Holism and Entrenchment in Climate Model Validation

J. Lenhard and E. Winsberg, draft March 7, 2008

Recent work in the domain of the validation of complex computational models reveals that modelers of complex systems, particularly modelers of the earth's climate, face a deeply entrenched form of confirmation holism. Confirmation holism, as it is traditionally understood, is the thesis that a single hypothesis cannot be tested in isolation, but that such tests always depend on other theories or hypothesis. It is always this collection of theories and hypotheses *as a whole*, says the thesis, that confront the tribunal of experience. But in contrast to the way the problem of confirmation holism is typically understood in the philosophy of science, the problems faced by climate scientists are not merely logical problems, and nor are they confined to the role of anything that can suitably be called auxiliary hypotheses. Rather, they are deep and entrenched problems that confront the scientist who works with models whose component parts interact in such a complex manner, and have such a complex history, that the scientist is unable to evaluate the worth of the parts in isolation.

In what follows, we want to argue for two central claims about complex computational models—with a particular emphasis on models of the earth's climate. The first claim is about *holism*. We will argue that recent efforts in the sphere of climate model *inter-comparison* reveal that modern, state-of-art climate models are what we call "analytically impenetrable." We will have to spell out this notion with more care in the sequel, but the intuitive idea is that, as a practical matter, it has become impossible for climate scientists to *attribute*[1] the various sources of relative successes and failures to particular modeling assumptions.

The second claim is about *entrenchment*. In particular, we argue that entrenchment can be identified as one of the principal causes of holism. Here, we want to argue that climate models are, in interesting ways, products of their specific histories. Climate models are developed and

[1] The word "attribution" also occurs in the prominent phrase "attribution of climate change" which stands for the question whether observed climatic change is caused by humans. We do not use the word in this way in this paper.

adapted to specific sets of circumstances, and under specific sets of constraints, and their histories leave indelible and sometimes inscrutable imprints on these models.

The validation of complex computation models is the central issue of the epistemology of computer simulation. The computer science literature often distinguishes between verification and validation as two aspects of the evaluation of simulation. We will speak somewhat more coarsely and treat validation and evaluation as the same. How do we know when a complex computer model is good enough, or reliable enough, for a task for which we hope to depend on it? The issue of the validation of simulations is a particularly interesting one for the epistemology of science, because issues of validation take center stage in simulation in a way in which they rarely do in other modalities in the sciences. It brings to light features of the epistemology that might be absent, but more likely simply hidden, in other modeling and theoretical practices.

To a first approximation, we can think of the validation of a model in the following way: a model is validated when we are convinced that there is an appropriate fit[2] between the dynamics of the model, on the one hand, and the dynamics of the real world system to be modeled, on the other. To be sure, such a conception of the validation of simulation models is somewhat simplified. In particular, simulations are often used to generate predictions about phenomena in domains where data are sparse. Hence, while appropriate fit is of course what we want in a model, we want more than fit with those features of the real world system that are immediately observationally accessible to use. That a model is valid, therefore, is rarely established solely by comparing it to the world. As we have argued elsewhere (Winsberg, 1999, 2001, Lenhard, forthcoming), the sanctioning of simulation models depends on a number of features in addition to fidelity of the simulation's output to known real-world data. It also depends on fidelity to theory, to accepted computation method, and a host of other factors. In this paper, however, we want to set these complications aside, and focus, in particular, on the role of comparison with data in the validation of simulations. We also want to focus, in this paper, on a particular facet of validation. We want, in particular, to think about situations in which models fail to be

[2] "Appropriate" in the sense that, for the intended purpose of the model, the model is close enough to the world in the intended respects and to the intended degree of accuracy.

adequately validated—at situations, in other words, where the behavior of the model is known *not* to be close enough to the behavior of the world for its intended purpose.

This, after all, is the state of affairs known to obtain with regard to most global climate models. There exist several of such models run by research centers worldwide. Each has its specific strengths and weaknesses in certain respects. The series of assessment reports of the Intergovernmental Panel on climate change (IPCC) documents how adequacy of the overall picture is thought to be produced by a synopsis of a plurality of models. In such cases, the issue of model validation is, in effect, the issue of *model improvement*. To put the central question succinctly: when a complex models fails to be adequate, is it possible to identify the various components of the model that contribute to its relative successes and failures?

