Climate model genealogy: Generation CMIP5 and how we got there

Reto Knutti,[1] David Masson,[2] and Andrew Gettelman[1,3]

Received 12 February 2013; accepted 12 February 2013; published 26 March 2013.

[1] A new ensemble of climate models is becoming available and provides the basis for climate change projections. Here, we show a first analysis indicating that the models in the new ensemble agree better with observations than those in older ones and that the poorest models have been eliminated. Most models are strongly tied to their predecessors, and some also exchange ideas and code with other models, thus supporting an earlier hypothesis that the models in the new ensemble are neither independent of each other nor independent of the earlier generation. On the basis of one atmosphere model, we show how statistical methods can identify similarities between model versions and complement process understanding in characterizing how and why a model has changed. We argue that the interdependence of models complicates the interpretation of multimodel ensembles but largely goes unnoticed.
Citation: Knutti, R., D. Masson, and A. Gettelman (2013), Climate model genealogy: Generation CMIP5 and how we got there, *Geophys. Res. Lett.*, *40*, 1194–1199, doi:10.1002/grl.50256.

1. Introduction

[2] Global climate models are ubiquitous and irreplaceable tools for projections of future climate change. They evolve and improve, but few people really understand exactly how and why. Model developers have scientific reasons for why they focus on improving on one process or component and not others, but the internal decision making processes for model development are rarely documented publicly. As a result, although new models are presented in detail in the literature and compared with observations, they remain massive and complex black boxes to many users, with many questions remaining unanswered. For example, why were certain parameterizations changed but not others? Which of those changes had the largest impact? Is the model "better" in terms of agreement with observations, or just "better" in terms of a more comprehensive description of the processes? Which variables and data sets were used to evaluate a given model?

[3] Because formal methods to quantify uncertainties in projections are complex and direct observational constraints often absent [*Knutti et al*., 2010; *Tebaldi and Knutti*, 2007; *Weigel et al*., 2010], the spread of an ensemble of models is often used as a first-order estimate of projection uncertainty [*Meehl et al*., 2007]. This assumes that the models are approximately a representative sample of our uncertainty in how to best describe the climate system given limited observations, imperfect understanding, and finite computational resources [*Knutti*, 2008; *Yokohata et al*., 2012]. It also assumes that there are not too many similarities that would bias the results. Of course, all models are similar because they describe the same system, but their biases, omissions of processes, simplifications, parameterizations of processes, and numerical approximations are also similar. In other words, they are often similarly biased with regard to reality, in some but not all cases for the same reasons (e.g., high mountains are not resolved in all models). This does not invalidate the use of the ensemble as a first-order estimate of uncertainty but complicates the interpretation.

[4] *Masson and Knutti* [2011, MK11 hereafter] produced a "family tree" of the Coupled Model Intercomparison Project Phase 2/3 climate models, which documents the similarities between models in an ensemble. For simplicity, we define model similarity as similarity in the model simulated fields because it is unclear how to define similarity of a model code or the underlying process assumptions. The term "model independence" is not used in a sense of statistical independence but loosely to express that the similarity between models sharing code is far greater than between those that do not. Models from the same centers were shown in MK11 to often be very similar in their present day climatology, and models in different centers sharing the same atmospheric model (even in different versions) were also closely related. MK11 argued that such similarities result from the fact that models evolve from their ancestors by modification and by exchange of ideas and code with other groups. Successful pieces are kept, improved, and shared, and less successful parts are replaced. Here, we present an analysis of the newest generation of models to supports this hypothesis.

2. Results

[5] We used data from the most recent World Climate Research Programme Coupled Model Intercomparison Project Phase 5 (CMIP5) [*Taylor et al*., 2012], along with data from the earlier CMIP3 and CMIP2 intercomparisons. Model similarity is defined as in MK11 (details in the Supporting Information of MK11) by a Kullback-Leibler divergence, a distance metric that considers the spatial field of monthly values in a control simulation without external forcing. It takes into account the seasonal cycle, the interannual variations, and the spatial correlation. The method and data from CMIP2/3 and observations are identical to those used by MK11. The only difference is that for Figures 1 and 3, the metric now also includes differences

All Supporting Information may be found in the online version of this article.

