
Volume 13, Issue 5, August 2010
ISSN 1462-9011
ELSEVIER
Environmental Science & Policy
Editor-in-Chief
James Briden

ENVIRONMENTAL SCIENCE & POLICY 13 (2010) 351–361

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/envsci

What does it mean when climate models agree? A case for assessing independence among general circulation models

Zachary Pirtle [a], *Ryan Meyer* [a,b,*], *Andrew Hamilton* [b]

[a] *Consortium for Science, Policy, and Outcomes, Arizona State University, PO Box 875603, Tempe, AZ 85287-5603, USA*
[b] *School of Life Sciences, Arizona State University, PO Box 874501, Tempe, AZ 85287-4501, USA*

ARTICLE INFO

Published on line 18 May 2010

Keywords:
Models
Prediction
Climate science
Science policy
Robustness

ABSTRACT

Climate modelers often use agreement among multiple general circulation models (GCMs) as a source of confidence in the accuracy of model projections. However, the significance of model agreement depends on how independent the models are from one another. The climate science literature does not address this. GCMs are independent of, and interdependent on one another, in different ways and degrees. Addressing the issue of model independence is crucial in explaining why agreement between models should boost confidence that their results have basis in reality.

1. Introduction

Predicting the Earth's future climate is an important part of climate science (cf. CCSP, 2008; Shukla et al., 2009).[1] This task is undertaken by combining scenarios of socioeconomic change with highly complex models that simulate the atmosphere and other key Earth systems. Many believe that in order to be useful, these general circulation models (GCMs) will need to provide accurate, reliable, and timely results on a regional to local scale (for example, see a recent declaration by the World Modelling Summit for Climate Prediction (Shukla et al., 2009)). Some question whether climate prediction is a necessary ingredient for decision making,[2] but prediction-based decision making remains a powerful motivator for present and future investments in climate science. As the promised role of GCMs shifts from ''demonstrating warming'' to supporting important regional and local decisions with predictive models, uncertainty will rise, and so will the political stakes.

This paper focuses on a conceptual issue surrounding climate models and the interpretation of their results that is also relevant to science policy and epistemology. In short, we note that GCMs are commonly treated as independent from one another, when in fact there are many reasons to believe otherwise. The assumption of independence leads to increased confidence in the ''robustness'' of model results when multiple models agree. But GCM independence has not been evaluated by model builders and others in the climate science community. Until now the climate science literature has given only passing attention to this problem, and the field has not developed systematic approaches for assessing model independence.

* *Corresponding author at*: Consortium for Science, Policy, and Outcomes, Arizona State University, PO Box 875603, Tempe, AZ 85287-5603, USA Tel.: +1 415 830 3621.
E-mail address: ryan.meyer@asu.edu (R. Meyer).

[1] We use the word ''predict'' in a very specific sense, denoting the production of deterministic information about some future reality. This may be distinguished from the ideas of generalization, extrapolation, or physical laws with abstract predictive power.

[2] Even accurate and reliable predictions are not necessarily useful for policy and other kinds of decision making. See cases in Sarewitz et al. (2000), Hulme et al. (2009), and Dessai et al. (2009). Rather than address this argument, our paper focuses on a fundamental issue related to the generation of reliable and accurate predictions. The question of whether increased predictive accuracy among GCMs would be useful to one group or another is an important area of study.

1462-9011/$ – see front matter
doi:10.1016/j.envsci.2010.04.004

In raising the question, ''what does it mean when climate models agree?'' this paper aims to:

1. demonstrate that this question is of crucial importance to climate science;
2. point out that the question has nonetheless been largely ignored by climate scientists;
3. evaluate what few attempts have been made to address the question thus far; and
4. discuss what it would take to address the question more thoroughly.

As we make clear below, we cannot provide a definitive answer to the question raised in our title. We argue only that a satisfactory answer must provide an account of the independence of models from one another, and that coming up with such an answer will take the combined efforts climate scientists, and perhaps scholars in other fields as well.

In the next section we spell the independence problem out in detail, and in Section 3 we discuss the few instances in which the climate literature has addressed the topic. In Section 4 we use the history of GCM development, along with some contemporary examples of particular model studies to illustrate various ways climate models may or may not be independent from one another. In a final section, we address the broader significance of this issue for climate science, and draw conclusions. The main lesson is that agreement of model results is not necessarily robustness because agreement without independence is not confirmatory. It is crucially important for climate policy, and climate science policy, that modelers give attention to the relationships between the sets of models they consider.

2. Models and reality

The logistics of building GCMs that simulate the behavior of complex interacting natural systems with multiple feedbacks give rise to an important and well-known problem: in order to build a tractable model, it is often necessary to make dramatic assumptions and simplifications. This is a reality for any numerical model of a complex system. Examples from climate science are the continuum assumption,[3] and the assumption that landmasses on Earth reach only 3.5 m in depth, as in the GISS model.[4] These assumptions may be adequate, appropriate, and even necessary for modeling purposes, but they are clearly false. Simplifying assumptions are the price of tractability, and in some cases, they are also necessary for precision.

One challenge facing modelers and policy makers is to understand the ways in which the inherent limitations on models impact their relationship with reality. As population biologist Richard Levins (1966, p. 421) has pointed out, many modelers of biological systems generally do not even try to capture the world as it is, because doing so would mean ''using perhaps 100 simultaneous partial differential equations with time lags; measuring hundreds of parameters, solving the equations to get numerical predictions, and then measuring these predictions against nature.'' Furthermore, Levins notes, more realistic models would require attention to many more salient parameters than we can measure, as well as the use of equations that have no analytic solutions and no outputs that we know how to understand. According to Levins, these constraints leave mathematical population biologists working at some distance from the phenomena they are studying, and creates the problem that it is sometimes difficult to know whether model outputs reflect their system of study appropriately or depend instead on the ''details of the simplifying assumptions'' (Levins, 1966, p. 423). He has written ''...we attempt to treat the same problem with several alternative models each with different simplifications but with a common biological assumption. Then, if these models, despite their different assumptions, lead to similar results, we have what we call a robust theorem which is relatively free of the details of the model. Hence our truth is the intersection of independent lies'' (Levins, 1966).

