Climate model genealogy

D. Masson[1] and R. Knutti[1]

Received 30 January 2011; revised 16 March 2011; accepted 22 March 2011; published 23 April 2011.

[1] Climate change projections are often given as equally weighted averages across ensembles of climate models, despite the fact that the sampling of the underlying ensembles is unclear. We show that a hierarchical clustering of a metric of spatial and temporal variations of either surface temperature or precipitation in control simulations can capture many model relationships across different ensembles. Strong similarities are seen between models developed at the same institution, between models sharing versions of the same atmospheric component, and between successive versions of the same model. A perturbed parameter ensemble of a model appears separate from other structurally different models. The results provide insight into intermodel relationships, into how models evolve through successive generations, and suggest that assuming model independence in such ensembles of opportunity is not justified.

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

1. Introduction

[2] Uncertainty in climate model projections is often characterized by some measure of spread across an ensemble of simulations [*Furrer et al.*, 2007; *Tebaldi et al.*, 2005; *Tebaldi and Knutti*, 2007]. The results thus depend on the range of responses covered by the models, and the distribution of the models within that range. The most recent coordinated ensemble used here is from the World Climate Research Project (WCRP) Coupled Model Intercomparison Project Phase 3 (CMIP3) [*Meehl et al.*, 2007]. A common assumption in such ensembles is that the set of models reflects the uncertainty in how to best describe the climate system in a model, arising partly from the difficulty in defining a unique model quality metric [*Parker*, 2006; *Tebaldi and Knutti*, 2007; *Knutti*, 2008; *Knutti et al.*, 2010b, 2010a]. Whether models span the full uncertainty range is hard to verify or falsify. The other assumption is that the models can be considered as independent in the sense that every model contributes additional information. All models of course contain common elements (e.g., the equations of motion) because they describe the same system, and they produce similar results. But if they make the same simplifications in parameterizing unresolved processes, use numerical schemes with similar problems, or even share components or parts thereof (e.g., a land surface model), then their deviations from the true system or other models will be similar. In the extreme case, a model run at two resolutions, or the same model run twice with two initial states provide very little additional information about climatology or a decadally averaged projection. We qualitatively define an additional model as dependent if it provides little insight into why and how models differ from each other in the existing ensemble, and from observations. While statistically convenient, the assumption of independence is unlikely to be fully justified. Successful concepts in models are often copied or inherited, some institutions have used whole components from other models. Models are evaluated against the same observations, often using similar metrics. In CMIP3 several modeling groups have submitted two or three models. Most intercomparisons are thus ensembles of opportunity in which the sampling and dependence in the model space is unknown.

2. Method

[3] The metric used to quantify the distance between unforced control simulations of two models is based on the Kullback-Leibler divergence that takes into account the full spatial field of monthly values in a control simulation. It thus considers the mean state, the seasonal cycle, the interannual variations, as well as the spatial correlation. A hierarchical clustering applied to the distance matrix of pairwise model dissimilarities produces a ‘family tree’ of the models. The position at which the tree connects two models (relative to zero) characterizes the disagreement between the simulated control climate of two models, the vertical ordering of the branches is arbitrary. In addition, three reanalysis datasets (ERA, NCEP and MERRA) and two precipitation datasets (GPCP, CMAP) are treated like additional models. The details of the statistical method as well as the CMIP3 and the reanalysis and observation datasets are described in the auxiliary material. Note that in contrast to earlier work [*Jun et al.*, 2008a; *Knutti et al.*, 2010b; *Pennell and Reichler*, 2011] this method does not analyze pairwise correlation of model errors to observations, but simply the similarity of two models as expressed by the similarity of their temperature and precipitation fields. Observations are included here as ‘additional models’ just for illustration. The method and all conclusions are independent of whether the ensemble is interpreted as models being centered around truth or models and truth being indistinguishable, because the method only uses pairwise distances between models [*Knutti et al.*, 2010a; *Annan and Hargreaves*, 2010].

3. Results and Discussion

[4] The tree in Figure 1 shows that models from the same institution in almost all cases are very similar (e.g., GISS, MIROC, CCCMA, GFDL, UKMO, CSIRO). The degree of similarity varies and is more pronounced for example for GFDL than for GISS. Some of these pairs are not surprising, e.g., the two CCCMA models only differ in resolution. Others like the two UKMO for temperature are more surprising. Some characteristics seem to be preserved that keep

[1] Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland.