It is precisely in these contexts, however, in which a serious form of confirmational holism rears its ugly head. On the common understanding of this thesis, a result of the so-called Quine-Duhem problem, it is thought to have two features. First, the problem of confirmational holism is typically associated with the idea of *auxiliary hypotheses* having to do with *observation*. Suppose, for example, that we have the hypothesis that all metal rods expand when heated. An alleged falsification of this hypothesis comes from the observation of rod being heated and not expanding. Confirmational holism comes from the realization that such an observation's credibility depends on a sound understanding, grounded in certain theories or hypotheses, of thermometers and measuring instruments. Any seeming conflict between our original hypothesis and our data could either be the fault of the original hypothesis, or it could be the fault of these auxiliary hypotheses—hypotheses associated with measuring instruments. Second, the problem of confirmational holism is often thought to be a *logical problem*. In other words, on a common understanding of the Quine-Duhem problem, and of confirmational holism, what we are supposed to conclude is that *logic alone* never dictates whether a single hypothesis or theory is confirmed or falsified by a collection of data. But it is usually supposed that good judgment (what Duhem called "bon sense") can decide between such rival possibilities. This is often supposed on the basis of the belief that auxiliary hypotheses that are used in observation can be independently tested. It is usually supposed, in other words, that the Quine-Duhem problem is a philosophical problem without actual practical implications for the working scientist.

But unlike in the conventional picture of how Quine-Duhem is supposed to operate, the holism that arises in climate modelling is wholly independent of whatever hypotheses or theories sanction the reliability of the observational base upon which validation occurs. Even in situations in which the reliability of the data against which simulation output is being compared are not in doubt—that is, even if we imagine a situation where, for example, the data concerning historical record of ice-ages against which the simulation's output will be compared are not open to question—where there is no concern about the reliability of the auxiliary hypothesis used to generate these data— there is still a serious problem of confirmational holism.

Climate simulation

Suppose, for example, that we have a computer simulation of the climate whose simulated dynamics can be compared to its real world counterpart—our planet's climate— in at least important respects. An iconic example of this kind of comparison is the purported fit between the history of the global mean temperature and the output of various global climate models, applied to the past. Everyone has seen this image, cf. image 1.

[locate image 1 here]

Of course, the "real world" side of the comparison—the historical mean temperature—is itself a re-construction out of a vast array of different sources. It is, to use a bit of technical terminology coined by Woodward and Bogen, a "*phenomenon*"—a highly massaged and negotiated description of the behavior of the world that is *inferred* from a variety of sources (Bogen and Woodward, 1988). A corresponding vast array of theoretical and instrumental resources stand behind the line on the graph that is labeled "real world climate." And of course, whether agreement or disagreement between model and world count as evidence for or against the model depends entirely on the credibility of the data conferred by those resources. As we noted earlier, the Quine-Duhem problem, and the problem of confirmation holism, is typically thought to be about these very theoretical resources that stand behind the inferences to these "phenomena." But we shall not be concerned with those issues here—we are more concerned with issues related to the relationships between the models themselves, on the one hand, and the fully reconstructed "phenomena," whatever they turn out to be, on the other.

When it comes to climate models, one cannot overemphasize the degree to which the credibility and assumed reliability of the models comes precisely from the good fit between the output of these models and this reconstructed historical record. Image 2 displays a case of such output as reported in IPCC's Third Assessment Report (2001).

[locate image 2 here]

A variety of political, economic, and policy scenarios is part of this complex picture (d). Graphics (a)-(c) display a variety of scenarios that determine the boundary conditions of the simulations. Climate scientists, themselves, of course, are not in the business of making political and economic forecasts. What they do, instead, is to make a variety of simulation-based predictions of global temperature using a variety of particular assumptions about future greenhouse gas emissions. This accounts for part of the uncertainty in the predictions. There is still, however, another kind of uncertainty lurking in the background of part (d). Even for one particular scenario—one set of assumptions about economics, politics and policy (in short, about emissions)—a range of possible values is reported, not a single predicted temperature. This range stems from a plurality of individual models that can be run given one particular political/economic scenario. Each of these models gives a different forecast and the collection of forecasts gives the range reported here. The entire process is much like that of "collecting all the opinions of valuable witnesses". First you canvass political/economic/policy experts about what to expect in terms of emissions, and then you canvass a second panel of expert—the models—about what kind of climate change to expect in response. The first kind of plurality is generally appreciated by science and the public; at the same time it is acknowledged to be irreducible – there are simply no trustworthy oracles to consult about emissions. However, the second kind of plurality is the more important and at the same time more hidden one. And it is precisely this aspect of the uncertainty that climate scientists would like, in principle, to reduce.