The challenges described by Levins also face climate modelers, whose GCMs are at a far enough remove from reality that it is sometimes difficult to discern whether particular results are artifacts of the modeler's practice or whether the model captures and explains real patterns and processes. Indeed, GCMs arguably have more parameters, more equations in need of solutions, and more solutions that have to be tested against nature than do the models of population biology that Levins discusses.

Climate modelers, of course, are well aware of all this, so they not only compare their model results against observations, they also check their models against other models. In a recent paper, for instance, Seager et al. (2007) assert that eighteen of nineteen general circulation models (GCMs) agree that the Southwestern United States will experience increased aridity and drought over the next one hundred years. The notion that agreement across models increases confidence in particular aspects of model projections is common. Take, for instance, this statement about climate warming from the fourth assessment report of the Intergovernmental Panel on Climate Change (IPCC AR4):

> ...models are unanimous in their prediction of substantial climate warming under greenhouse gas increases, and this warming is of a magnitude consistent with independent estimates derived from other sources, such as from observed climate changes and past climate reconstructions. (Solomon et al., 2007 p. 601)

Here the IPCC authors are reporting agreement among models, and counting this agreement as one reason (among

[3] Commonly used in fluid mechanics, the continuum assumption ignores the fact that physical matter is actually made up of individual particles that collide with one another, and assumes that properties such as pressure and temperature are consistent. This is necessary for the application of the most common mathematical equations, without which problem solving would in many cases be impossible.

[4] Information on the model developed at the Goddard Institute for Space Studies, along with other models referenced throughout this paper can be accessed through the Program for Climate Model Diagnosis and Intercomparison (http://www-pcmdi.llnl.gov/ipcc/model_documentation/ipcc_model_documentation.php). That website includes specifications for all 24 of the models used in the IPCC's Fourth Assessment Report.

many others) to believe that the models are correct in their projections. In a rough survey of the contents of six leading climate journals since 1990,[5] we found 118 articles in which the authors relied on the concept of agreement between models to inspire confidence in their results. The implied logic seems intuitive: if multiple models agree on a projection, the result is more likely to be correct than if the result comes from only one model, or if many models disagree.

As the above quote from Levins indicates, this logic only holds if the models under consideration are independent from one another. In Levins' terms, using multiple models to analyze the same system is a ''robustness'' strategy. Every model has its own assumptions and simplifications that make it literally false in the sense that the modeler knows that his or her mathematics do not describe the world with strict accuracy. When multiple *independent* models agree, however, their shared conclusion is more likely to be true. As Levins puts it, ''our truth is the intersection of independent lies.''

By contrast, the more dependent the models are on one another, the less one should be impressed by their intersecting results. No one is or should be surprised when very similar microscopes yield very similar observations after being trained on the same slide. This consideration has not gone completely unnoticed among climate modelers. In an extensive review of climate model reliability, Raisanen (2007, p. 9) laments the ''limited number of quasi-independent climate models,'' noting later that ''the risk that the uncertainty in the real world exceeds the variation between model results is obvious: even if all models agreed perfectly with each other, this would not prove that they are right'' (see also Tebaldi and Knutti, 2007).

A common response to this problem is to compare model results to past climate observations, but this measure of performance presents its own set of analytical problems. A match between model results and observations guarantees neither an ability to predict future system behavior, nor that the model accurately captures underlying causal mechanisms (Oreskes et al., 1994; Tebaldi and Knutti, 2007).[6] While there are many ways to assess the performance of a GCM, there is little consensus on which are most effective, or on how to determine the most appropriate measures for a given set of circumstances (Gleckler et al., 2008).

As evidenced in the results of our literature search (mentioned above), and its use by the IPCC, agreement across models has become an important tool for evaluating the robustness of model outputs. Many see agreement among models as support for the assumption that they represent the world in the relevant ways.

3. Robustness in the GCM literature

How much and in what ways should models differ from one another for modelers to have increased confidence in convergent results?[7] What does it mean to say that a model or set of models is robust? The climate modeling literature has devoted little attention to the problem of model interdependence. We have found no discussion of the ways and degrees to which climate models *should* be independent of one another for agreement among them to constitute confirmatory evidence of their outputs. Though some have recognized this as an issue (e.g. Abramowitz, 2010; Collins, 2007; Raisanen, 2007; Tebaldi and Knutti, 2007), we have found only one limited attempt (Abramowitz and Gupta, 2008, more on this later) at developing a method for evaluating the independence of models. The IPCC's Fourth Assessment Report only brushes against concepts related to robustness and independence.[8]

Below we summarize three instances in which scientists have addressed model independence: (1) the ensemble approach, which assumes that models differ due to errors in coding and processing (e.g. ''truncation errors''); (2) in the use of differing model outputs to indicate model independence; and (3) an argument that the range of model parameterizations should accurately reflect current uncertainties surrounding those parameters.

3.1. *Ensembles*

For the most part, discussions of model independence occur in the context of model ensembles. Ensembles, in which results are averaged across a group of models, are increasingly used for their ability to reproduce observed climate behavior accurately (Hagedorn et al., 2005; Lambert and Boer, 2001; Palmer et al., 2004; Wang and Swail, 2006). As with comparison of model results (such as with the Seager et al. study mentioned previously), the ensemble approach relies on the assumption of independence among models. Hagedorn et al. (2005) describe this assumption:

> Every attempt to represent nature in a set of equations, resolvable on a digital computer, inevitably introduces inaccuracy. That is, although the equations for the evolution of climate are well understood at the level of partial differential equations, they have to be truncated to a finite-dimensional set of ordinary differential equations, in order to be integrated on a digital computer... The basic idea of the multi-model concept is to account for this inherent model error in using a number of independent and skilful models in the hope of a better coverage of the whole possible climate phase space.

This account implicitly assumes model independence exists due to different cut-off errors resulting from the coding of differential equations within each model. But this assump-

[5] Geophysical Research Letters; the Journal of Geophysical Research-Atmospheres; the Journal of Atmospheric Sciences; the Journal of Climate; Climate Dynamics; and Climatic Change.

[6] Indeed, models, or their subcomponents, may be ''tuned'' specifically to match past observations. See discussions of tuning in Tebaldi and Knutti (2007).

[7] Simply assessing whether or not (or the degree to which) models agree with one another is difficult and complicated. See Raisanen (2007), and citations therein for more on this. It is also worth noting that as models become more complex, so too does the task of assessing their independence. With simple models one can see much more easily the ways and degrees to which models are independent (IPCC, 1997).

