On the Number of Effective Climate Models

Christopher Pennell and Thomas Reichler

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

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

Friday, April 23, 2010

Abstract

Projections of future climate change are increasingly based on the output of many different models. Typically, the mean over all model simulations is considered as the optimal prediction, with the underlying assumption that different models provide statistically independent information evenly distributed around the true state. However, there is reason to believe that this is not the best assumption. Coupled models are of comparable complexity and are constructed in similar ways. Some models share parts of the same code and some models are even developed at the same center. This contributes to the well-known problem of common model biases and possibly to an unrealistically small spread in the outcomes of model predictions.

This study attempts to quantify the extent of this problem by asking how many models there effectively are and how to best determine this number. Quantifying the number of effective models is achieved by evaluating 24 state-of-the-art models and their ability to simulate broad aspects of 20th century climate. Using two different approaches, we calculate the amount of unique information in the ensemble and find that the effective ensemble size is much smaller than the actual number of models. As more models are included in an ensemble the amount of new information diminishes in proportion. We further find that this reduction goes beyond the problem of "same-center" models and that systemic similarities exist across all models. Our results suggest that current methodologies for the interpretation of multi-model ensembles may lead to overly confident climate predictions.

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 to the public. The projections in the IPCC's most recent 4th assessment report (IPCC-AR4) were mostly based on simple multi-model averages over the different participating models (IPCC 2007). The underlying assumption here and in similar studies is that models are more or less statistically independent from each other (e.g., Abramowitz and Gupta 2008) and that averaging over the individual models will cancel out random errors. Probabilistic approaches of combining multi-model projections also require the assumption of model independence in order to produce tractable results (Furrer et al. 2007; Tebaldi et al. 2005). However, if the assumption of model independence in the above examples is not met, resulting predictions are likely to be biased towards some artificial consensus. In this case, the number of effective models, as defined by the amount of statistically independent information in the simulations, is less than what is suggested by the actual number of models.

Another undesirable result of having a relatively low number of effective models is that the uncertainty in multi-model derived climate projections could be significantly underestimated. Considering too many similar models in the calculation of the standard error of the ensemble mean leads to unrealistically narrow confidence-intervals. Underestimating uncertainty has important consequences, for example on climate impact studies, which rely on a realistic understanding of the range of potential climate outcomes (Knutti 2009).

Previous research has shown clear evidence that the current generation of models considered by the IPCC has common biases and thus violates the assumption of independence (Reichler and Kim 2008; Jun et al. 2008b). A plausible explanation for this lack of independence is the fact that models are designed in similar ways, for example by utilizing similar resolutions, numerical schemes and parameterizations. In actuality, models often even share large parts of the same code (Pincus et al. 2008), and some institutions contribute simulations from more than one version of the same model. This is certainly true of the third Coupled Model Intercomparison Project (CMIP3) (Meehl et al. 2007) data set, where model output is generally accepted from any center willing to participate. This data set has been described as an 'ensemble of opportunity', which is likely to include shared model biases (Tebaldi and Knutti 2007).

It could be argued that effects of model similarity on multi-model climate prediction are negligible because climate simulations are typically very long. In this case, the highly nonlinear nature of the climate system coupled with even relatively subtle variations in code structure could produce large differences in model behavior within an ensemble. However, it still remains to be determined to what extent this argument holds.

In the present study, we seek to determine how statistically independent the models in the CMIP3 ensemble actually are from one another. In order to proceed, we develop the concept of a number of effective models. Although this concept may appear intuitive, some complications arise. First, one must clearly state what is meant by "model independence". Clearly, it does not mean that models produce different solutions, but rather that models arrive at their respective solutions in unique ways. Next, one must develop a suitable metric. In other words, under what criteria is such a metric constructed? Possible measures could consider a model’s error, the magnitude of its simulated variability, but there are a countless number of defensible choices. Although they are all perhaps sensible, they may not necessarily lead to similar conclusions. A further difficulty lies in how robust a particular estimate of ensemble similarity is given the limited amount of available data.

In what follows, we explore similarity in the CMIP3 ensemble by establishing a measure of error that relates to how well these models simulate present day mean climate (sections 2 and 3). Then, we describe two distinct methods which aid in quantifying the degree of similarity within the ensemble (section 4). Next, we present our results based on these two methods and we explore their sensitivity with respect to particular models and quantities (section 5). And finally, we summarize our results and discuss the potential impacts on current strategies for ensemble prediction (section 6).

2. Data

In this study, we examine climatological mean (1979 – 1999) data based on the output of $M = 24$ climate models from the 20th century experiments (20C3M) of the CMIP3. For each different model, we consider the simulation of $Q = 35$ climate quantities during each of the four seasons (DJF, MAM, JJA and SON). The different quantities, which are shown in TABLE 1, are chosen based on the availability of suitable observations as well as standard practices in climate model validation.

<TABLE 1 about here.>

We begin by formulating a data set based on errors in simulating observed present-day mean climate. Specifically, we examine the differences between simulated (f)

and observed (o) fields on a uniform grid, which are expressed for model m, grid-point n, season s and quantity q as $(f_{n,m,q,s} – o_{n,q,s})$. We normalize these differences based on the observed standard deviation at grid-point n, $\sigma_{n,q,s}$, written as

$$e_{n,m} = (f_{n,m} - o_n)/\sigma_n \, , \tag{1}$$

dropping the s and q subscripts for clarity here and hereafter. These differences are then non-dimensional and comparable across quantity. We emphasize that, although all subsequent analysis is performed separately for the four seasons, we will primarily focus on annual means, given by the mean over all seasons. In addition, we examine results individually for three regions of interest: the northern hemisphere (30°N – 90°N), the tropics (30°S – 30°N) and the southern hemisphere (90°S – 30°S).

