snow coverage	global	SNW	%	NSIDC
	air temperature	zonal mean	TA	K	AIRS
dynamics	specific humidity	zonal mean	HUS	g/kg	ERA
	zonal/meridional wind 200 hPa	global	U200, V200	m/s	ERA
	stream function 200 hPa	global	ψ200	10^6 m^2s^{-1}	ERA
	velocity potential 200 hPa	global	χ200	10^6 m^2s^{-1}	ERA
	temperature 200 hPa	global	T200	K	ERA
	geopotential 500 hPa	global	Z500	gpm	ERA
	stationary waves 500 hPa	global	SW500	gpm	ERA
	zonal/meridional wind 850 hPa	global	U850, V850	m/s	ERA
	zonal mean zonal/meridional wind	zonal mean	UA, VA	m/s	ERA
	mean meridional mass streamfunction	zonal mean	MMC	10^9 kg/s	ERA
oceans	sea surface height	ocean	ZOS	m	GRACE-DOT
	sea ice content	ocean	SIC	%	GICE
	sea surface salinity	ocean	SO	‰	NODC
	sea surface temperature	ocean	TOS	K	GISST

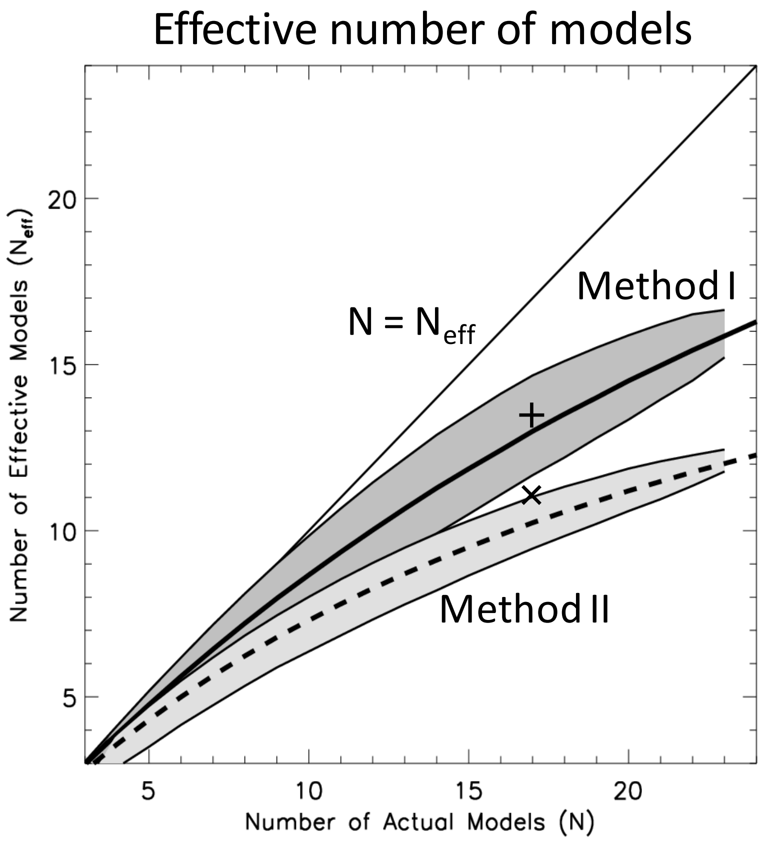

Figure 1. Effective number of models (N_{eff}) and model similarities for the northern hemisphere (30°-90°N). N_{eff} as a function of N for Method I (solid) and Method II (dashed) (see text). Grey shading indicates 95% confidence intervals. + and x symbol denote N_{eff} after excluding seven specific companion models [GISS-ER, GISS-EH, UKMO-HadCM3, MIROC3.2(medres), GFDL-CM2.0, CGCM3.1(T47), and CSIRO-Mk3.0].

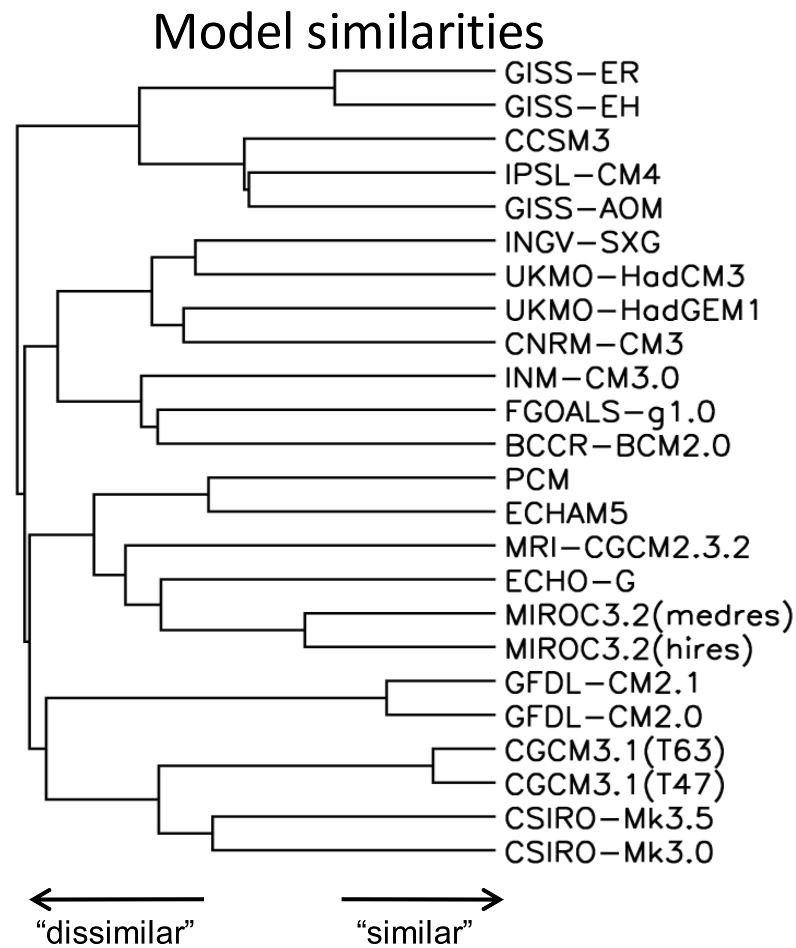

Figure 2. Hierarchical clustering scheme based on model correlations. Similar (dissimilar) models merge closer to the right (left).