[8] See the discussion and justification of the use of multi-model ensembles in IPCC AR4 (Solomon et al., 2007, pp. 805–806).

tion ignores the important question of whether the use of different coding schemes, with correspondingly different truncation errors, is enough to constitute independence. Truncation error is not the only significant source of error and uncertainty within our understanding of the climate, nor is it the sole dimension along which models can be deemed independent.

The focus of our argument is on modelers' claims that agreement among models increases the robustness of their results. The ensemble approach would benefit from a clearer account of model independence, which is only implicitly assumed in the ensemble literature.

3.2. *Independence based on differing results*

In one of the few direct examinations of model independence, Abramowitz and Gupta (2008) argue that, given that models may share similar biases, the use of multiple model ensembles should involve some understanding of the extent to which models differ from one another, and thus the extent to which they may share the same biases. To demonstrate, they analyze different land-component models using a state-space approach, examining how each model behaves under different combinations of initial conditions. Abramowitz and Gupta define a "distance," or degree of independence between models based upon the divergence of their outputs. This enables a quantitative measure of the "distance" between the different land-use models examined in the study.

The underlying motivations of Abramowitz and Gupta are directly in line with our central argument. However, their solution only evaluates differences in model *results*, rather than differences in the way the models represent the world and its causal structure. They essentially treat models as black boxes, ignoring the causal reasons for disagreement between models. It is possible that two models could agree with respect to outputs despite their having different causal assumptions, but such a result, using this approach, would falsely indicate model "dependence," because these models would yield the same output despite the fact that they make different and possibly conflicting claims about the underlying mechanisms. A more thorough analysis of why models disagree should complement the output-oriented independence metric of Abramowitz and Gupta with an understanding of the causal assumptions of the models, though we recognize that the technical and cognitive difficulties involved in doing so might be significant.

Abramowitz and Gupta encounter another difficulty in balancing concerns for independence with concerns to accurately predict known observations. If two models are accurate in predicting observed results, then they would be deemed dependent on one another by an outcome-oriented metric, and thus the significance of their agreement would be minimized. As Abramowitz and Gupta acknowledge, it is a difficult task to balance concerns for independence with concerns for performance. To try to resolve this, they state that the utility of a given model is a function of the model's performance multiplied by how different it is from all of the other models, divided by the utility given for all of the other models. Again, this makes the problem clearer but does not resolve it. If two models have a large "distance" between them because their outputs disagree in a number of hypothetical cases, but the models agree with observed results, does this increase confidence that their agreement is significant? Or does it indicate that their agreement might be a special case (or perhaps the result of error)? Looking at where the models diverge under different possible conditions and understanding why they disagree would provide the key to understanding the significance of their agreement.

3.3. *Parameters and uncertainty*

Discussion by Tebaldi and Knutti (2007) points to a third way of examining independence among multiple models, by focusing on the different parameters that are encoded within models. They raise a concern that the parameters for a given process within a GCM may not represent the full range of scientific uncertainty about that process:

> Once scientists are satisfied with a model, they rarely go back and see whether there might be another set of model parameters that gives a similar fit to observations but shows a different projection for the future. In other words, the process of building and improving a model is usually a process of convergence, where subsequent versions of a model build on previous versions, and parameters are only changed if there is an obvious need to do so...[I]t is probable that the range covered by the multi-model ensemble covers a minimum rather than the full range of uncertainty. (Tebaldi and Knutti, 2007, p. 2069)

Tebaldi and Knutti point to different parameters as sources of independence among models, suggesting that it could be possible to choose different parameters within model ensembles in ways that better reflect the scientific uncertainties surrounding model parameters. An example of this is the perturbed physics ensembles approach, in which very large ensembles are used to investigate the effects of slight changes in parameters on model output (e.g. Ackerley et al., 2009; Stainforth et al., 2005). Pushing different models to reflect different probability distributions for specific parameters can generate more context for understanding uncertainty in model predictions, and perhaps even differences across GCMs. However, as Tebaldi and Knutti recognize, this cannot fully address the problem, as it ignores structural uncertainties stemming from the somewhat subjective art of selecting parameters for inclusion in the model.

The treatment of models as independent from one another is widespread. Yet this brief review of the literature reveals only a limited recognition that this assumption needs investigation, and an almost complete absence of methods for doing so. The nascent existing literature on model independence supports our argument that the climate science community needs to address this issue more explicitly.

4. Climate models in historical and conceptual perspective

In addition to these concerns about independence raised by climate modelers, both the history of climate modeling and a look at cases of ensemble-modeling show further reasons for skepticism about the independence of many GCMs from one another, and for a better conceptual grasp on independence. In the next subsection we draw on the work of Paul Edwards to offer a quick history of climate modeling, pointing to some of the ways the various projects shared data, personnel, and computer code with one another. Sections 4.2 and 4.3 contain detailed examinations of two ensemble-modeling exercises and shift concern about independence in the abstract toward a discussion of independence in particular cases.

4.1. The history of GCM development

While serious attempts at numerical weather prediction date from early in the 20th century, working GCMs were not developed until the mid-1950s. These early models were apparently developed more or less independently of one another, and all were based on based on Lewis Fry Richardson's simplified versions of Vilhelm Bjerknes' seven ''primitive equations'' (Bjerknes, 1921). These equations could, in principle, capture atmospheric physics well enough to predict general circulation. By the 1960s, there were four efforts to develop realistic GCMs, two of which eventually became institutionalized as the present-day Geophysical Fluid Dynamics Laboratory and the National Center for Atmospheric Research. In the next decades, many more modeling groups developed around the world, but crucially, they did so in close relationships with the existing groups. According to Paul Edwards (2000, p.79):

> GCMs and modeling techniques began to spread by a variety of means. Commonly, new modeling groups began with some version of another group's model. Some new groups were started by post-docs or graduate students from one of the three original GCM groups. Others built new models from scratch.

By the mid-1980s, still more groups were building GCMs with increasing levels of sophistication. GCMs then included coupled ocean–atmosphere models, representations for clouds, higher resolution grid scales, and new, faster numerical methods for writing and solving equations. Building GCMs *de novo*, however, would have been expensive and very time consuming, so virtually no one did. *De novo* codings remain rare.