For model m, the errors in (1) form spatial patterns expressed as vector $\mathbf{e}_m = (e_{1,m}, e_{2,m}, \ldots, e_{N-1,m}, e_{N,m})$, where N is the number of grid-points in the regional domain. Here, bold notation represents spatial error vectors, or patterns. It is well established that climate models have similar biases and that these biases result in correlated error patterns (e.g., Reichler and Kim 2008; Jun et al. 2008b; Knutti et al. 2009). Similar biases are typically characterized by the multi-model error pattern (MME), which can be written as $\overline{\mathbf{e}} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{e}_m$. In FIG. 1, an example of the MME is shown in the top-right panel for precipitation. In order to demonstrate the effect of normalization in (1), we also present in the top-left panel averages of non-normalized (“raw”) precipitation errors. Since the amount and variability of precipitation is much larger in the tropics than the extratropics, normalization allows for more comparable errors in all regions. The middle row of FIG. 1 shows individual precipitation errors for the GFD21 and MRICM models. These two particular models were chosen because, amongst all models, they exhibit the highest (81%) and lowest correlation (52%) with the MME, respectively. Clearly, both models' error patterns bear a strong resemblance to the MME, demonstrating the tendency for models to share similar large-scale biases. In order to examine model similarity beyond these overarching commonalities, we next remove from all model errors the portion of the MME that is congruent with the errors corresponding to each individual model.

Removing the relevant portion of the MME entails that these fields are constructed by standardizing both, the model error fields and the MME, and then subtracting the scaled standardized MME pattern from the standardized model error field, i.e., $\mathrm{d}_m = \mathrm{e}_m^* - r\,\overline{\mathrm{e}}^*$, where $(\cdot)^*$ indicates statistical standardization, and *r* is the correlation between the m^{th} model's error field and the MME. For simplicity, we refer to the construction of fields as "removing the MME" or "controlling for the MME". The correlation between the MME and each individual error pattern is now zero by construction. The result of removing the MME on the GFD21 and MRICM model errors can be seen in the bottom row of FIG. 1. Removing the MME tends to make model errors more dissimilar from each other, as exemplified by the correlations between the two models before (36%) and after (-13%) this procedure. The resulting collection of such *M x Q* error patterns will hereafter be referred to as "SPATIAL-data".

<FIG. 1 about here>

3. Calculating Model Similarity

FIG. 2 demonstrates the effects of removing the MME in SPATIAL-data for all quantities and models. The top curve ("Model vs. MME") shows correlations between individual model errors and the MME (i.e., $corr(\mathbf{e}_i, \overline{\mathbf{e}})$, where $corr(\cdot,\cdot)$ is the correlation function between two vectors), averaged over all models and seasons, as a function of quantity. These correlations tend to be highly positive demonstrating the extent of common systematic biases.

The middle curve (“With MME”) in FIG. 2 shows the mean over all correlations between model pairs when the MME is retained (i.e., $corr(\mathbf{e}_i, \mathbf{e}_j)$). The correlations are now smaller (~20-60%) but still quite positive. The largest correlations are found in precipitation quantities (pr, prw), cloudiness (clt), and surface temperature (ts). One can show that these correlations are roughly the square of the ones seen in the top curve, suggesting that similarities between model and MME imply strong correlations between model pairs.

The bottom curve (“Without MME”) shows the mean over all correlations between model pairs in SPATIAL-data when the MME is removed (i.e., $corr(\mathbf{d}_i, \mathbf{d}_j)$). In this instance, controlling for the MME produces correlations that are near zero, indicating that the effect of the MME has largely been mitigated. A slight negative bias results because the MME is an average of model error and, therefore, bears some likeness to individual model error fields. As the number of models in the ensemble increase, this bias tends toward zero because the resemblance between the MME and individual model errors decreases. We note that removing the MME greatly reduces regional variation across quantity and that there is little seasonal variation amongst correlations (not shown).

<FIG. 2 about here>

For SPATIAL-data, model error patterns are used directly to formulate an estimate of model similarity. For example, if the strength of linear relationship between two models' error fields is large, then we will interpret the two models as being similar.

We note that whenever we average correlations in this study, we first take the mean over the Fisher's z transformed correlation coefficients (i.e., corresponding z-values). The Fisher's z transformation (Wilks 2006) as a function of correlation coefficient, *r*, is defined as

$$z(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right). \qquad (2)$$

We then use the functional inverse of (2) to derive an average correlation. This is standard practice for averaging correlation coefficients.

4. Defining the Number of Effective Models

The number of effective models (M_{eff}) is defined in the following way: M_{eff} equals one if an ensemble consists of completely correlated error structures since the model members' error fields have identical features; alternatively, if all error fields are uncorrelated, then M_{eff} equals the actual number of models (M) in the ensemble. To measure M_{eff}, we utilize two methods that are both based on determining the effective degrees of freedom or effective sample size from a given data set. We apply these methods to our SPATIAL-data providing us with two different estimates of M_{eff}. We tested these two methods, along with others, on artificial data sets and found both to be particularly reliable (not shown). The two methods are as follows:

a. Z-method

Our first method ("Z-method") has been proposed in prior literature (van den Dool and Chervin 1986; Wang and Shen 1999). This method employs an inverse procedure based on analytical properties of the Fisher's z transformed correlation coefficient distribution (van den Dool 2007). As explained in the literature, if two independent variables are Gaussian distributed, then their Fisher's z transformed correlation coefficient, or z-value, is approximately Gaussian distributed with variance 1/(M - 3).

For the Z-method, we first calculate the sample variance of z-values as $S_z^2 = \frac{1}{K-1}\sum_{i<j}(z_{i,j} - \bar{z})^2$, where K denotes the total number of unique (i, j) pairs. M_{eff} is then estimated by equating S_z^2 with the expected population variance as $S_z^2 = (M_{eff} - 3)^{-1}$. Given that z-values are symmetric about some mean value, larger similarities amongst model errors will subsequently result in larger sample variability thereby reducing M_{eff}.