0094-8276/11/2011GL046864

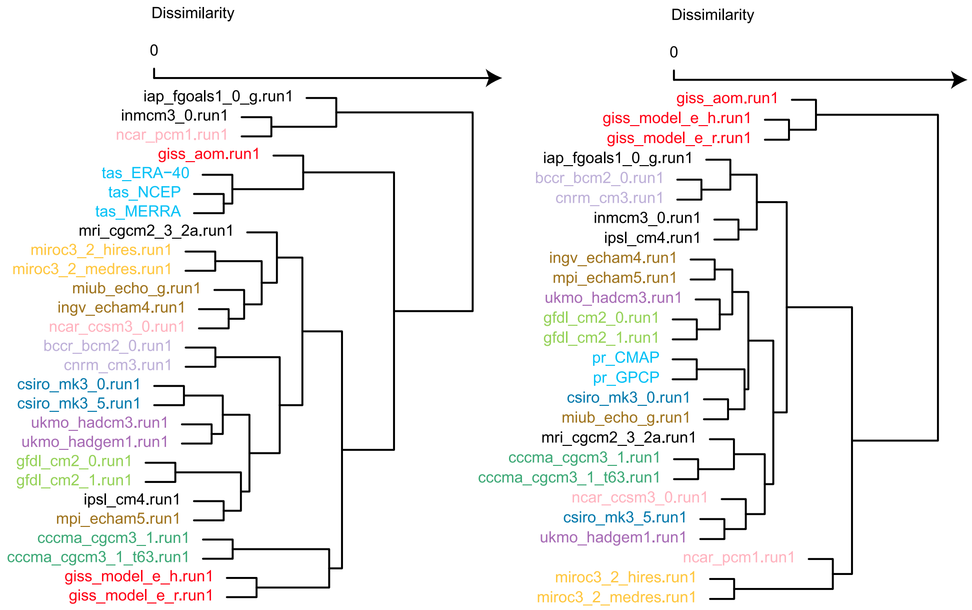

Figure 1. Hierarchical clustering of the CMIP3 models for (left) surface temperature and (right) precipitation in the model control state. Models from the same institution and models sharing versions of the same atmospheric model are shown in the same color. Observations also are marked by the same color. Models without obvious relationships are shown in black.

these two models close, despite significant changes that were made to most components of the model. But relationships go beyond the "same modeling center" attribute. MIUB and ECHO-G (both based on an ECHAM4 atmosphere) cluster for temperature, and INGV-ECHAM4 and MPI-ECHAM5 (both ECHAM based but with different versions) cluster for precipitation. A less evident pair, BCCR-BCM2-0 and CNRM-CM3, is identified for both temperature and precipitation. These models share the same atmosphere and land components.

[5] *Pennell and Reichler* [2011] performed a similar analysis using hierarchical clustering but with a different distance metric based on model biases and 35 climate variables. While their results for CMIP3 are similar to those presented here, we show that a single variable (thus avoiding normalization) is sufficient to reveal most of the dependency structure, and that the key elements of dependence are similar for both surface temperature and precipitation. Observation and reanalysis datasets are not needed for the analysis, but when included like additional models they also cluster together, with some distance to the models, but well within the bulk of the simulations.

[6] The picture gets even more interesting when the QUMP perturbed physics ensemble [*Collins et al.*, 2010] and the previous generation of models in CMIP2 is included, shown in Figure 2. The CMIP2 and CMIP3 models from the same institution also tend to cluster. For precipitation for example, the old NCAR CSM, PCM1 and the NCAR-WM models are close. The newest NCAR CCSM3 in CMIP3 however was developed almost independently from earlier NCAR models and appears separated. Qualitatively, the history can be traced back further for most models [*Edwards*, 2010]. But given the rapid development, the increase in resolution in the models, the inclusion of new processes and the availability of more observations, we believe the connections between successive model versions are unlikely to persist over more than one or two generations.

[7] In most of the trees, there is no clear separation into two or three clusters that are far apart, i.e., there is no evidence for multiple classes of models, different mutually exclusive theories or philosophies in how to build a model, or a clear separation between CMIP2 and CMIP3. The climate model landscape rather resembles an evolutionary process. Individual models take small steps compared to the size of the model space, successful pieces of a model are kept, inherited and copied and less successful parts go extinct. Existing models adapt to new environments (computer architecture and capacity, new observations, improved understanding of the climate system), although by deliberate rather than random modifications. New models rarely are written from scratch but evolve from combining, modifying and improving existing parts and ideas.