[1]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland.
[2]Federal Office of Meteorology and Climatology, MeteoSwiss, Zurich, Switzerland.
[3]National Center for Atmospheric Research, Boulder, Colorado, USA.

Corresponding author: R. Knutti, Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland. (reto.knutti@env.ethz.ch)

©2013. American Geophysical Union. All Rights Reserved.
0094-8276/13/10.1002/grl.50256

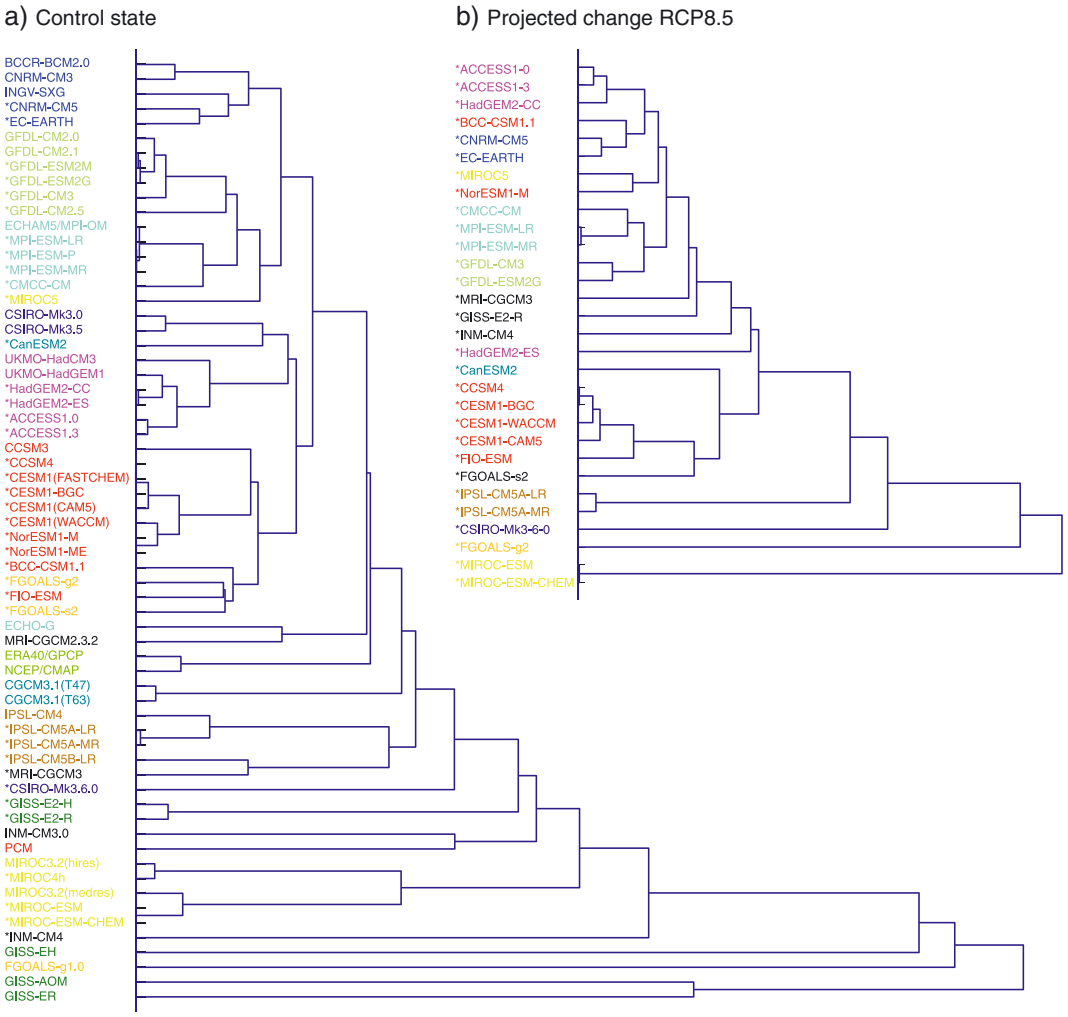

Figure 1. (a) The model "family tree" from CMIP3 and CMIP5 (marked with asterisks) control climate plus observations (ERA40/GPCP and NCEP/CMAP), shown as a dendrogram (a hierarchical clustering of the pairwise distance matrix for temperature and precipitation fields, see text). Some of the models with obvious similarities in code or produced by the same institution are marked with the same color. Models appearing in the same branch are close, and similarity is larger the more to the left the braches separate (for a detailed description of the method, see *Masson and Knutti* [2011]). (b) Same but based on the predicted change in temperature and precipitation fields for the end of the 21st century in the RCP8.5 scenario relative to the control.