Using SPATIAL-data, we correlate error vectors located at two different grid-points. Defining a vector at grid-point n, consisting of M different model error elements as $\mathbf{g}_n = (d_{n,1}, d_{n,2}, ..., d_{n,M-1}, d_{n,M})$, the correlation between two vectors located at unique grid-points u and v can be expressed as $r_{u,v} = corr(\mathbf{g}_u, \mathbf{g}_v)$. The corresponding z-value, $z_{u,v} = z(r_{u,v})$ is then one sample from the distribution considered by the Z-method. When calculating these correlation coefficients from the error patterns, we exercise caution. Two nearby grid-points will produce spuriously high correlation since there is significant spatial dependency in these data. Some quantities, such as precipitation (pr), exhibit smaller spatial dependency than others, such as upper air temperature (t200). In order to produce more accurate M_{eff} estimates, it is necessary to only consider grid-point pairs which are located a large enough distance apart as to minimize the effect of spatial autocorrelation. Examining spatial error patterns, we calculate de-correlation length-scales (DCLS) based on the distance for which autocorrelation extends below a threshold of 1/*e*. A representative DCLS is constructed by averaging the DCLS for all available models. We only correlate two grid-points when they are located at a distance greater than one DCLS unit apart from each other.

b. EIGEN-method

For our second method ("EIGEN-method"), we consider the eigenvalues that result from an eigenanalysis of correlation matrix [R] (Bretherton et al. 1999), where [R] is a symmetric *M x M* matrix composed of correlation coefficients. The element of [R] corresponding to model index pair (*i, j*) is defined as $r_{i,j} = corr(\mathbf{d}_i, \mathbf{d}_j)$.

Collecting the *M* eigenvalues that result from the eigenanalysis of [R], M_{eff} can then be calculated as $M_{eff} = \frac{(\sum_{i=1}^{M} \lambda_i)^2}{\sum_{i=1}^{M} \lambda_i^2}$, where λ_i represents the i^{th} eigenvalue. If model error structures are collectively independent, then all eigenvalues will have the same value and $M_{eff} = M$. However, if all error structures are identical, then there will exist only one non-zero eigenvalue and $M_{eff} = 1$. Here, M_{eff} is bounded inclusively between one and the number of models *M*.

5. Results

In the first part of this section we present the results for the number of effective models. In the second part, we provide a breakdown of model similarities as a function of model and quantity in order to shed additional light on our findings.

a. Number of Effective Models

We now derive two different estimates for the number of effective models, M_{eff}, in the CMIP3 ensemble by applying our two methods (section 4) to our SPATIAL-data (section 3). More precisely, we calculate M_{eff} for an increasing number of M models, ranging between 3 and 24, which allows us to quantify the amount of statistical independence for ensembles of various sizes.

FIG. 3 presents the outcomes for (a) the Z-method and (b) the EIGEN-method. We estimate M_{eff} by creating 100 ensembles of randomly selected models (i.e., bootstrap without replacement) for an increasing number of M. We repeat this procedure for each quantity and season. The grey lines in FIG. 3 indicate the outcomes for each of 35 individual quantities, averaged over all trials and seasons. Sampling only 100 ensembles is sufficiently precise because we fit the empirical results to semi-logarithmic functions. The M_{eff} estimates for individual quantities differ widely (ranging between ~3 and ~15 for $M = 24$), indicating that results derived from single quantities are not representative for overall model similarity. Averaging these fitted curves over all quantities produces the thick solid black curve shown in both panels. The 95% confidence bounds displayed as grey shadings around the solid black curve are generated by averaging over the 35x4 (quantities x seasons) individual confidence intervals.

The two methods reveal that the number of effective models is substantially lower than the number of actual models. Averaged over all quantities, M_{eff} is only about 9 (Z-Method) and 7.5 (EIGEN-method) for all 24 models. Though these results only apply to the northern hemisphere, we arrive at similar estimates for the other two regions (not shown). Both methods produce slightly higher estimates (by ~2 models for $M = 24$) in the tropics. In other words, the models are more dissimilar over the tropics than over the extratropics.

The curves in both panels depict a characteristic concavity over M. This concavity indicates that procedurally adding new models provides less and less new information as the ensemble grows. To some degree, diminished returns on adding new models is to be expected since the chance for an added model having commonalities with the preceding models increases with ensemble size. In other words, this behavior is a consequence of biases shared across models.

<FIG. 3 about here>

Extrapolating from the semi-log equation defining the average number of effective models, we can speculate as to how much unique information could be gained by adding another hypothetical CMIP3 model. For the Z and the EIGEN-method, adding one model yields an information increase of ~1% and ~1.5%, respectively. If the models simulations were independent, we would expect a value of 4.2% (1/24).

Contrasting the outcomes from the two methods (FIG. 3a. and b.), we find that they arrive at quite similar results despite the differences in their respective approaches. Only for small ensembles, the Z-method often produces unrealistically high estimates, which are likely due to the uncertainties inherent in utilizing small sample sizes.

We also applied our methods to the error fields that were not controlled for the MME, denoted by $\mathbf{e}_\mathrm{m}$ above. For the EIGEN-method, the dashed curve in FIG. 4b shows the results when the MME is retained. In this instance, the M_{eff} estimates are now far lower and range between three and four in all three regions. This result implies that model similarities are largely encapsulated by the MME. For the Z-method, retaining the MME only slightly impacts the results and is therefore not shown. The reason for the slight difference is that the MME is subtracted by construction via the correlation operation.

b. Sensitivity to specific models and quantities

The results from the previous section raise a number of important questions. For example: What causes the considerable reduction of the number of effective models? What is the contribution of individual models and quantities to this reduction? And, what role do models from the same institution play? In the following, we try to shed some light on these questions.