The earth's climate can be thought of as system consisting of a variety of subsystems: the atmosphere, oceans, ice cover, etc. The overall climate dynamics is brought about by the interaction of all these subsystems. Climate models, in turn, are correspondingly modular. There are model modules for the oceans, ice cover, cloud formation, rain dynamics, etc. And so one way to think about climate model improvement is in terms of the contributions to model

output that come from these various modules. However, the approach to improve each module separately has to face serious and even insurmountable problems.

Some words about modularity and climate models. The historical origins of climate analysis are rooted in models of the circulation of the atmosphere – general circulation models (GCMs) that have been developed since the mid 1950ies. The theoretical core of these models is built by the so-called fundamental equations, a system of partial differential equations from the physics of motion and thermodynamics. With the growing interest in climate change in the 1980ies, a process of substantial growth of these models was starting, because more and more facets of the climate system had to be included while aiming at a comprehensive picture. The growth both included the resolution of more sub-processes, like the dynamics of aerosols in the atmosphere, and also the addition of sub-processes in parameterized form. In short, there exists a large variety of paths of growth and the different climate models followed different paths during their development.

Modularity and Pluralism

One aspect of the development of more comprehensive models is of particular importance. A multitude of sub-models had to be included into the atmospheric GCMs that had little to do with the theoretical physical basis of the atmospheric circulation, e.g. ice cover, circulation of the oceans, or land use. The coupling of atmospheric and oceanic circulation models is recognized as one of the milestones of climate modeling because both components had their independent modeling history, including the independent calibration of model performance. Putting them together was a difficult task because the performance of the two sub-models now interfered one with the other. Today, atmospheric GCMs have lost there central place; coupled models entertain a deliberatively modular architecture and comprise a number of highly interactive sub-models (cf. Küppers and Lenhard 2006). The results of these modules are not gathered independently and after that get synthesized, rather data are permanently exchanged between all modules during the runtime of the simulation. Thus the overall dynamics of one global climate model is the complex result of the interaction of the modules—not the interaction of the results of the modules.

Against this background of the modularity of climate models we want to describe the problem of validation, i.e. the question of how well a model simulates the actual climate dynamics. One particular model can of course be compared with certain aspects of the observed climate history. The most prominent one is the global mean temperature. The model can simulate its course over past times and the output can be compared with the reconstruction of climate (temperature) history. Paul Edwards (2001) has pointed out that this reconstruction cannot be derived from data directly but depends on models in various ways. However, as has been said, we simplify matters, neglect the issue of conformational holism in the traditional sense and assume that this reconstruction is straightforward.

But GCMs can also be checked against more local and recent patterns, such as the intensity of tropical winds, precipitation patterns, etc. Relatively speaking, these comparisons are a straightforward validation strategy that can assess systematic errors of the simulation model and enables stepwise improvement. This strategy is well established in simulation modeling practice.

A central problem arises, however, as the complexity, multi-dimensionality and modularity of the models grow. An achievement with respect to one metric of model comparison, produced by complicating the model with a new feature, say a tropical precipitation adjustment, or by substitution one module with another, may not lead to amelioration with respect to another metric or may even make comparisons on that metric impossible or meaningless. Changing the model in such and such a way may improve prediction of tropical winds, but it may simultaneously degrade prediction of precipitation patterns, or even make it now impossible to compare model output with regard to cloud cover.

There are, furthermore, many possible avenues to pursue for improving model performance. Each modeling group follows their own path. In the end, there is a variety of GCMs on the market: major climate research institutions tend to have one or even several of their 'own' GCMs. And each one has its characteristic successes and failures. Adopting a John-Stuart-Mill-kind of view, this plurality can be seen as a virtue to foster competition and to end up with even better results even if unanimity is not attained.

Wendy Parker, in her recent paper "Understanding Pluralism in Climate Modeling", presents us with an illuminating discussion of model pluralism along these lines. She acknowledges that the up-to-date complex climate models cannot be compared in a

straightforward manner: “they represent physical processes acting in climate system in mutually incompatible ways and produce different simulations of climate.”(350) That means, according to Parker, that modellers have different opinions of how to represent the relevant physical processes. Furthermore, she rightly remarks, there are insufficient data to be able to resolve the plurality using criteria of empirical adequacy. Here, we would simply like to direct attention towards an additional cause of the observed plurality.