[8] The perturbed versions of the HadCM3 [*Collins et al.*, 2010] model separate themselves from the rest of the CMIP models. For some aspects, a large PPE can span a "model space" similar or larger than CMIP3, e.g., for the range of feedbacks and climate sensitivity [*Sanderson et al.*, 2010; *Collins et al.*, 2010; *Stainforth et al.*, 2005]. However, if the full spatiotemporal fields are considered, the underlying model structure (grid, numerical scheme, parameterizations, resolved processes) appears to be important. Note that parameter perturbations in the QUMP ensemble are chosen to maximize the spread in feedbacks but ensure good agreement with climatology for each member (see auxiliary material). Very different unconstrained model versions are likely to exist, and those may well fall outside the QUMP cluster.

4. Conclusions

[9] Our analysis of spatial and temporal variations in surface temperature and precipitation shows strong similar-

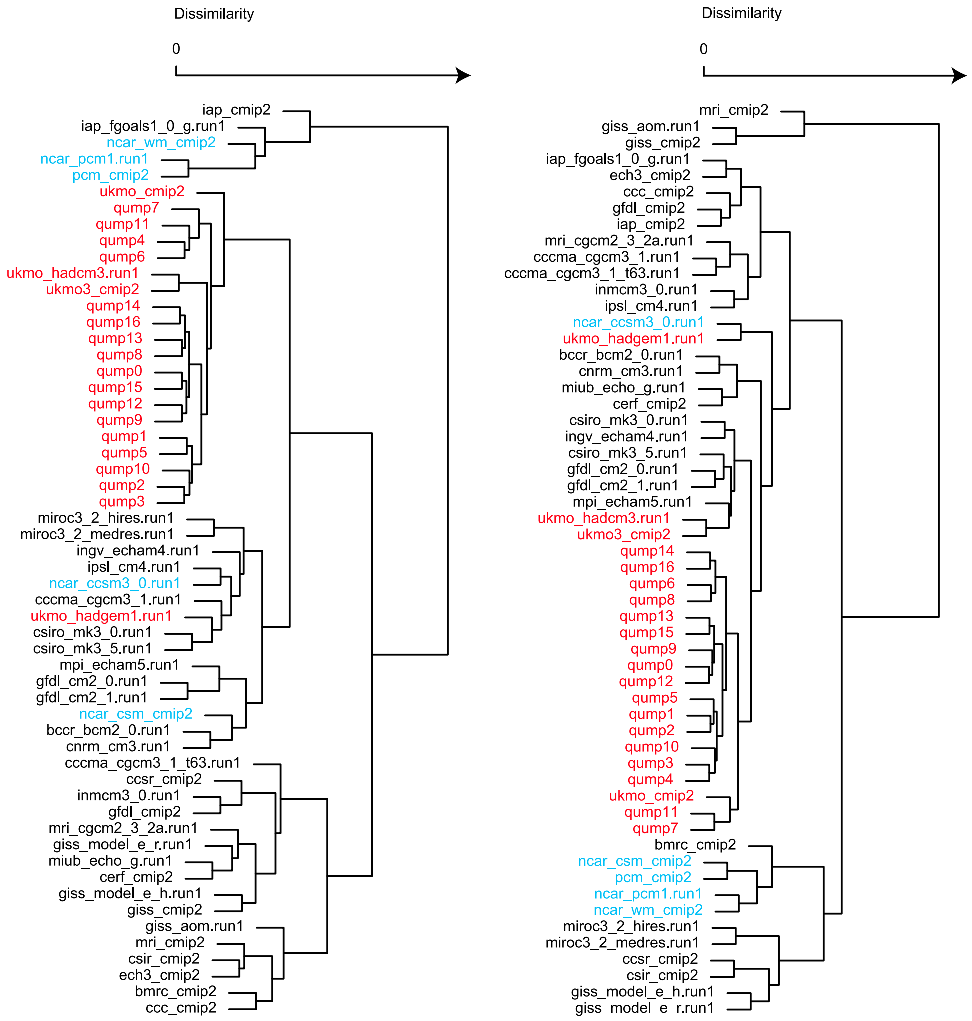

Figure 2. Hierarchical clustering of the CMIP3, CMIP2 and QUMP perturbed physics ensemble for (left) surface temperature and (right) precipitation in the model control state. Models developed by the UK Metoffice Hadley Centre are shown in red, models developed by NCAR are marked in blue.

ities in groups of two or three models, supporting earlier claims that the effective number of independent models is smaller than the actual number of models in the multi model ensemble [*Pirtle et al.*, 2010; *Pennell and Reichler*, 2011; *Jun et al.*, 2008b; *Tebaldi and Knutti*, 2007; *Knutti*, 2008; *Knutti et al.*, 2010b; *Knutti*, 2010]. Models developed at the same institution show the most striking similarities, but dependencies are even visible between two models that use different versions of the same atmosphere. Ensembles of different generations as well as observations largely overlap, suggesting a gradual evolutionary development and refinement of models. For the metric and variables chosen here, structural model differences seem to be important. We interpret this as an indication that sampling different model structures is important to capture the full range of model behavior.