in the annual mean climatology and that we use the mean square of the temperature and precipitation distance to define overall similarity rather than presenting the single variables as in MK11. None of those choices affect the main conclusions. The similarity metric is defined from the unperturbed preindustrial model control state and is strongly determined by biases in the present day climatology and seasonal cycle, that is, by model differences rather than initial conditions or forcing. The pairwise distance between models is used to construct "family trees" by a hierarchical clustering as in MK11. The interpretation of the trees is that models appearing in the same branch are close to each other in terms of the defined metric. Two branches or nodes are more similar the farther to the left the branching point is located. Note that all results presented here are only based on model output; they assume no knowledge about the structure, parameterizations, or code of any model. A similar tree is constructed for the projected change in the RCP8.5 scenario.

[6] The clustering of the CMIP3 and the new CMIP5 models for the control climate is shown in Figure 1a and confirms several connections. For example, the new MPI-ESM remains close to its predecessor because the atmosphere models ECHAM6 and ECHAM5 are quite similar. Two of the CSIRO models are close, and CCSM4 is close to CCSM3. Surprisingly, the CESM1 model versions are still close to CCSM4, although most of the major parameterizations were changed going from CCSM4 to CESM1 (see below). IPSL-CM5A is an only slightly modified IPSL-CM4 and appears close in the tree, IPSL-5A-LR/MR differ in resolution, whereas IPSL-CM5B involved substantial changes in the atmospheric model. The GISS-E2-H/R models differ in their ocean components and remain close but appear separated from the older GISS-E-H/R models in CMIP3 despite similar physics. The main reason is much higher resolution in both ocean and atmosphere. The GFDL-ESM2M/G models differ in

their ocean and have a new land surface and vegetation component, but their atmosphere is similar to that used in GFDL-CM2.0/2.1. GFDL-CM3 and CM2.5 are surprisingly close to the other GFDL models despite substantial changes to the atmosphere. All Hadley Centre models cluster as well, despite many years of development between HadCM3 and HadGEM2.

[7] Sharing of model code occurred in CMIP3 (e.g., the Italian INGV model using an ECHAM atmosphere, or the Norwegian BCCR using the French ARPEGE atmosphere also present in CNRM) but has become more widespread in CMIP5. The Australian ACCESS models are based on the HadGEM2 atmosphere, which is nicely picked up in the tree. NorESM is built with key elements of CESM1, as is FIO-ESM. BCC is based on a CCSM3 atmosphere, and the FGOALS atmosphere uses several parameterizations from CCSM. CNRM and EC-EARTH are both based on the ARPEGE/IFS/ECMWF atmosphere, and CMCC-CM uses an MPI ECHAM5 atmosphere. The use of similar ocean models appears to be less relevant to the surface climatology, but relationships exist as well. BCCR and NorESM use the Miami-based MYCOM ocean version, whereas its successor HYCOM is used in the GISS-EH models. Many European models (e.g., IPSL and CNRM) use an ocean based on the NEMO, ORCA, or OPA family. Some ocean codes in fact have left a remarkable legacy. The ocean codes by Bryan, Cox, and Semtner [*Bryan*, 1969], developed into the legendary MOM and POP models, improved versions of which are still used today in the Hadley and the NCAR models. Further details on model components and references are given in the Supporting Information.

[8] For the projected changes, some models remain close (e.g., those which share a similar atmosphere), whereas other similarities do not persist (e.g., if the new version of a model has a different climate sensitivity). This is consistent with the fact that the projected climate change is often not related to the climatological mean bias in an obvious way, a difficulty seen in many studies [*Knutti et al.*, 2010]. A key challenge and requirement is to find metrics where certain biases in observables can be attributed to specific parts of the model code or related to the projected changes, to constrain the projections, and to take into account the model dependency.