On Parker’s analysis, mutually conflicting assumptions lead to what she calls an “ontic competitive pluralism” (362) This account at least implicitly suggests that we can accurately identify the causes of the various differences in models outputs in terms of the differences in the assumptions the model authors make about physical processes. It is precisely this, however, that we want to deny. We think this view oversimplifies matters and we will argue that incompatibility is brought about by the very process of complex computational modelling. Our claim about conformational holism is in effect a scepticism about whether the researchers are really able to identify these cause. And thus, we are suggesting there are simulation-specific reasons, reasons having to do with the ways in which computation models are actually implemented, as opposed to reasons having simply to with basic climate science, for model pluralism... Thus confirmational holism is making the multi-model approach unavoidable and is brought about more by the exigencies of dealing with complex simulation models than by rational, though conflicting, choices of researchers.

Analytical understanding impossible

The complex internal composition and massive modularity of climate models is principal source of the problem. Climate models are made up of a variety of modules and submodels. There is a module for the general circulation of the atmosphere, a module for cloud formation, for the dynamics of sea and land ice, for effects of vegetation and many more. In addition, each of them includes a mixture of principled science and parameterizations. And it is the interaction of these components that brings about the overall observable dynamics in simulation runs.

Putting the modules together, moreover, is no easy task. Typically, the specific form of the model that integrates these submodels is crafted over a long process of piecemeal mutual adjustments of the parameters, changes in parameterization schemes, and algorithmic

implementations of the different components. The course of development of these models is close to organic—it would not be a stretch to liken to their development to an evolutionary process. Like in evolution, function is optimized to the particular circumstances, the particular data sets available for comparison, and particular criteria of evaluation, under which optimization occurs.

We argue that the best way to understand the historical nature of GCM optimization is in terms of a concept introduced by William Wimsatt in his recent book: that of "generative entrenchment". Wimsatt's discussion of this concept arises in the context of understanding how techniques from adaptive design function as "a way of increasing the reliability of structures built with unreliable components" (Wimsatt 2007, 133) According to Wimsatt: "Adaptive design is a layered organization of kludged adaptations acquired sequentially and assembled on the fly…" (2007, 133)

The term "kludge" or "kluge" initially stems from programmers' colloquial language and is an extremely useful one here. Andy Clark stresses the important role played by kluges in complex modular computer modelling in general. A Kluge is "an inelegant, 'botched together' piece of program; something functional but somehow messy and unsatisfying", it is—Clark here quotes Sloman: "a piece of program or machinery which works up to a point but is very complex, unprincipled in its design, ill-understood, hard to prove complete or sound and therefore having unknown limitations, and hard to maintain or extend". (Clark 1987, 278)

Kluges have been incorporated into the body of philosophy of science by scholars like Clark and Wimsatt who are inspired both by computer modelling and evolutionary theory. The important point in our present context is that kluges typically function only in the context of a whole system, i.e. for the performance of an entire GCM simulation, whereas they have no meaning in relation to the submodels and modules considered in isolation, or, perhaps more importantly, in relation to that module's potential employment in some other GCM. "What is a kludge considered as an item designed to fulfill a certain role in a large system, may be no kludge at all when viewed as an item designed to fulfill a somewhat different role in a smaller system."(1987, 279)

Suppose, in other words that I want to improve the predictive accuracy of my GCM by coupling a sub-model of ice cover to my existing model. I may begin with some principled

assumption about the physics of ice formation and melting. But what is typical in climate modeling is that by the end of the day, I will incorporate features into the sub-model, or into the interface of the sub-model and the rest of the GCM, that are "complex, unprincipled in [their] design, ill-understood, hard to prove complete or sound and therefore having unknown limitations". The modules of GCMs, in short, inevitably become "kludged," and the fact that they increase the accuracy of one GCM is no guarantee whatsoever that would work as well or at all in another.

The notion, therefore, of generative entrenchment is particularly useful way of understanding the epistemological situation in which climate models often find themselves. Wimsatt explains it as follows: "A deeply generatively entrenched feature of a structure is one that has many other things depending on it because it has played a role in generating them."(2007, 133)

The multitude of possible parameterization schemes and choices of parameters and their balanced interaction in modular models are classic examples of kludged adaptations that are tied, in a fundamental way, to modeling features that have become generatively entrenched. Such features contribute to the difficulties of gaining what we call *analytic understanding* of complex simulation models—an understanding of which sub-components of a simulation are responsible for its various successes and failures--because during the modeling process, the kernel of code, the choice and adjustment of parameterizations, and the peculiarities of controlling the interaction of modules typically get *adapted to* generatively entrenched features of the particular GCM for which they have been crafted.