The history of GCMs indicates that a few prominent research groups have contributed greatly to defining the field. Some of the historically prominent modeling groups still exist today, along with new groups from around the world, and have made major contributions to the IPCC-AR4. The long-standing and tight relationships among many GCMs, combined with the constant interaction among researchers at contemporary modeling centers, raises questions about independence and interdependence. If these models are independent, it is not because they are separately housed or administered: modeling groups often share ideas, data, personnel, and computer code. At the same time, increased collaboration among scientific research groups may aid in the generation of new and different approaches, and it may encourage criticism and increased accuracy in different models. We cannot fully answer what the overall effect is, but it is worth highlighting that the societal interconnections among research groups may add to and at the same time take away from the independence among different modeling groups.

4.2. GCMs used in the IPCC fourth assessment report (IPCC-AR4)

IPCC's AR4 devotes considerable attention to models, their evaluation, and their results. But it does not discuss which kinds and degrees of differences across models point to the kind of model independence that would make agreement between model predictions significant. This presents a problem to anyone who would make use of model results, whether for further research, or for important policy decisions. The authoritative and comprehensive account of climate science gives very little attention to model independence beyond citing a few articles relating to the justification of multi-model ensembles. Directly examining the models used in the IPCC-AR4 strengthens the argument that model independence needs to be more thoroughly explored. The IPCC AR4 drew upon the results of the 23 GCMs, listed in Table 1. Analysis of the table and the models it references reveals several ways in which the models are both similar and dissimilar: there are variations in grid resolution (though some are the same); some model the basic physical equations using different methods (finite differencing as against the spectral-transform approach), though some do it the same way; there are probably also subtle differences in coding, though many models use the same or very similar basic code; some models use different parameterizations for specific phenomena, and sometimes differ in the choice of which physical variables to include.

A look at forcing variables included in the AR4 GCMs yields similar results. In Table 2 one can see that some features, such as forcing from the main greenhouse gases, are shared across almost all models. Greater differences occur across representations of aerosol forcings, with considerable variation in the components included, and the manner in which they are represented. For example, almost all models include sulfate, while very few have incorporated urban aerosols.

In addition to these *prima facie* similarities and differences, deeper study of the components of the models raises further questions about independence. Two models considered from the Hadley Centre for Climate Prediction and Research, UKMO-HadGEM1 and UKMO-HadCM3 are very closely related: the former is a newer version of the latter and some of the major personnel overlap. The same is true in the case of the two models (PCM and CCSM3) from the National Center for Atmospheric Research. Other models, like those considered from JAMSTEC and CGCM3.1, are high- and medium-resolution variants of each other, meaning that their parameterizations and basic approaches are identical, but they differ in their spatial resolutions. Still others, like GFDL-CM2.0 and GFDL-CM2.1, appear to vary with respect to discretization schemes for advection, damping schemes, and the value of time steps used for ocean components, but to share everything else.

Table 1 – List of general circulation models included in the IPCC AR4, with key characteristics.

Model details: Originating group(s)	Country	CMIP3 I.D.	Atmospheric resolution (lat/long)	Atmospheric Layers	Oceanic Resolution [1 at/long)	Oceanic Vertical Layers	Grid Type
Beijing Climate Center	China	BCC-CM1	1.9° × 1.9°	16	1.9° × 1.9°	30	
Bjerknes Centre for Climate Research	Norway	BCCR-BCM2.0	1.9° × 1.9°	31	0.5–1.5° × 1.5°	35	
National Center for Atmospheric Research	USA	CCSM3	1.4° × 1.4°	26	0.3–l° × 1°	40	Eulerian spectral transform
Canadian Centre for Climate Modelling and Analysis	Canada	CGCM3.1(T47)	3.75° × 3.75°	31	1.9° × 1.9°	29	Spectral transform
Canadian Centre for Climate Modelling and Analysis	Canada	CGCM3.1(T63)	2.8° × 2.8°	31	1.4° × 0.94°	29	Spectral transform
Meteo-France/Centre National de Recherches Meteorologiqu.es	France	CNRM-CM3	~1.9° × .9°	45	0.5–2° × 2°	31	Semi-lagrangian semi-implicit time integration with 30 mn time-step, 3 h timestep for radiative transfer;
CSIRO Atmospheric Research	Australia	CSIRO-Mk3.0	~1.9° × 1.9°	18	0.8° × 1.9°	31	Spectral for some variables, lagrangian for others, leapfrog
Max Planck Institute for Meteorology	Germany	ECHAM5/MPI-OM	~1.9° × l.9°	31	1.5° × 1.5°	40	Spectral transform method, leapfrong timestep scheme
University of Bonn (Germany), KMA (Korea) and M&D Group[a]	G/K	ECHO-G	~3.9° × –3.9[a]	19	0.5°–2.8° × 2.8°	20	
LASG/Institute of Atmospheric Physics	China	FGOALS-g1 0	~2.8° × 2.8°	26	1° × 1°	16	Finite difference, semi-implicit time
US Dept. of Commerce/NOAA/Geophysical Fluid Dynamics Laboratory	USA	GFDL-CM2.1	2.0° × 2.5°	24	0.3–1° × 1°	->	B-grid scheme
US Dept. of Commerce/NOAA/Geophysical Fluid Dynamics Laboratory	USA	GFDL-CM2.0	2.0° × 2.5°	24	0.3–1° × 1°		B-grid scheme
NASA/Goddard Institute for Space Studies	USA	GISS-AOM	3° × 4°	12	3° × 4°	16	C-grid scheme
NASA/Goddard Institute for Space Studies	USA	GISS-EH	4° × 5°	20	2° × 2°	16	Arakawa B-grid, among others
NASA/Goddard Institute for Space Studies	USA	GISS-ER	4° × 5°	20	4° × 5°	13	Arakawa B-grid, among others
Institute for Numerical Mathematics	Russia	INM-CM3.0	4° × 5°	21	2° × 2.5°	33	Finite difference (arakawa 1972), semi-implicit
Institut Pierre Simon Laplace	France	IPSL-CM4	2.5° × 3.75°	19	2° × 2°	31	Finite difference equations, leapfrog time approach
University of Tokyo, NIES, and JAMSTEC[b]	Japan	MIROC3.2(hires)	~1.1°×1.1°	56	0.2° × 0.3°	47	Spectral transform
University of Tokyo, NIES, and JAMSTEC[b]	Japan	MIROC3.2(medres)	2.8° × 2.8°	20	0.5–1.4° × 1.4°	43	Spectral transform
Meteorological Research Institute	Japan	MRI-CGCM2.3.2	2.8° × 2.8°	30	0.5–2.0° × 2.5°	23	Spectral transform method, leapfrong timestep scheme, semi-implicit.
National Center for Atmospheric Research	USA	PCM	~2.8° × 2.8°	26	0.5–0.7° × 1.1°	40	Eulerian spectral transform
Hadley Centre for Climate Prediction and Research/Met Office	UK	UKM0-HadCM3	2.5° × 3.75°	19	1.25° × 1.25°	20	Arakawa B-grid, hybrid vertical coordinates. Eulerian advection scheme.
Hadley Centre for Climate Prediction and Research/Met Office	UK	UKMO-HadGEMI	~1.3° × 1.9°	38	0.3–1.0° × 1.0°	40	Arakawa C-grid horizontally, Chaney Phillips grid vertically