[9] The distance between models from the same institutions is often a factor of three to ten smaller than that to other models. This is not surprising but rarely taken into account. In most studies (including in the IPCC reports), all models are treated equally, thus giving more weight to those who have submitted multiple versions or share their code with

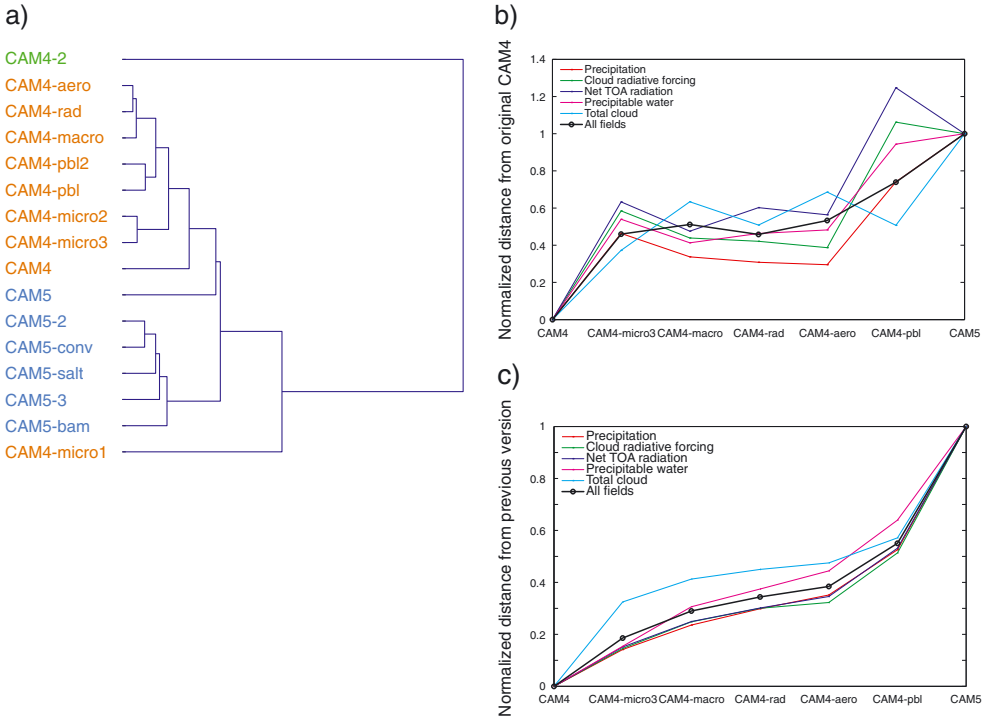

Figure 2. (a) Dendrogram of the ensemble tracking the different steps from CAM4 to CAM5 in the NCAR CESM1 based on aggregating the Kullback-Leibler divergence (see text) for precipitation, cloud radiative forcing, net top of atmosphere radiation, precipitable water, and total cloud amount. CAM4 perturbations are shown in orange, and CAM5 sensitivity tests are shown in blue. CAM4-2 is run from a different code base (different sea ice albedo specified at the surface). (b) Normalized absolute distance from CAM4 along the development path from CAM4 to CAM5 for different variables (colors) and all variables aggregated (black). (c) Same as Figure 2b, but for the accumulated distance from CAM4. The latter indicates how much the model has changed in one step compared with its predecessor when adding a new component and increases monotonously, whereas Figure 2b measures the distance to the base model CAM4, which can increase or decrease. The biggest changes to the model appear with the change in the shallow convection (CAM4-pbl to CAM5). The second biggest step was the change in the microphysics (CAM4 to CAM4-micro3) or the boundary layer (CAM4-aero to CAM4-pbl), depending on the variable (see main text).

others. Some models have evolved strongly from CMIP3 to CMIP5, whereas in other centers much of the effort has gone into additional components. Shared code or concepts may lead to similarity of the output, but the degree depends of course on what effect the shared code has on the simulated field and less on the amount of code. For example, a shared atmosphere produces more similarity than a shared ocean when looking at a precipitation field. Similarity may also arise from "fitting" to common data sets (see below). Shared code and data sets reduce the effective degrees of freedom in a multimodel ensemble.