The point, in sum, is that comprehensive climate models—from the first atmospheric GCMs up to the coupled versions of Earth System Models—have grown organically over several decades of development. And the growth has been a process of give and take between theoretical motivation and practical exigency. Whether a new module adds to or subtracts from the overall reliability of the model may have more to do with some generatively entrenched features of the model than it does with that module's generic "goodness of fit", considered in isolation. When a vegetation module is added to a GCM and adds to the GCM's reliability, how much of this should we attribute to the general features of the module itself, as it might be abstractly characterized, and how much should be attributed to very locally tailored attributes of

the module—the kludges—that have been used to fit and adapt the module to the generatively entrenched features of the GCM? Features, which, presumably, will not necessarily play a role in competing climate models.

Back to the validation of GCMs: If our claim about holism and entrenchment is correct they should visibly shape the way GCMs are validated. It is possible, of course, to test the performance of these models under a variety of conditions. And different models perform better under certain conditions than others. But if model A performs better at making predictions on condition A, and model B performs better under condition B, then optimistically, one might hope that a hybrid model—one that contained some features of model A and some features of model—would perform well under both set of conditions. But what would such a hybrid model look like?

Ideally, to answer that question, one would like to attribute the success of each of the models A and B to the success of particular ones of their submodels—or components. One might hope to believe, for example, that a GCM that is particularly good at prediction of precipitation is one that has, in some suitably generalizable sense, a particularly good rain module. We call success in such an endeavor, the process of teasing apart the sources of success and failure of a simulation, "analytical understanding" of a global model. We would say that one has such understanding precisely when one is able to indentify the extent to which each of the submodels of a global model is contributing to its various successes and failures.

Unfortunately, analytic understanding is hard or even impossible to achieve. The complexity of interaction between the submodels in GCMs, and the degree to which these submodels are adapted, via kludges, to generatively entrenched features of the GCM, is so severe that it becomes impossible to independently assess the merits or shortcomings of each submodel. One cannot trace back the effects of assumptions because the tracks get covered during the klugeing together of complex interactions.. That complex climate models are sometimes characterized as "balance of approximations" (Lambert and Boer 2001, cited in Parker 2006, 359) is in line with our analysis. The ideal of analytic understanding is profoundly impeded by what appears to be a particularly vicious form of confirmational holism. A closer look at model validation as it is actually done in climate science and especially in the so-called model intercomparison projects will support these conclusions.

Validation of climate models

With the growing prominence of climate issues in the public, there has been a great deal of pressure coming from the policy arena to make the process of model validation more rational, and more open to public scrutiny. In particular, policy makers are keen to get from their climate scientists not only prediction, but predictions that are accompanied by quantitative assessments of margins of error and of uncertainty (QMU). As a result of these pressures, specific model comparison projects have been launched. Because prediction uncertainty has been linked to model plurality, (nothing highlights uncertainty more than a plurality of predictions), the community has had to find ways to deal with validation that take into account the existing plurality of models—and the plurality of predictions that emerge from these models.

A key site where these sorts of activities have taken place is the Lawrence Livermore National Laboratory. There, the "Program for Climate Model Diagnosis and Intercomparison" (PCMDI) has been set up in 1989, with the goal of using model intercomparison as a method of supplementing existing modes of validation.

The official PCMDI website states: "The PCMDI mission is to develop improved methods and tools for the diagnosis and intercomparison of general circulation models (GCMs) that simulate the global climate. The need for innovative analysis of GCM climate simulations is apparent, as increasingly more complex models are developed, while the disagreements among these simulations and relative to climate observations remain significant and poorly understood. The *nature and causes of these disagreements must be accounted for in a systematic fashion* in order to confidently use GCMs for simulation of putative global climate change." (PCMDI 2008, our emphasis)

In other words, the goal of intercomparison is to uncover significant differences between models, and to analyze those difference in such a way as to *understand the sources of those differences*. The hope is that this could lead to model improvement on the basis of such improved understanding. Prima Facie, this expressed hope stands in tension with our claim that entrenchment and holism preclude analytical understanding. So let us view at some examples in a bit more detail. Among the intercomparison projects that have been launched at Livermore are

the Atmospheric Intercomparison Project (AMIP), its follower, the Coupled Model Intercomparison Project (CMIP) and the Aqua-Planet Experiment Project (APE).