Sources: IPCC Table 8.1 (Solomon et al., 2007, pp. 597–599) and (PCMDI http://www-pcmdi.llnl.gov/ipcc/model_documentation/ipcc_model_documentation.php).

[a] Meteorological Institute of the University of Bono, Meterological Research Institute of KMA, and Model and Data group.

[b] Center fot Climate System Research (The University of Tokyo), National Institute fot Environment Studies, and Frontier Research Center for Global Change (JAMSTEC).

Table 2 – Model forcing agents. Y indicates that forcing agent is included, 0 indicates that it is not.

Model details	Forcing agents						Forcing agents											
CMIP3 I.D.	Green house gases						Aerosols										Other	
	CO_2	CH_4	N_2O	Stratospheric ozone	Troposphric ozone	CFCs	SO_4	Urban	Black carbon	Organic	Nitrate	1st Indirect	2nd Indirect	Dust	Volcanic	Sea salt	Land use	Solar
BCC-CM1	Y	Y	Y	Y	Y	Y	Y	0	0	0	0	0	0	0	Y	0	Y	Y
BCCR-BCM2.0	Y	Y	Y	Y	Y	Y	Y	Y	0	0	0	0	0	Y	0	Y	Y	Y
CCSM3	Y	Y	Y	Y	Y	Y	Y	0	Y	Y	0	0	0	Y	Y	Y	0	Y
CGCM3,1(T47)	Y	Y	Y	Y	Y	Y	Y	0	0	0	0	0	0	Y	Y	Y	Y	Y
CGCM31(T63)	Y	Y	Y	Y	Y	Y	Y	0	0	0	0	0	0	Y	y	Y	Y	Y
CNRM-CM3	Y	Y	Y	Y	Y	Y	Y	Y	G	0	0	0	0	Y	0	Y	0	0
CSIRO-Mk30	Y	Y	Y	Y	Y	Y	Y	0	0	0	0	0	0	Q	0	0	0	Y
ECHAM5/MPI-OM	Y	Y	Y	Y	Y	Y	Y	0	0	0	0	Y	0	0	0	0	0	0
ECHO-G	Y	Y	Y	Y	Y	Y	Y	0	0	0	0	Y	0	0	Y	0	0	Y
FGOALS-g1.0	Y	Y	Y	Y	Y	Y	Y	0	0	0	0	0	0	0	0	0	0	Y
GFDL-CM2.1	Y	V	Y	Y	Y	Y	Y	0	Y	Y	0	0	0	Y	Y	Y	Y	Y
GFDL-CM2.0	Y	V	Y	Y	Y	Y	Y	0	Y	Y	0	0	0	Y	V	Y	Y	Y
GISS-AOM	Y	Y	Y	Y	Y	Y	Y	0	Y	Y	Y	0	Y	Y	y	Y	Y	Y
GISS-EH	Y	Y	Y	Y	Y	Y	Y	0	Y	Y	Y	0	Y	Y	Y	Y	Y	Y
GISS-ER	Y	Y	Y	Y	Y	Y	Y	0	Y	Y	Y	0	Y	Y	Y	Y	Y	Y
INM-CM30	Y	Y	Y	Y	Y	0	Y	0	0	0	0	0	0	0	Y	0	0	Y
IPSL-CM4	Y	Y	Y	0	0	Y	Y	0	0	0	0	Y	0	0	0	0	0	0
MIROC3.2(hires)	Y	V	Y	Y	Y	Y	Y	0	Y	Y	0	Y	Y	Y	Y	Y	Y	Y
MIROC3.2(medres)	Y	Y	Y	Y	Y	Y	Y	0	Y	Y	0	Y	Y	Y	Y	Y	Y	Y
MRI-CGCM2.32	Y	Y	Y	Y	Y	Y	Y	0	0	0	0	0	0	0	Y	0	0	Y
PCM	Y	Y	Y	Y	Y	Y	Y	0	0	0	0	0	0	0	Y	0	0	Y
UKM0-HadCM3	Y	Y	Y	Y	Y	Y	Y	0	G	0	0	Y	0	0	Y	0	0	Y
UKM0-HadGEM1	Y	Y	Y	Y	Y	Y	Y	0	Y	Y	0	Y	Y	0	Y	Y	Y	Y

Modified from IPCC AR4 Table 10.1 in (Solomon et al., 2007, p. 756). This version ignores nuances in the original which distinguished different data sets or modeling approaches that were used to account for different forcing agents.

4.3. *Regional case studies*

Though we have focused so far on the IPCC and on *global* climate change, regionally focused studies drawing on the IPCC models are becoming common. The recent paper on aridity and drought in the American Southwest (Seager et al., 2007) mentioned in Section 2, for instance, received a great deal of attention in the popular media, especially in the American Southwest. That paper uses agreement among models as strong evidence for an impending transition to a more arid regional climate. This "broad consensus," the authors note, has implications for water and development policy. Indeed it does, but consensus and unanimity can be achieved in more than one way. What agreement means in this case is particularly important in the light of a similar study by (Christensen and Lettenmaier, 2006) that points to a slight decrease in overall precipitation (on the order of 1%), but also projects substantial precipitation increase during the winter months. The relationship of precipitation to climate change in the Southwest remains poorly understood, and which set of results is most convincing *should* depend in part on the case that can be made for the independence of models from one another in the Seager et al. study.