[10] The detailed steps from one model version to the next are often not obvious. Exceptions are the MIROC model [*Watanabe et al*., 2012] and the evolution from the NCAR CCSM4 (CAM4) to CESM1 (CAM5), which is documented in detail by *Gettelman et al*. [2012] and illustrates steps between CESM with two different versions of the atmosphere model: CAM4 and CAM5. *Gettelman et al*. [2012] created an ensemble of different experiments to step from CAM4 to CAM5, by sequentially adding new microphysics (micro), macrophysics (macro), radiation (rad), aerosols (aero), planetary boundary layer (pbl), and finally the shallow convection scheme to reach CAM5 (all runs labeled CAM5). As discussed by *Gettelman et al*. [2012], the biggest change in climate sensitivity results from the change to the shallow convection scheme, which increases shortwave cloud feedbacks. As is clear from the tree shown in Figure 2a, the CAM5 experiments cluster together, with three single perturbation experiments similar to the base CAM5 experiment. The sequential changes between CAM4 and CAM5 also cluster together (with macro, rad, aero, and pbl added in that order). The micro1–3 series represent different tuning adjustments to get a better radiation balance (micro3 is in approximate balance). The same is true for the pbl1–2 experiments. Experiment CAM4_2 was run with different sea ice albedo specified at the surface, which may partially explain the separation. Thus, the CAM5 experiments cluster, and there is a break point in the differences. The perturbation experiments also cluster, with some of the single perturbation (tuning) experiments closest together. In general, the CAM models with the most similar physics packages cluster closest together.