AMIP

The AMIP project was launched in 1989, the same year as PCMDI, as a worldwide undertaking under the auspices of the World Climate Research Programme. It "undertook the systematic validation, diagnosis and intercomparison of the performance of atmospheric general circulation models. For this purpose all models were required to simulate the evolution of the climate during the decade 1979-1988, subject to the observed monthly-average temperature and sea ice and a common prescribed atmospheric CO2 concentration and solar constant." (Gates et al. 1998)

The simulations were run with certain prescribed boundary conditions – standard scenarios - to make the performances of different simulation models comparable. The simulation output (whose volume can be measured in terabytes), including the calculation of certain diagnostic measures of performance for all contributing models, were then made available in a standard format by the Livermore Lab. AMIP was quickly accepted as a project of the global climate science community and "virtually the entire international atmospheric modeling community (…) contributed the required standard output …" (Gates et al. 1998) The computational phase ran for several years until the data were completed in 1993. After that, a couple of diagnostic subprojects began use these data for validation purposes. Optimism ran high:

"AMIP offers an unprecedented opportunity for the comprehensive evaluation and validation of current atmospheric models, and is expected to provide valuable information for model improvement." (Gates 1992)

It came out, for instance, that "a large-scale error common to all current atmospheric GCMs is colder than observed air in the lower troposphere in the tropics and in the upper troposphere in higher latitudes." (Gates 1992) However, results of this kind were thought to be only the first preliminary step. Based on the observed differences in model performance, the important thing was to make inferences about the performances of the various sub-components of the models and to attribute the diagnosed strengths and weaknesses of the different models.

This, however, turned out to be much more difficult than initially expected. The process of intercomparison took several years and helped to locate and diagnose differences in performance – that was surely a success (and a huge organisational effort). In discussing the "present status" of AMIP in (1992), Gates noted that "while much important information on the model's individual and collective performance will be provided by these statistics, insight into the models' portrayal of specific physical mechanisms requires a deeper and more revealing diagnosis of the results." The question of *attribution*, however, of which particular mechanisms implemented in the models—for instance particular parameter choices or parameterization schemes--where responsible for performance remained largely unsolved – even in later years.

Nevertheless, attribution remained a core goal of AMIP, and the more optimistic stance remained common that intercomparison was the right way to proceed: "In such endeavors, attempts to attribute differences among the simulations to specific model properties require, as a minimum prerequisite, the accurate and comprehensive documentation of these features." (Phillips 1996, PCMDI report No. 24)

While documentation proceeded, difficulties with *attribution*, and with what we have called *analytic understanding* of the models persisted. In their voluminous 1998 review of AMIP, Gates et al. conceded that there where still errors revealed by the intercomparison. Some had been reduced during the last years, but many remained nearly the same. The goal of using intercomparison to understand the nature of these errors remained a goal, but it was postponed until the next project. They wrote programmatically:

"In order to understand better the nature of these errors and to accelerate the rate of model improvement, an expanded and continuing project (AMIP II) is being undertaken in which analysis and intercomparison will address a wider range of variables and processes, using an improved diagnostic and experimental infrastructure."

To summarize the AMIP project, it had two goals:

- First, comparison: make available a technical platform at the Livermore Lab, based on standardized data of model performance so that models' performance could be compared.

- Second, attribution: conduct an analysis that could attribute differences in performance to differences in the model components and mechanisms.

While the first goal was a success, the second was a failure. Our diagnosis of this failure is that it is best understood as a form of confirmational holism arising from the need modelers face to adapt their efforts, often with kludges, to generatively entrenched features of GCMs.

CMIP

The conclusions we draw from our study of AMIP persist as we shift our focus to its more recent sibling: CMIP, the "Coupled Model Intercomparison Project" (CMIP), another one of PCMDI's intercomparison projects. It followed similar lines as the AMIP, but used the up-to-date flagships of simulation modelling, and used coupled atmosphere-ocean models. Phase III of this project provided the data to be shown in the newly released Fourth Assessment Report of the IPCC (AR4, 2007). The project description stressed the organizational and networking aspect for the climate science community. One of the central original goals–deepened understanding of simulation mechanisms via attribution– disappeared nearly entirely from the proposals. What this seems to indicate is that the climate science community has began to tacitly accept a kind of holism about complex simulations that renders analytic understanding of these models out of reach. We admit that there is no complete proof for this claim. It is of course possible that time and effort had not been sufficient yet to reach the kind of understanding that we are suggesting is practically impossible. But we find this unlikely, hence we hold that the conclusions of CMIP 3 reflect a kind of disillusion on the part of climate scientists with regard to attribution, and, in short, believe that acceptance of a very deep kind of confirmation holism is inevitable.