Regional studies also pose a model-selection problem. Often, studies of particular regions or processes cannot make use of the full set of AR4 model output due to resource constraints or, more commonly, because some models simply do not represent the system under examination in an appropriate manner. In these cases, researchers often proceed with a selection of models that best fit the study at hand.

Such is the case in a study reported by Kripalani et al. (2007), which investigates the effects of climate change on the South Asian summer monsoon. Beginning with twenty-two coupled GCMs, the authors argue that "confidence in climate model regional precipitation projections will depend on how well the models are able to simulate the 20th century monsoon rainfall." Thus, through a selection process using statistical procedures to identify the models best suited to their study, they arrive at a group of six models (in Tables 1 and 2: BCCR-BCM2.0; CGCM3.1(T47); CNRM-CM3; ECHAM5/MPI-OM; MIROC3.2(hires); and UKMO-HadCM3), all of which project an increase in mean summer monsoon precipitation as a result of long-term climate change.

Our question here is the same as before: what significance does agreement among the six selected models have? Did the process of narrowing down to a subgroup of model results eliminate or preserve the independence necessary to add robustness to the study's findings? We cannot answer this question comprehensively. A preliminary comparison based on the data in Tables 1 and 2 does not reveal any obvious interdependencies, but there are many other ways to investigate this set of models (Gleckler et al., 2008). For example, error profiles resulting from various probabilistic and deterministic skill measures, as described in Hagedorn et al. (2005), would be another approach. In any case, the authors would strengthen their case for increased confidence with a clear account of what these six models share that would make them useful for this study, and an argument for independence despite this overlap.

In another example, Wang and Overland (2009) investigate Arctic sea ice extent using a subset the IPCC AR4 models. Their selection among the models relies on an "observational constraint," which eliminates any model lying outside a certain range in its ability to simulate past sea ice extent. This strategy yields six models (as did the monsoon study mentioned above), which are then used to predict future sea ice extent. This approach is justified because: a.) confidence in IPCC projections is related to the ability of models to reproduce past observations, and b.) eliminating outlier models will reduce uncertainty in future projections.[9] The authors do not offer a causal explanation of why these models can be expected to perform better than the others with respect to sea ice extent,[10] but do point to similarity in their results as a source of confidence in the predictions.

Only one model is shared between the two studies described above (Kripalani et al., and Wang and Overland), each of which settled on six models that best reproduce the processes under investigation. This alone should raise concerns about such studies: given that these processes play out in the same open, highly complex, and interconnected system, what is the rationale for using one set of models to predict future sea ice extent, and an almost entirely different set to predict monsoon behavior? It may be reasonable to assume that there is something shared by each subset of models which makes them suitable to the task. How, then, can the authors treat them as independent?

We readily acknowledge that different models have strengths and weaknesses when it comes to modeling particular processes in particular regions. The question here relates to the use of *groups* of models, and the comparison of their results. The results of each of these studies may not be wrong, but increased confidence requires an account of independence within each set of models, and an explanation of model selection that takes into account the underlying causal structure of the models.

The central point of the preceding analysis and discussion is that there are compelling reasons to be concerned about the independence of models that share histories, computer code, causal assumptions, and other essential features if we are to use agreement across models as a reason to believe that our models are correct or approximately correct. Text Box 1 presents a list of the various dimensions we have identified, along which GCMs might be examined in an assessment of independence. This is a preliminary list, no doubt subject to further additions and modifications. If we are to understand the relationships between models, and between models and reality, a great deal more work will need to be done. Our task so far has been to raise a simple but crucially important question that should accompany studies such as those highlighted above: how independent are these models from one other?

In tables and figures throughout the literature analyzing GCM output, one regularly sees distinct, apparently independent statistics or graphical elements for each model, when

[9] Though it does reduce the spread among model results, it is far from clear how such an approach would reduce overall uncertainty.

[10] They note only that three of the six contain a "sophisticated sea ice physics and dynamics package."

Box 1. Sources of independence and interdependence among models.

Basic physics: The underlying assumptions about how the climate works. Many models are similar with respect to fundamental issues such as principles of conservation of mass and energy but there can be different assumptions about how the climate works

Parameters: Assumptions about how aspects of the system operate, such as the effect of cloud cover on global warming. Parameters are specific variables which can be represented with code, as opposed to equations which can reflect more general laws.

Scope and idealizations: The way that models resolve the system can be different. Models may use different resolutions, have different system boundaries, or make different idealizations. GCMs differ in their resolutions while making roughly similar boundary assumptions. GCMs use different idealizations (such as whether the model assumes that ground soil is 3.5 m or 5.5 m) which may be literally false but predictively useful.

Data: Observations (e.g. from satellites or ground-based sensors) used for validating or tuning models, or model subcomponents. Using data from different time periods or physical sources could make model analyses independent from one another.

Numerical coding: Software and computer processes that drive models. If equations are resolved through different coding schemes, this can be a source of independence.

History and sociology: Origins and evolution of the institutions, personnel, cultural norms and technical approaches to modeling. If modeling programs were created based upon another program's source code and approach, this may be a source of interdependence. Professional exchange among modelers can encourage shared strategies; at the same time, academic norms of openness and transparency can foster the criticism of ideas and proliferation of different approaches.

some may be, for all intents and purposes, either the same or at least highly interdependent. For agreement to lead to robustness, it must be at least possible for the models to disagree.

5. Conclusions: toward an interdisciplinary understanding of models and their meanings

Understanding agreement among multiple models is one area in which conceptual clarity is crucial, and more work is needed. We have argued here that there are reasons to be concerned about multi-model analyses that claim robustness because there is presently no consensus on how to evaluate such claims. We have not offered a metric of our own because we recognize that we are pointing to a hard problem. Indeed, following Gleckler's et al. (2008) survey of model *performance* measures, there probably is no single metric of independence that will prove appropriate in all cases. We hope these arguments encourage those who make robustness claims to pay attention to the relationships between the models in their analyses and to address the ways in which the models they study are independent of one another. Our argument is not that the IPCC's models are incapable of supporting robustness claims or that regional studies that rely on multi-model ensembles do not withstand scrutiny. We are arguing that the situation is worse than this, because there is presently no way to judge the quality of analyses based on multiple models.