Going back to FIG. 3, the number of effective models determined from individual quantities differs widely. Upon closer inspection, we find that quantities associated with “smooth”, large-scale error fields (e.g., t200, chi, psi) tend to produce small estimates while the opposite holds for quantities associated with “noisy”, small-scale error fields (e.g., pr, va, v850). In other words, there exists an inverse relationship between the characteristic spatial scale for a quantity and its associated M_{eff} estimate. Perhaps, this is not too surprising since small-scale features imply a much larger range of possible simulation outcomes.

We now investigate in more detail how error structures from individual models are related to each other. To this end, FIG. 4 shows correlations between error fields for all possible combinations of model pairs. Each circle represents the average correlation over all quantities and seasons for a given model pair. Each correlation appears twice in FIG. 4; once in each column representing the two models.

FIG. 4 shows that most correlations are quite evenly distributed around zero, which is primarily a consequence of removing the MME (see FIG. 2). However, FIG. 4 also illustrates nine model pairs that are significantly correlated at the 95% level ($r > 0.28$), as indicated by the larger outlined circles. According to the CMIP3 model documentation (www-pcmdi.llnl.gov), pairings at this level indicate models that are developed at the same center or share parts of the same code ("same-center" models): CCSM/PCM11, ECHM5/INGV4, GISSA/GISSR, BCM20/CNRM3, C3T47/ C3T63, CSR30/CSR35, GFD20/GFD21, GISSH/GISSR, MIROM/MIROH. Some of the less obvious pairings are CCSM3 and PCM11 (developed at the National Center for Atmospheric Research), ECHM5 and INGV4 (based on the same atmospheric model), and BCM20 and CNRM3 (share the same atmospheric component). Most of these nine pairings are also found over the other two regions (not shown). One notable exception is the GFDL model pair, which, consistent with earlier findings (Gnanadesikan et al. 2006), is quite dissimilar over the southern hemisphere. Lastly, FIG. 4 contains a few large negative correlations (e.g., GFD20 and INGV4), but these do not appear across all regions and are therefore not robust.

We now explore the influence of the same-center models, as identified above, on M_{eff}. We accomplish this in two ways. First, we remove CNMR3, C3T47, CSR30, GFD20, GISSR, MIROM from our ensemble and repeat the M_{eff} -analysis explained in section 5a. These particular models are removed because they are associated with the largest same-center correlations seen in FIG. 4. The outcome is presented by the thick dash-dotted curves in FIG. 3. As shown, the remaining $M = 18$ models lead to an increase in M_{eff}, but this increase is quite small. Next, we repeat this analysis, retaining only the six models and their same-center counterparts (dotted curves in FIG. 3). As expected, there is now a decrease in M_{eff}. The relative size of this decrease is larger than the relative increase when removing same-center models. This makes sense given the uneven distribution of the two groups of models. We also examined if same-center relationships are connected to specific groups of quantities. We find (not shown) that it is generally impossible to identify similarities that belong to a specific quantity or groups of quantities. Instead, it appears that each given model pair is well correlated across most quantities.

We now provide a summary view of the similarities amongst different models. To this end, we convert the correlations (r) seen in FIG. 4 into a distance metric and enter them into a hierarchical clustering scheme. The clustering scheme groups models at different levels based on the distances between models. The outcome of this analysis is graphically depicted by the "dendrogram" shown in FIG. 5.

<FIG. 5 about here>

From FIG. 5, one can see that the same-center model pairs, identified earlier, merge at relatively short distances. Clusters containing more than two models tend to merge at insignificant correlations ($r < 0.28$), denoted by grey shading, and therefore appear to arise by construction. The only exception is GISSA, GISSR, and GISSH shown at the top of the dendrogram. These three models are all developed at the same center, suggesting that this merger is meaningful. Also, it is interesting that FGOAL has the largest merging distance with any other model, which is consistent with findings from Jun et al. (2008a) that this model is most independent from the CMIP3 ensemble. Additionally, although the two Hadley Centre models (HADGM and HADCM) merge at a rather large distance, it is clear that the two have more in common with each other than with any other model.

All of our results, thus far, are based on single member simulations from each model (usually “run1”). Looking at multiple members from the same model we find that the respective outcomes are very similar. This is exemplified in FIG. 5 for two members of GFD21 (GFD21-A and GFD21-B). They exhibit a correlation of ~93%, which is higher than between any two different models.

6. Conclusion

To our knowledge, this study represents the first attempt at explicitly determining the number of effective models from an ensemble. This is accomplished by calculating spatial errors in simulating present-day climatological mean fields for 35 different quantities. We then construct a data set based on these errors and utilize two distinct methods to quantify the amount of statistical independence in the ensemble. Using both methods, we find that the number of effective models (M_{eff}) is considerably smaller than the actual number (M), and as the number of models increases, the disparity between the two widens considerably. For the full 24 member ensemble, this leads to an M_{eff} that, depending on method, lies only between 7.5 and 9 when we control for the multi-model error (MME). These results are quantitatively consistent with that from Jun. et al. (2008a, 2008b), who also found that CMIP3 cannot be treated as a collection of independent models.

As explained before, we consider the number of effective models to be a useful measure of model independence. The demonstrably low number of effective models suggests that CMIP3 is not a very diverse ensemble. Due to this lack of diversity, we discover diminishing returns on adding models to a growing ensemble. Regarding the northern hemisphere, for example, twelve models on average account for about 75% of the total information (FIG. 3). In other words, the CMIP3 ensemble gives the false impression of having more models than there actually are. As discussed in the introduction, this may imply that the CMIP3 ensemble underestimates the real extent of climate prediction uncertainty.