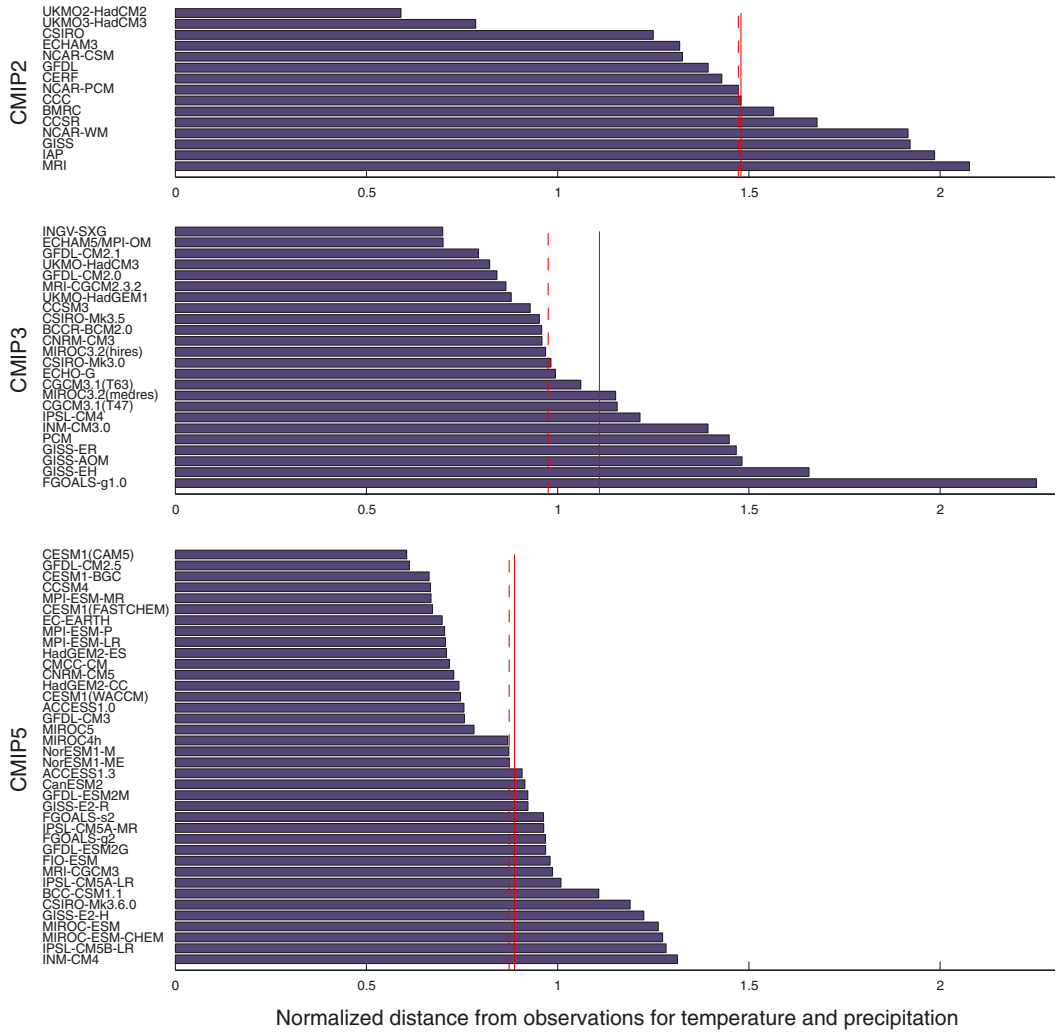

Figure 3. Normalized distance from observations in the CMIP2, CMIP3, and CMIP5 models. The distance metric is calculated as the root mean square of the surface temperature and precipitation distance as in Figure 1 but relative to observations (NCEP, ERA40, and MERRA for temperature; GPCP and CMAP for precipitation, see MK11). Mean and medians for the different ensembles are indicated by red solid and dashed lines, respectively. Note that most models in CMIP2 (including HadCM2, but not HadCM3) used flux corrections.

[11] Another way to depict the development path is to show the distance of each version from the starting CAM4 version (Figure 2b) and the cumulative distance along the path (Figure 2c). The final change in the shallow convection introduces the largest change (last step in panel c), although it brings CAM5 closer to the original CAM4 in some variables (panel b). The second largest changes are caused by the microphysics and boundary layer schemes, depending on the variable.

[12] We can also use the distance metric of Figure 1 as one example to quantify distance from observations for different CMIP generations (shown in Figure 3), but with the strong caveat that linking model performance metrics to model quality or skill is difficult, subjective, and strongly metric dependent [*Gleckler et al*., 2008]. The ranking within the ensemble should therefore not be over interpreted and it differs depending on the metric. CMIP5 continues the trend of better agreement with observations [*Reichler and Kim*, 2008], with the mean distance from observations reduced by approximately 20% from CMIP3 to CMIP5. Not unexpectedly, progress becomes harder at higher performance levels, that is, the "worst" models have improved most or are no longer used. The contribution of internal unforced variability (quantified using multiple segments of a control simulation) is small for this metric, but observational uncertainties (estimated from differences across multiple data sets) may explain some of the remaining model biases. The typical distance between two reanalysis temperature data sets (e.g., MERRA and ERA40) is about half the distance as the typical distance between a model and the "observations." Nevertheless, as shown already by MK11, the different observation and reanalysis data set cluster together (see Figure 1a), and the rankings in Figure 3 are nearly identical irrespective of the chosen data set.

3. Conclusion and Discussion

[13] We propose that one reason some models are so similar is because they share common code. Another explanation for the similarity of successive models in one institution may be that different centers care about different aspects of the climate and use different data sets and metrics to judge model "quality" during development. In practice, that hypothesis cannot be tested with the currently available models but could be explored by calibrating structurally different and computationally inexpensive models to different data sets using different metrics.

[14] Confidence in model projections does not come from the sheer amount of code and data and cannot be demonstrated by repeated verification. Confidence is greatest in those aspects of climate change that we understand and can link back to known physical processes and simpler models and concepts (e.g., the global energy balance). Scientific insight into the models and their development is crucial. We argue that transparency in the model development process is helpful in understanding model evaluation and projections and contributes to that insight. Efforts are already made to document models and tuning (see Supporting Information). Shared code should be communicated clearly, and the way the model is evaluated and calibrated (metrics and data sets) should be documented where possible. Such information will help to make the best use of the massive amounts of data, for example, by specifically selecting subsets of models for certain applications, and to better understand model differences and uncertainties.