APE

A third intercomparison project reflects the disillusion and tries to maintain the goal of understanding/attribution by reducing complexity. The "Aqua-Planet Experiment Project" (APE) arose out of the problems the researchers had run into with AMIP (cf. Neale and Hoskins 2000a). The APE proposal tries to solve that problem by radically simplifying the boundary conditions: the whole simulated planet—"aqua-planet"—now is covered by water. "In this way, the model's physical interactions are retained whilst the complexity associated with many surface

inhomogeneities are discarded." (Neale and Hoskins 2000b, 108) It is the basic approach of APE to keep the parameterization schemes and simplify solely the boundary conditions. The updated documentation of APE formulates quite cautious. Again, the authors stress the value of obtaining a benchmark for comparison whereas the more important goal – the understanding of the causes of differences in model performance, in short: attribution – is postponed to a later stage (see APE 2008).

Conclusions

The original goal of these projects had been to diagnose strengths and weaknesses of different climate simulation models on the market. But it was precisely in this context that the concrete problem of confirmation holism emerged. The overall performance of the various models could be compared, but the model comparison projects had hoped to do more. They had hoped to be able to identify which, among the various modules, submodels and parameterization schemes that were being employed by the various complete models, were responsible for the various aspects of the successes and failures of the complete models. But this proved not to be feasible. It was impossible to re-trace differences and to single out the culprit of a particular property in terms of modeling assumptions, module inclusion or exclusion, or algorithm implementations. The complex interaction of simulation modules, including kluged adaptations, during which the climate dynamics evolves, covered the tracks. This is an important reason, so we argued, why observed differences in model behavior between various models could not be successfully attributed to flaws or successes of the various sub models. It is well-known, for example, which GCMs are good in reproducing wind patterns, but it is not possible to locate the cause for this in code. And hence the researchers were not able to improve part of their models by the knowledge gained through comparison with other models.

We can now bring together two of the central claims of this paper: The first claim is that climate modelers confront a particularly intractable form of confirmation holism—their complex and highly modular models of the earth's climate are analytically impenetrable. The second claim is about entrenchment as a putative cause for this holism: the various ways in which particular climate models succeed and fail, the ways in which they exceed and lag their peers in

performing the predictive tasks to which they are put, is a product of their history—of the circumstance under which they were developed.

Another concept from Clark's work is useful here: what he calls the principle of the "historical snowball", an informal principle formulated by geneticist and physician Francois Jacob: "Simpler objects are more dependent on (physical) constraints than on history. As complexity increases, history plays the greater part."(Clark 1987, 280)

Think of Dumbo the Elephant, the Disney elephant character whose ears grow so large that he could fly. We know, of course, that in the real world, elephants will never fly. Even though there are various evolutionary adaptations which enable certain creatures to fly, none of these will ever work for an elephant. That is because there are other features of elephants (in particular: their bulk)—features that evolved in particular evolutionary circumstances in response to particular environmental pressures—which make adaptations like wings (or big, floppy ears) useless. A wing is an adaptation for an insect, but not for an elephant.

We propose to see climate models and the efforts of the various model intercomparison projects in a similar fashion. A particular module which is "adaptive" for one GCM (in the sense that, given the present barrage of benchmarking tests available: it improves performance) may not be adaptive for another GCM—indeed it may degrade performance. And it is the particular histories of the GCMs, the "environmental pressures" these models faced as they were developed (read: what the modelers were trying, in particular, to get the models to achieve, and the particular data sets they were using to benchmark their models as the models were being developed) that explain these differences. The features of those models that became generatively entrenched through those histories are the features that make the elephants unable to fly and the insects unable to knock down trees—no matter how many wings we give the elephant, or how many tusks we give the insect.