Previous studies have provided convincing evidence that averaging over the outcomes from many models (multi-model mean) generally outperforms any individual model (e.g, Reichler and Kim 2008, Gleckler et al. 2008). It has further been argued that the superiority of the multi-model mean is due to the inclusion of a large number of diverse models, which tend to reduce the effects of natural climate variability and cancel offsetting errors (Pierce et al. 2009). However, we do not find CMIP3 to be as diverse as suggested by its ensemble size, further limiting its potential usefulness for multi-model projections.

Common model biases are an obvious explanation for this lack of diversity. However, it is important to emphasize that we removed the MME in our calculations and still, the number of effective models is surprisingly small. Another possible explanation for small M_{eff} may be that some CMIP3 models were developed at the same centers, and such models tend to differ little in implementation (e.g., Delworth et al. 2006; Schmidt et al. 2005; Hasumi and Emori 2004). However, we find that same-center models only have a modest impact; eliminating them from the ensemble increases M_{eff} by less than 10%. This suggests that despite removing the MME, considerable overarching commonalities remain amongst the models. Apparently, removing the MME does not entirely eliminate such commonalities.

A potential caveat to this study is that model similarity is determined from present-day mean climate. Some studies indicate that there may be little relationship between the ability of models to simulate mean climate and their simulation of trends (Jun et al. 2008b, Pierce et al. 2009, Knutti et al. 2009, Reifen and Toumi 2009). However, in the present study we are merely interested in similarities in error patterns and not in the magnitude of errors. The strong similarities in model error structures found in our study indicate a considerable lack of model diversity. It is reasonable to suspect that such model similarities translate into a limited range of climate change projections.

Acknowledgments

We acknowledge Huug van den Dool for useful discussions, Junsu Kim for providing data and code, the modelling groups for providing the CMIP3 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 and NOAA grant OAR-OGP-2006-2000116.

References

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

Bretherton, C. S., M. Widmann, V. P. Dymnikov, J. M. Wallace and I. Bladé, 1999: The effective number of spatial degrees of freedom of a time-varying field. *J. Clim.* **12,** 1990

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

Furrer, R., R. Knutti, S. R. Sain, D. W. Nychka and G. A. Meehl, 2007: Spatial patterns of probabilistic temperature change projections from a multivariate Bayesian analysis. *Geophys. Res. Lett.,* **34,** L06711

Gleckler, P. J., K. E. Taylor and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.*, **113**, D06104, doi:10.1029/2007JD008972.

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

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

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

Jun, M., R. Knutti and D. W. Nychka, 2008: Spatial Analysis to quantify numerical model bias and dependence: How many climate models are there? *J. Amer. Statistical Assoc.*, **103(483),** 934-947

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

Knutti, R., 2009: Should we believe model predictions of future climate change? *Phil. Trans. R. Soc.*, **366**, 4647-4664

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak and G. A. Meehl, 2009: Challenges in combining projections from multiple climate models. *J. Clim.*, (submitted)

Meehl, G. A. et al., 2007: *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)

Meehl, G. A., C. Covey, K. E. Taylor, T. Delworth, R. J. Stouffer, M. Latif, B. McAvaney and J. F. B. Mitchell, 2007: The WCRP CMIP3 multimodel dataset: A new era in climate change research. *Bull. Amer. Meteor. Soc.,* **88,** 1383

Min, S.-K. and A. Hense, 2006: 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

Murphy, J., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins and D. A. Stainforth, 2004: Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature,* **429,** 768

Pierce, D. W., T. P. Barnett, B. D. Santer and P. J. Gleckler, 2009: Selecting global climate models for regional climate change studies. *PNAS*, **106,** 8441-8446

Pincus, R., C. P. Batstone, R. J. P. Hofmann, K. E. Taylor and P. J. Gleckler, 2008: Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models. *J. Geophys. Res.,* **113,** D14209

Räisänen, J., L. Ruokolainen and J. Ylhäisi, 2009: Weighting of model results for improving best estimates. *Clim. Dyn.* DOI:10.1007/s00382-009-0659-8

Reichler, T. and J. Kim, 2008: How well do coupled models simulate today's climate? *Bull. Am. Meteorol. Soc.,* **89,** 303

Reifen, C. and R. Toumi, 2009: Climate projections: Past performance no guarantee of future skill? *Geophys. Res. Lett.*, **36,** L13704

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

Tebaldi, C., R. L. Smith, D. Nychka and L. O. Mearns, 2005: Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the analysis of multimodel ensembles. *J. Climate,* **18(10)**, 1524

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

van den Dool, H. and R. M. Chervin, 1986: A comparison of month-to-month persistence of anomalies in a general circulation model and in the earth's atmosphere. *J. Atmos. Sci.*, **43,** 1454-1466

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

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

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

Figure and Table Captions

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 of all available observations for one quantity is taken as validation.

FIG. 1. Errors in climatological mean precipitation for DJF. (top) Shown are multi-model errors, (middle) GFD21 and MRICM errors (corresponding to e_m^* in the text) and (bottom) GFD21 and MRICM errors with the multi-model error removed (corresponding to d_m in the text). For visualization purposes only, spatial/mass weighting is not applied in this figure. "Raw MME" errors are in units of mm/day, while all other errors are unitless.

FIG. 2. Error pattern correlations by quantity. Results shown are for the northern hemisphere; results for the other regions are similar (not shown). Top curve ("Model vs. MME") shows annually-averaged correlation between model error and multi-model error. Bottom two curves depict annually-averaged correlation in errors amongst different models when the multi-model error is retained ("With MME") and when the multi-model error has been removed ("Without MME"). Quantity labels are ordered by increasing correlations ("With MME").