[10] Correlations between the control state and the projected change across models are generally weak [*Knutti et al.*, 2010b], implying that a mapping of the dependence structure into projections is difficult (see auxiliary material for a discussion of clustering projections), i.e., two models that have similar biases in the present may not have similar projection errors in the future. It is therefore unclear whether the dependence in the control state implies that uncertainty in projections is underestimated, or that the number of effective models in the future is the same as in the present.

[11] Using equal weights for all models to create a most likely projection fails to take into account model dependencies. If a group of similar models is part of the ensemble, either from small changes in parameters or resolution, this poses a risk of double counting and giving undue weight to the structure underlying the group. Given the large cost of model development, the availability of open community models and the broader availability of supercomputers, model variations of existing models and perturbed physics ensembles [*Sanderson et al.*, 2010; *Collins et al.*, 2010] may become more common in the future, making this dependence a much bigger issue.

[12] The goal for an ensemble should be to maximize diversity in models yet ensure good performance for all members, and minimize dependency. In principle, this could be achieved with a sufficiently large and broad ensemble to start with, and an appropriate weighting that takes into account two distinct factors: performance metrics measuring model skill, and the dependence to other models to account for sampling problems demonstrated above. In practice, this

proves to be very difficult. For the former, the obvious problem is the lack of repeated verification to define skill for the forecast quantity of interest [*Knutti et al.*, 2010b; *Tebaldi and Knutti*, 2007; *Knutti et al.*, 2010a]. Skill therefore has to be indirectly determined by relating the forecast to metrics based on past trends and climatology. With few exceptions [e.g., *Stott et al.* 2006] such relationships are weak [*Knutti et al.*, 2010b], making the definition of model weights ambiguous. If weights are incorrectly specified, the forecast is likely to be worse than if no weighting was used, in particular for small ensembles [*Weigel et al.*, 2010]. Accounting for model performance may be possible in some cases, e.g., in the Arctic where several metrics of present day climate and past trends are clearly related to future warming and sea ice decline, and where the underlying processes are well understood [*Boé et al.*, 2009b, 2009a]. If the identified model biases lead to biased predictions [*Stroeve et al.*, 2007], it would seems stupid not to consider the observed evidence to improve projections and estimate uncertainties.

[13] Even if the general formulation of an unambiguous weighting scheme for various regions, variables and timescales that takes into account model performance and dependence appears to be a long way off, a few conclusions are obvious. First, there is a lively debate in the community on the point of model weighting [*Knutti*, 2010], but the issue of sampling in ensembles has received very little attention. Second, diversity is critical. The number of structurally different models is small, and maintaining a sufficiently large set of reasonably independent models that span a wide range of plausible assumptions and scientific viewpoints is important both to quantify uncertainty and to understand model differences [*Knutti*, 2010]. Eliminating a model from an analysis is easy, extrapolation beyond the range covered by the ensemble is nearly impossible. Third, models are rarely built with lasting value as a primary goal [*Held*, 2005] and are superseded by newer models. Yet to understand why models and their projections differ, archiving results from older model versions and common scenarios would help. Fourth, conclusions drawn from ensembles should at least test the sensitivity to how models are selected in the ensemble. Current coordinated model experiments are like asking the same question to a small number of people, without thinking about how to select those people, how many to ask, and how to account for the fact that they may have similarly biased opinions. This undoubtedly makes the interpretation of the answers challenging.

[14] **Acknowledgments.** We thank Thomas Reichler for CMIP2 and Mat Collins for the Hadley Centre QUMP data. We acknowledge the modelling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled Modelling (WGCM) for their roles in making available the WCRP CMIP3 multi-model dataset. Support of this dataset is provided by the Office of Science, U.S. Department of Energy (DOE).

[15] The Editor thanks the two anonymous reviewers for their assistance in evaluating this paper.