A recent "Synthesis and Assessment Product" by the US Climate Change Science Program makes a distinction between the quality of climate models and our confidence in their ability to make accurate projections with basis in reality (CCSP, 2008). The quality-accuracy distinction goes to the heart of the robustness issue. If we merely strive to develop a suite of tools that can mimic observational data, then we can say that climate models have made great progress (though considerable room for improvement remains). The central problem addressed by Levins' concept of robustness, however, is that of relating abstract models to the world—an issue that, for climate models, "remains a subject of active research" (CCSP, 2008, p. 4). It seems, however, that climate science has not yet "actively" pursued this problem.

Governments continue to invest billions of dollars in GCMs, seeking to capitalize on the promise of better predictions with lower uncertainty, and increased relevance to decision makers. But GCMs continue to have limited applicability at local and regional scales. Many within the climate science community see such limitations as reason for further investment in the development of GCMs. We argue that in the likely event that such investments continue, considerable attention should be devoted to model independence.

In pointing out the problem of model independence, we are also pointing out one of many broader science policy problems facing climate science. A recent report in *Nature* suggests that the IPCC Fifth Assessment will show a wider spread of model projections than in AR4 (Hefferman, 2010). Is this an appropriate direction for climate science, given the overriding goal of informing decisions? Will the Fifth Assessment express a better grasp of model independence, and thus the meaning of agreement among models? The answer depends on research choices made by scientists and science policy makers.

Understanding model independence is likely to be an interdisciplinary venture. Scientists like Richard Levins have written on the topic, as have philosophers of science, though less directly (Cartwright, 1999; Giere, 1990; Norton and Suppe, 2001; Wimsatt, 2007). Scientists and philosophers have also written on the topic together (Oreskes et al., 1994; Orzack and Sober, 1993). Historians and social scientists have considered the history of climate models and what models mean in the real world (Edwards, 2000, 2001; Lahsen, 2005; Oreskes, 2000; Oreskes and Belitz, 2001; Oreskes et al., 1994; Parker, 2006; Shackley and Wynne, 1996; van der Sluijs et al., 1998). There are also broader questions about the role of models in shaping disciplines, and in influencing the outcomes of science (Robert, 2008; Shackley, 2000; Shackley et al., 1998). There is much more to do, and these scholars will have to work closely with modelers to understand the details of models and why they matter.

As with the general problem of uncertainty in climate science, there will be no silver bullet approach to the issue of independence and robustness in climate model results. In both cases, however, we think that science and society stand to benefit from a consistent and concerted effort to understand and communicate transparently, especially as researchers become bolder in their attempts to make specific regional predictions, and more insistent in their demands for the substantial resources needed for this kind of work (e.g. Shukla et al., 2009). Discourse is essential for maintaining transparency and avoiding the pitfalls of overconfidence and overreliance on models.

Acknowledgments

A spirited exchange with several anonymous reviewers has considerably improved our presentation of these ideas. In addition, we thank our colleagues Roger Pielke Jr, Daniel Sarewitz, Jason Robert, and Clark Miller for helping us to think through the robustness problem. Melissa Baker contributed considerable time and energy in gathering data for our literature review.

REFERENCES

Abramowitz, G., 2010. Model independence in multi-model ensemble prediction. Australian Meteorological and Oceanographic Journal 59, 3–6.

Abramowitz, G., Gupta, H., 2008. Toward a model space and model independence metric. Geophysical Research Letters 35 (5), 4.

Ackerley, D., Highwood, E.J., Frame, D.J., 2009. Quantifying the effects of perturbing the physics of an interactive sulfur scheme using an ensemble of GCMS on the climateprediction.net platform. Journal of Geophysical Research-Atmospheres 114 (D1), D01203.

Bjerknes, V., 1921. On the Dynamics of the Circular Vortex. Geofysiske Publikasioner 2 (4), 1–88.

Cartwright, N., 1999. The Dappled World: A Study of the Boundaries of Science. Cambridge University Press, New York.

Bader, D.C., Gutowski Jr., W.J., Held, I.M., Kunkel, K.E., Miller, R.L., Tokmakian, R.T., Zhang, M.H., 2008. CCSP ‘‘Climate Models: An Assessment of Strengths and Limitations. A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research. Department of Energy, Office of Biological and Environmental Research, Washington, DC, p. 124.

Christensen, N., Lettenmaier, D.P., 2006. A multimodel ensemble approach to assessment of climate change impacts on the hydrology and water resources of the Colorado river basin. Hydrology and Earth System Sciences 3, 3727–3770.

Collins, M., 2007. Ensembles and probabilities: a new era in the prediction of climate change. Philosophical Transactions of the Royal Society A: Mathematical: Physical and Engineering Sciences 365 (1857), 1957–1970.

Dessai, S., Hulme, M., Lempert, R.J., Pielke Jr., R.A., 2009. Do we need better predictions to adapt to a changing climate? Eos 90 (13), 111–112.

Edwards, P.N., 2000. A brief history of atmospheric general circulation modeling. In: Randall, D.A. (Ed.), General Circulation Model Development, Past Present and Future: The Proceedings of a Symposium in Honor of Akio Arakawa. Academic Press, New York, pp. 67–90.

Edwards, P.N., 2001. Representing the global atmosphere: computer models. data, and knowledge about climate change. In: Edwards, P.N., Miller, C. (Eds.), Changing the Atmosphere: Expert Knowledge and Environmental Governance. pp. 31–65.

Giere, R., 1990. Models and theories. In: Explaining Science: A Cognitive Approach, University of Chicago Press, Chicago, pp. 62–91.

Gleckler, P.J., Taylor, K.E., Doutriaux, C., 2008. Performance metrics for climate models. Journal of Geophysical Research 113.

Hagedorn, R., Doblas-Reyes, F.J., Palmer, T.N., 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting. Part I. Basic concept. Tellus 57A, 219–233.

Hefferman, O., 2010. The climate machine. Nature 1014–1016.

Hulme, M., Pielke Jr., R.A., Dessai, S., 2009. Keeping Prediction in Perspective 0911 .