FIG. 3. Number of effective models as a function of actual models (northern hemisphere). Thick solid lines are averages over all quantities and models, and grey shading indicates 95% confidence intervals. Thick dotted is for same-center models only, and dash-dotted is for excluding six same-center models. Thin grey lines are quantity-specific estimates.

Dashed curve in bottom panel indicates including the multi-model error. Straight diagonal line shows $M_{eff} = M$. All curves are fits of the actual data to semi-logarithmic functions ($M_{eff} = a + b\ln\left(M\right) + cM$), and extrapolating addresses the occasional problem of missing data.

FIG. 4. Correlation between model errors (northern hemisphere). Shown are averages over all 35 quantities and the four seasons. Larger filled circles indicate significantly positive values that are outside the one-tailed 95% confidence limit (r = 28%, assuming a Gaussian distribution with empirically estimated moments).

FIG. 5. Hierarchical clustering based on model error correlation (northern hemisphere). Similar models merge closer to the right. The clustering scheme is based on the weighted pair-wise average distance algorithm developed for the Interactive Data Language. The distance between two models is given by $z(0.95) - z(r)$, with z being defined in (2) and with a value of 0.95 being an upper bound on correlation. Other distance metrics and methodologies produce similar inter-model relationships (not shown). Scale bar units indicate equivalent correlation.

Tables and Figures

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 of all available observations for one quantity is taken as validation.

	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
	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^{2}s^{-1}	ERA
	velocity potential 200 hPa	global	χ200	10^{6} m^{2}s^{-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
ocean	sea surface height	ocean	zos	m	GRACE-DOT
	sea surface salinity	ocean	so	‰	NODC
	sea surface temperature	ocean	tos	K	GISST

Figures

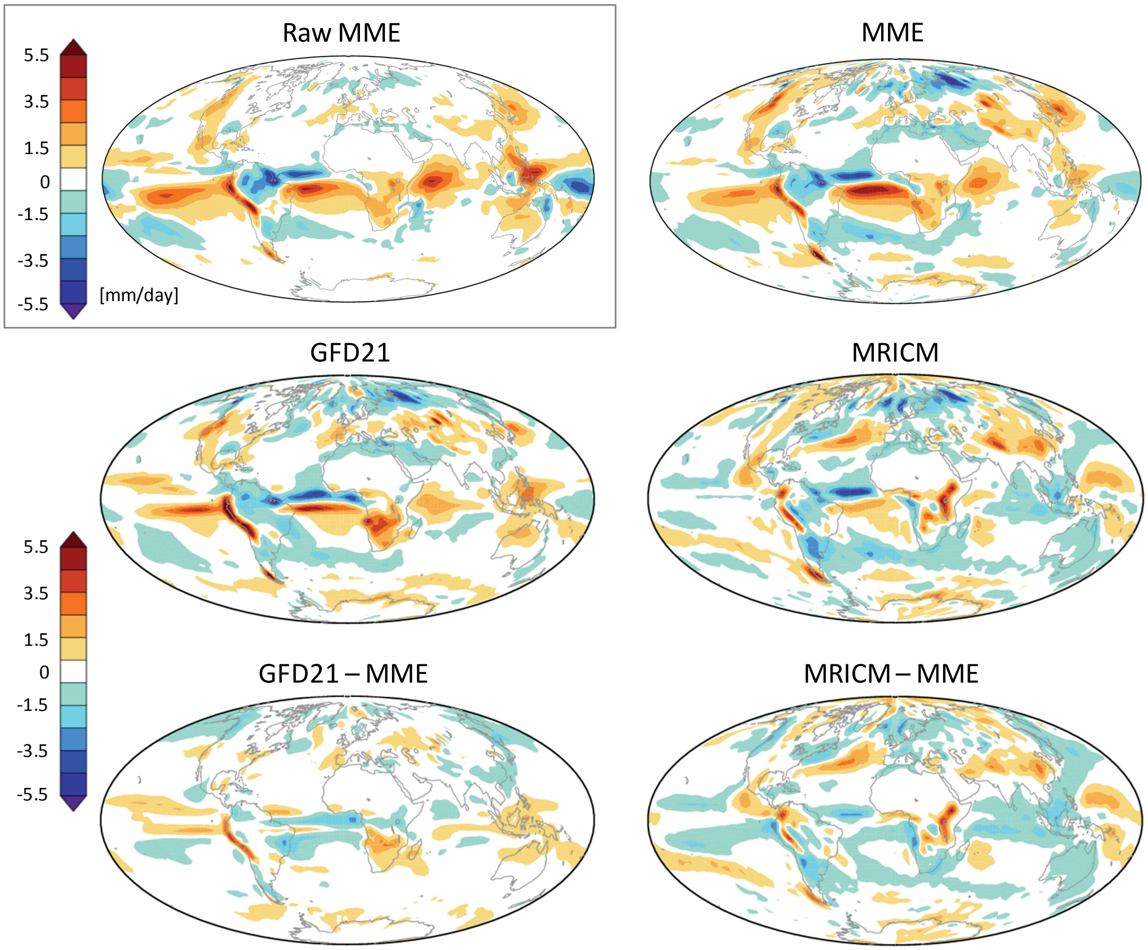

FIG. 1. Errors in climatological mean precipitation for DJF. Shown are (top) multi-model errors, (middle) GFD21 and MRICM errors (corresponding to e_m^* in the text) and (bottom) GFD21 and MRICM errors with the multi-model error removed (corresponding to d_m in the text). For visualization purposes only, spatial/mass weighting is not applied in this figure. "Raw MME" errors are in units of mm/day, while all other errors are unitless.