[15] The new generation of global climate models in CMIP5 supports the idea of a rather gradual evolutionary process by which models improve over time. The strengths and weaknesses of particular models, as they are evident in the biases in simulating present day climate, are partly passed on to newer model versions and to other models by the exchange of code and ideas. Sharing model components is not a problem, and we can learn much from it. Diversity is important, and reducing the ensemble to a few models would not be useful. However, approaches to weighting models in projections should not only consider metrics of model performance but could also down weight models that have very similar control biases to avoid biased projections from near duplicate models. Pairwise distance metrics as used here could be one approach, but each projection will require a careful analysis of which metric and variables are important and how the distance in the present relates to the distance in predicted changes.

[16] CMIP5 appears to be a "better CMIP3" rather than a radically new ensemble, also in its climate change response [*Knutti and Sedláček*, 2012]. The results point to a remarkable consistency and robustness in many aspects of simulated present day and future climate, but it also suggests that convergence to reality is slow and radical breakthroughs are hard to achieve. However, the fact that the ensemble compares more favorably with observations, despite the enhanced complexity of many of the models, is perhaps one indication that this development strategy of using enhancements in computational power to add complexity is successful in better representing the current climate system.

[17] **Acknowledgments.** The authors acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and the climate modeling groups for producing and making available their model output. For CMIP, the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. The National Center for Atmospheric Research is supported by the U.S. National Science Foundation.

References

Bryan, K. (1969), A numerical method for the study of the circulation of the world ocean (Reprinted from the Journal of Computational Physics, vol 4, pg 347–376, 1969), *J. Comput. Phys.*, *4*, 347–376.

Gettelman, A., J. Kay, and K. Shell (2012), The Evolution of Climate Sensitivity and Climate Feedbacks in the Community Atmosphere Model, *J. Climate*, *25*(5), 1453–1469.

Gleckler, P., K. Taylor, and C. Doutriaux (2008), Performance metrics for climate models, *J. Geophys. Res.-Atmos.*, *113*(D6), D06104.

Knutti, R. (2008), Should we believe model predictions of future climate change? *Philosophical Transactions of the Royal Society A-Mathematical Physical and Engineering Sciences*, *366*(1885), 4647–4664.

Knutti, R., and J. Sedláček (2012), Robustness and uncertainties in the new CMIP5 climate model projections, *Nature Climate Change*, doi:10.1038/NCLIMATE1716.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl (2010), Challenges in Combining Projections from Multiple Climate Models, *J. Climate*, *23*(10), 2739–2758.

Masson, D., and R. Knutti (2011), Climate model genealogy, *Geophys. Res. Lett.*, *38*, L08703, doi:10.1029/2011GL046864.

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*, edited by S. Solomon, D. Qin, M. Manning, Z. Chen,

M. Marquis, K. B. Averyt, M. Tignor, and H.L. Miller, Cambridge Univ. Press, Cambridge, UK, and New York, NY, USA.

Reichler, T., and J. Kim (2008), How well do coupled models simulate today's climate? *Bull. Am. Meteorol. Soc.*, *89*(3), 303–311.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl (2012), A Summary of the CMIP5 Experiment Design, *Bull. Amer. Meteor. Soc*, *93*, 485–498.

Tebaldi, C., and R. Knutti (2007), The use of the multi-model ensemble in probabilistic climate projections, *Philosophical Transactions of the Royal Society A-Mathematical Physical and Engineering Sciences*, *365*(1857), 2053–2075.

Watanabe, M., H. Shiogama, T. Yokohata, Y. Kamae, M. Yoshimori, T. Ogura, J. Annan, J. Hargreaves, S. Emori, and M. Kimoto (2012), Using a Multiphysics Ensemble for Exploring Diversity in Cloud-Shortwave Feedback in GCMs, *J. Climate*, *25*(15), 5416–5431.

Weigel, A., R. Knutti, M. Liniger, and C. Appenzeller (2010), Risks of Model Weighting in Multimodel Climate Projections, *J. Climate*, *23*(15), 4175–4191.

Yokohata, T., J. D. Annan, J. C. Hargreaves, C. S. Jackson, M. Tobis, M. Webb, and M. Collins (2012), Reliability of multi-model and structurally different single-model ensembles, *Clim. Dyn.*, *39*, 599–616.