References

Annan, J. D., and J. C. Hargreaves (2010), Reliability of the CMIP3 ensemble, *Geophys. Res. Lett.*, *37*, L02703, doi:10.1029/2009GL041994.

Boé, J., A. Hall, and X. Qu (2009a), Deep ocean heat uptake as a major source of spread in transient climate change simulations, *Geophys. Res. Lett.*, *36*, L22701, doi:10.1029/2009GL040845.

Boé, J. L., A. Hall, and X. Qu (2009b), September sea-ice cover in the Arctic Ocean projected to vanish by 2100, *Nat. Geosci.*, *2*(5), 341–343.

Collins, M., B. B. B. Booth, B. Bhaskaran, G. Harris, J. Murphy, D. Sexton, and M. Webb (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.

Edwards, P. N. (2010), *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*, 528 pp., MIT Press, Cambridge, Mass.

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, doi:10.1029/2006GL027754.

Held, I. M. (2005), The gap between simulation and understanding in climate modeling, *Bull. Am. Meteorol. Soc.*, *86*(11), 1609–1614, doi:10.1175/BAMS-86-11-1609.

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

Jun, M. Y., R. Knutti, and D. W. Nychka (2008b), Spatial analysis to quantify numerical model bias and dependence: How Many climate models are there?, *J. Am. Stat. Assoc.*, *103*(483), 934–947.

Knutti, R. (2008), Should we believe model predictions of future climate change?, *Philos. Trans. R. Soc. A*, *366*(1885), 4647–4664.

Knutti, R. (2010), The end of model democracy?, *Clim. Change*, *102*(3–4), 395–404.

Knutti, R., G. Abramowitz, M. Collins, V. Eyring, P. J. Gleckler, B. Hewitson, and L. Mearns (2010a), Good practice guidance paper on assessing and combining multi model climate projections, in *Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections*, edited by T. Stocker et al., Univ. of Bern, Bern.

Knutti, R., R. Furrer, C. Tebaldi, and J. Cermak (2010b), Challenges in combining projections from multiple climate models, *J. Clim.*, *23*(10), 2739–2758, doi:10.1175/2009JCLI3361.1.

Meehl, G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer, and K. E. Taylor (2007), The WCRP CMIP3 multimodel dataset—A new era in climate change research, *Bull. Am. Meteorol. Soc.*, *88*(9), 1383–1394.

Parker, W. (2006), Understanding pluralism in climate modeling, *Found. Sci.*, *11*, 349–368, doi:10.1007/s10699-005-3196-x.

Pennell, C., and T. Reichler (2011), On the effective number of climate models, *J. Clim.*, doi:10.1175/2010JCLI3814.1, in press.

Pirtle, Z., R. Meyer, and A. Hamilton (2010), What does it mean when climate models agree?, *Environ. Sci. Policy*, *13*, 351–361.

Sanderson, B., K. Shell, and W. J. Ingram (2010), Climate feedbacks determined using radiative kernels in a multi-thousand member ensemble of AOGCMs, *Clim. Dyn.*, *35*, 1219–1236, doi:10.1007/s00382-009-0661-1.

Stainforth, D. A., et al. (2005), Uncertainty in predictions of the climate response to rising levels of greenhouse gases, *Nature*, *433*(7024), 403–406.

Stott, P. A., J. F. B. Mitchell, M. R. Allen, T. L. Delworth, J. M. Gregory, G. A. Meehl, and B. D. Santer (2006), Observational constraints on past attributable warming and predictions of future global warming, *J. Clim.*, *19*(13), 3055–3069.

Stroeve, J., M. M. Holland, W. Meier, T. Scambos, and M. Serreze (2007), Arctic sea ice decline: Faster than forecast, *Geophys. Res. Lett.*, *34*, L09501, doi:10.1029/2007GL029703.

Tebaldi, C., and R. Knutti (2007), The use of the multi-model ensemble in probabilistic climate projections, *Philos. Trans. R. Soc. A*, *365*(1857), 2053–2075.

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. Clim.*, *18*(10), 1524–1540.

Weigel, A., R. Knutti, M. Liniger, and C. Appenzeller (2010), Risks of model weighting in multi-model climate projections, *J. Clim.*, *23*(15), 4175–4191, doi:10.1175/2010JCLI3594.1.

R. Knutti and D. Masson, Institute for Atmospheric and Climate Science, ETH Zurich, Universitätstr. 16, CH-8092 Zurich, Switzerland. (reto.knutti@env.ethz.ch; david.masson@env.ethz.ch)