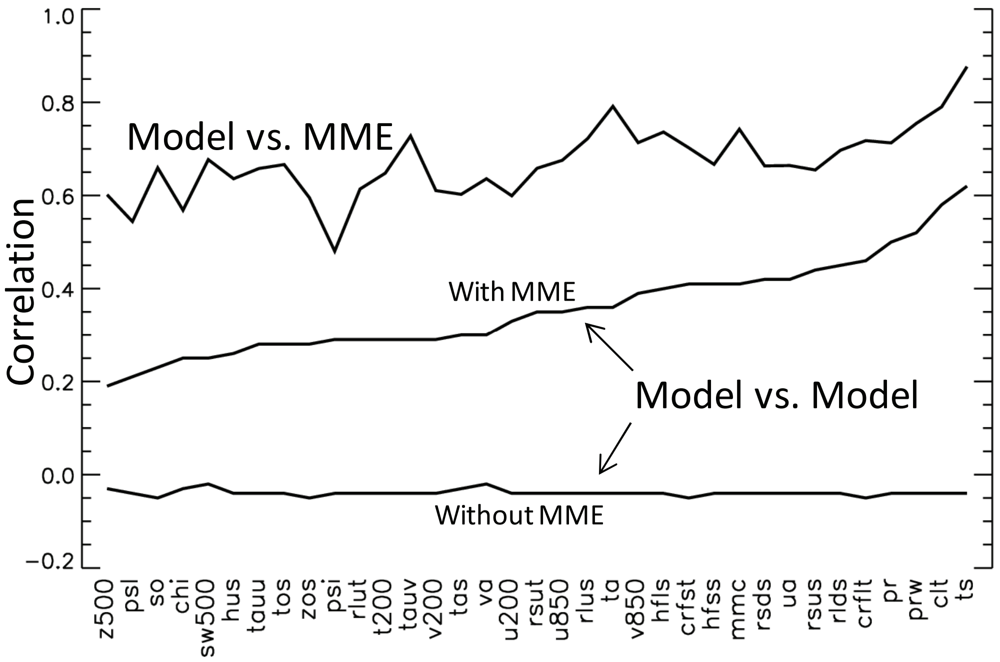

FIG. 2. Error pattern correlations by quantity. Results shown are for the northern hemisphere; results for the other regions are similar (not shown). Top curve ("Model vs. MME") shows annually-averaged correlation between model error and multi-model error. Bottom two curves depict annually-averaged correlation in errors amongst different models when the multi-model error is retained ("With MME") and when the multi-model error has been removed ("Without MME"). Quantity labels are ordered by increasing correlations ("With MME").

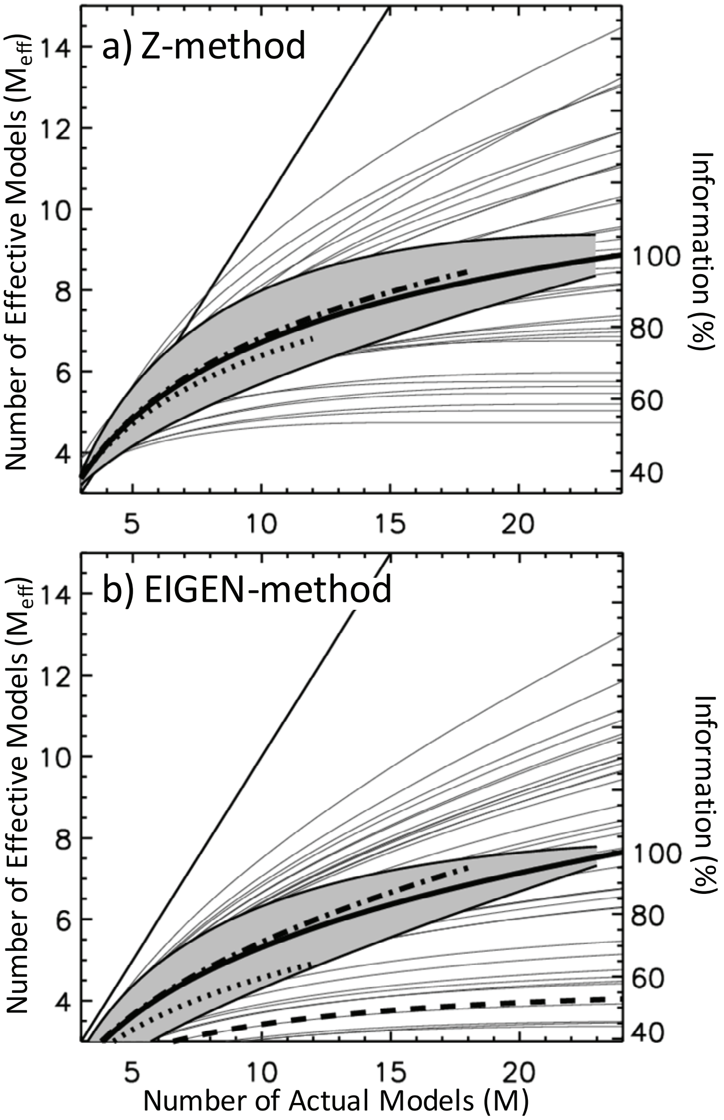

FIG. 3. Number of effective models as a function of actual models (northern hemisphere). Thick solid lines are averages over all quantities and models, and grey shading indicates 95% confidence intervals. Thick dotted is for same-center models only, and dash-dotted is for excluding six same-center models. Thin grey lines are quantity-specific estimates. Dashed curve in bottom panel indicates including the multi-model error. Straight diagonal line shows $M_{eff} = M$. All curves are fits of the actual data to semi-logarithmic

functions ($M_{eff} = a + b\ln\left(M\right) + cM$), and extrapolating addresses the occasional problem of missing data.