IPCC, 1997. An introduction to simple climate models used Int He Ipcc second assessment report. In: Houghton, J.T., Filho, L.G.M., Griggs, D.J., Maskell, K. (Eds.), Technical Paper II.. Intergovernmental Panel on Climate Change.

Kripalani, R.H., Oh, J.H., Kulkarni, A., Sabade, S.S., Chaudhari, H.S., 2007. South Asian Summer Monsoon Precipitation Variability: Coupled Climate Model Simulations and Projections under IPCC AR4. Theoretical and Applied Climatology 90 (3), 133–159.

Lahsen, M., 2005. Seductive simulations? Uncertainty distribution around climate models. Social Studies of Science 35 (6), 895–922.

Lambert, S.J., Boer, G.J., 2001. Cmip1 evaluation and intercomparison of coupled climate models. Climate Dynamics 17 (2), 83–106.

Levins, R., 1966. The strategy of model building in population biology. American Scientist 54 (4), 421–431.

Norton, S.D., Suppe, F., 2001. Why atmospheric modeling is good science. In: Edwards, P.N., Miller, C. (Eds.), Changing the Atmosphere: Expert Knowledge and Environmental Governance. p. 67.

Oreskes, N., 2000. Why predict? Historical perspectives on prediction in earth science. In: Sarewitz, D., Pielke, Jr., R.A., Byerly, Jr., R. (Eds.), Prediction: Science, Decision Making, and the Future of Nature. Island Press, Washington, DC, p. 23.

Oreskes, N., Belitz, K., 2001. Philosophical issues in model assessment. Model Validation: Perspectives in Hydrological Science 23.

Oreskes, N., Shrader-Frechette, K., Belitz, K., 1994. Verification, validation, and confirmation of numerical models in the earth sciences. Science 263 (5147), 641.

Orzack, S.H., Sober, E., 1993. A critical assessment of Levins's the strategy of model building in population biology (1966). Quarterly Review of Biology 533–546.

Palmer, T.N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M., Délécluse, P., Déqué, M., Díez, E., Doblas-Reyes, F.J., Feddersen, H., 2004. Development of a European multimodel ensemble system for seasonal-to-interannual prediction (Demeter). Bulletin of the American Meteorological Society 85 (6), 853–872.

Parker, W.S., 2006. Understanding pluralism in climate modeling. Foundations of Science 11 (4), 349–368.

Raisanen, J., 2007. How reliable are climate models? Tellus A 59 (1), 2–29.

Robert, J.S., 2008. The comparative biology of human nature. Philosophical Psychology 23 (3), 425–436.

Sarewitz, D., Pielke, Jr., R.A., Byerly, Jr., R. (Eds.), 2000. Prediction: Science, Decision Making, and the Future of Nature. Island Press, Washington, DC.

Seager, R., Ting, M., Held, I., Kushnir, Y., Lu, J., Vecchi, G., Huang, H.-P., Harnik, N., Leetmaa, A., Lau, N.-C., Li, C., Velez, J., Naik, N., 2007. Model projections of an imminent transition to a more arid climate in Southwestern North America. Science 316 (5828), 1181–1184.

Shackley, S., 2000. Epistemic lifestyles in climate change modelling. In: Miller, C., Edwards, P.N. (Eds.), Changing the Atmosphere: Expert Knowledge and Environmental Governance. MIT Press, Cambridge, MA, pp. 107–133.

Shackley, S., Wynne, B., 1996. Representing uncertainty in global climate change science and policy: boundary-ordering devices and authority. Science, Technology, and Human Values 21 (3), 275.

Shackley, S., Young, P., Parkinson, S., Wynne, B., 1998. Uncertainty, complexity and concepts of good science in climate change modelling: are gcms the best tools? Climatic Change 38, 159.

Shukla, J., Hagedorn, R., Hoskins, B., Kinter, J., Marotzke, J., Miller, M., Palmer, T.N., Slingo, J., 2009. Strategies: revolution in climate prediction is both necessary and possible: a declaration at the world modelling summit for climate prediction. Bulletin of the American Meteorological Society 90 (2), 175–178.

Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M., Miller, H.L., 2007. IPCC, 2007, Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK.

Stainforth, D.A., Aina, T., Christensen, C., Collins, M., Faull, N., Frame, D.J., Kettleborough, J.A., Knight, S., Martin, A., Murphy, J.M., 2005. Uncertainty in predictions of the climate response to rising levels of greenhouse gases. Nature 433 (7024), 403–406.

Tebaldi, C., Knutti, R., 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.

van der Sluijs, J., van Eijndhoven, J., Shackley, S., Wynne, B., 1998. Anchoring devices in science for policy: the case of consensus around climate sensitivity. Social Studies of Science 28 (2), 291.

Wang, M., Overland, J.E., 2009. A sea ice free summer arctic within 30 years? Geophysical Research Letters 36.

Wang, X.L., Swail, V.R., 2006. Climate change signal and uncertainty in projections of ocean wave heights. Climate Dynamics 26 (2), 109–126.

Wimsatt, W.C., 2007. Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality. Harvard University Press, Cambridge, MA.

Zachary Pirtle recently completed a Masters in Civil and Environmental Engineering at Arizona State University. He has BS in mechanical engineering and a BA in philosophy from ASU. Zach worked on this project as a research assistant for the Center for Earth Systems Engineering and Management. In the 2008-9 academic year, he was on a Fulbright Scholarship in Mexico. He also recently completed the Mirzayan Science Policy Fellowship at the National Academy of Engineer, and worked as a consultant for CSPO-DC.

Ryan Meyer is a Ph.D. candidate at the School of Life Sciences and Consortium for Science, Policy and Outcomes at Arizona State University. His work focuses on the connections between advancing knowledge and social problems such as climate change. His dissertation research focuses on institutional, political, and epistemological aspects of climate science, and their impacts on social outcomes.

Andrew Hamilton is Assistant Professor in the School of Life Sciences at Arizona State University. Dr. Hamilton's research focuses on the conceptual and theoretical foundations of the biological sciences, particularly evolutionary theory and systematics, as well as on the relationships between science and public policy. Two goals of this work are to use the tools of philosophy to clarify ideas and arguments with the hope of making progress in answering empirical questions and to bring careful thinking about science to discussions of values and policy in the classroom and in public forums. His work is supported by the National Science Foundation (SES-09083935 and SES-0824475).