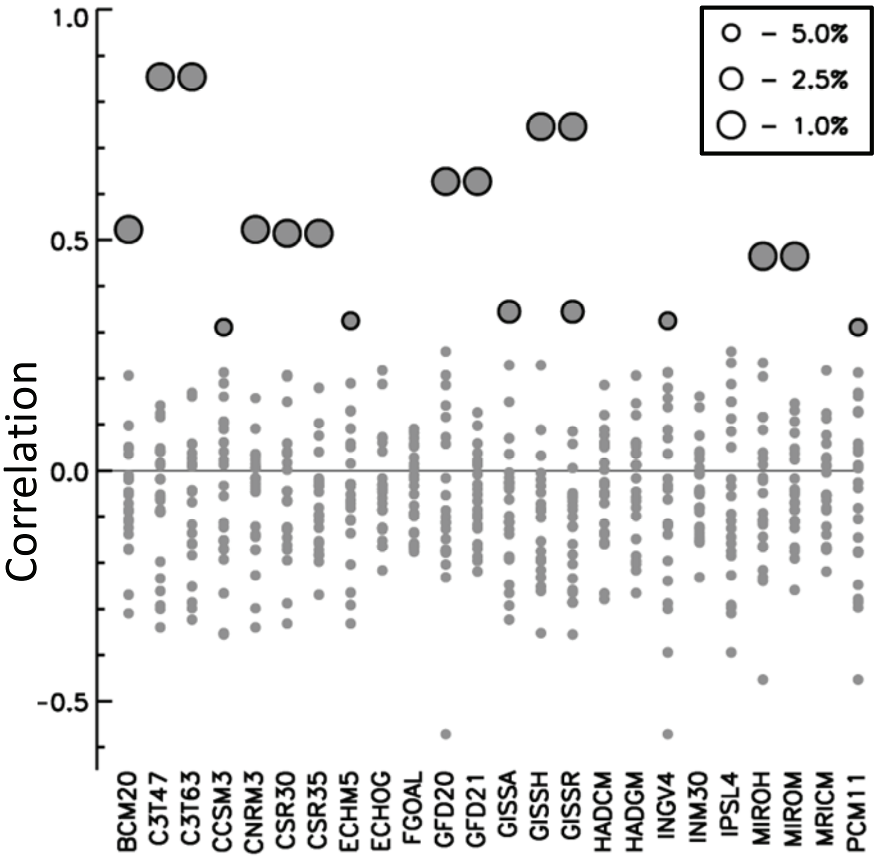

FIG. 4. Correlation between model errors (northern hemisphere). Shown are averages over all 35 quantities and the four seasons. Larger filled circles indicate significantly positive values that are outside the one-tailed 95% confidence limit (*r* = 28%, assuming a Gaussian distribution with empirically estimated moments).

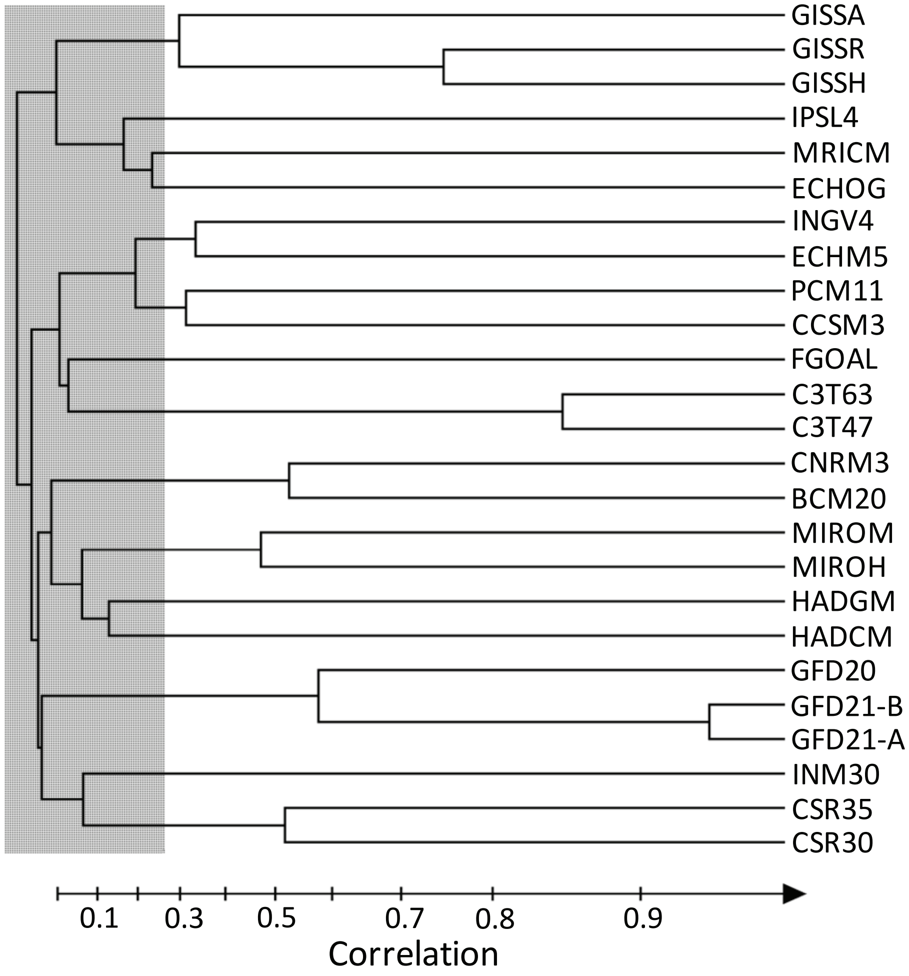

FIG. 5. Hierarchical clustering based on model error correlation (northern hemisphere). Similar models merge closer to the right. The clustering scheme is based on the weighted pair-wise average distance algorithm developed for the Interactive Data Language. The distance between two models is given by $z(0.95) - z(r)$, with z being defined in (2) and with a value of 0.95 being an upper bound on correlation. Other distance metrics and methodologies produce similar inter-model relationships (not shown). Scale bar units indicate equivalent correlation.