Put together, these two conclusions become particularly salient when we think about model pluralism and model uncertainty. Think, again, of the procedure of "collecting the opinions of all the valuable witnesses." There are recent trends in climate science which suggest that the range of predictions made by the available arsenal of climate models corresponds, in some way or another, to a probability measure over those various possible outcomes (cf. IPCC 2007 or, for a more skeptical position regarding the feasibility of this endeavor, Smith 2002). There is some

justification for this: the principal justification is that policy makers desperately need to know these probabilities, and we know of no other way to generate them.

But against a background of these practices, it is very important to remember the history that produced the particular arsenal we happen to have at our disposal, and to reflect on the possible effects this history has on that arsenal, and the epistemic limitations we face in uncovering and understanding those effects.

References

APE (2008), website of the Aqua-Planet Experiment Project, visited Feb 4, 2008. http://www-pcmdi.llnl.gov/projects/amip/ape/index.html

Bogen, James and James Woodward (1988), *Saving the Phenomena*, The Philosophical Review, Vol. 97, N. 3 (July, 1988).

Clark, Andy (1987), *The Kludge in the Machine*, Mind and Language 2(4), 277-300.

Edwards, Paul N. (2001), "Representing the Global Atmosphere: Computer Models, Data, and Knowledge about Climate Change", in Clark Miller and Paul Edwards (eds.), *Changing the Atmosphere: Expert Knowledge and Environmental Governance*. Cambridge, MA: MIT Press, 31-65.

Gates, W. Lawrence (1992), AMIP: The Atmospheric Model Intercomparison Project, PCMDI Report No. 7, (also published in *Bulletin of the American Meteorological Society*, 73, 1962-1970), visited Jan 16, 2008 at http://www-pcmdi.llnl.gov/publications/PCMDIrept7/index.html

Gates, W. Lawrence and Coauthors, 1999: An overview of the results of the Atmospheric Model Intercomparison Project (AMIP I). *Bull. Amer. Meteor. Soc.*, **80**, 29-55.

IPCC (2001), Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change (IPCC), J. T. Houghton, Y. Ding, D.J. Griggs, M. Noguer, P. J. van der Linden and D. Xiaosu (Eds.), Cambridge University Press.

IPCC (2007), Climate change 2007 – The Physical Science Basis, Contribution of Working Group I to the Fourth Assessment Report of IPCC, Cambridge University Press.

Küppers, Günter and Johannes Lenhard (2007), Simulation and a Revolution in Modelling Style: From Hierarchical to Network-like Integration, in: J. Lenhard, G. Küppers und T. Shinn (Eds.): *Simulation: Pragmatic Construction of Reality*, Sociology of the Sciences Yearbook 25, Dordrecht: Springer (2007): pp. 89-106.

Lambert, S. and G. Boer (2001), CMIP1 Evaluation and Intercomparison of Coupled Climate Models. *Climate Dynamics* 17, 83-106.

Lenhard, Johannes (in prep), Structural Underdetermination in Simulation Modeling, manuscript.

Neale, Richard B. and Hoskins Brian J. (2000a): "A Standard Test for AGCMs and Their Physical Parameterizations. I: The proposal." *Atmospheric Science Letters*, 1, 101-107.

Neale, Richard B. and Hoskins Brian J. (2000b): "A Standard Test for AGCMs and Their Physical Parameterizations. II: Results for the Meteorological Office Model". *Atmospheric Science Letters*, 1, 108-114.

Parker, Wendy (2006), *Understanding Pluralism in Climate Modeling*, Foundations of Science 11(4), 349-368.

PCMDI (2008), Statement of the website of PCMDI, visited January 16, 2008, http://www-pcmdi.llnl.gov/about/index.php

PCMDI report No. 24, visited Jan16, 2008 at http://www-pcmdi.llnl.gov/publications/PCMDIrept24/AMIPhtdoc.html

Smith, Leonard A. (2002), *What Might We Learn from Climate Forecasts?*, Proceedings of the National Academy of Sciences USA, 4(99), 2487–2492.

Wimsatt, William (2007), *Re-engineering Philosophy for Limited Beings. Piecewise approximations to reality*. Harvard University Press, Cambridge, MA and London, England

Winsberg, Eric (1999): "Sanctioning Models: The Epistemology of Simulation," Science in Context 12(2), 275-292.

Winsberg, Eric (2001), Simulations, Models, and Theories: Complex Physical Systems and Their Representations, Philosophy of Science 68 (PSA Proceedings), S442-454.

Image 1

Image: from IPCC Third Ass. Report (TAR 2001), the fourth report is not different in principle; we use the TAR here because the images are available, while AR4 (2007) is still in the process of being completely published.

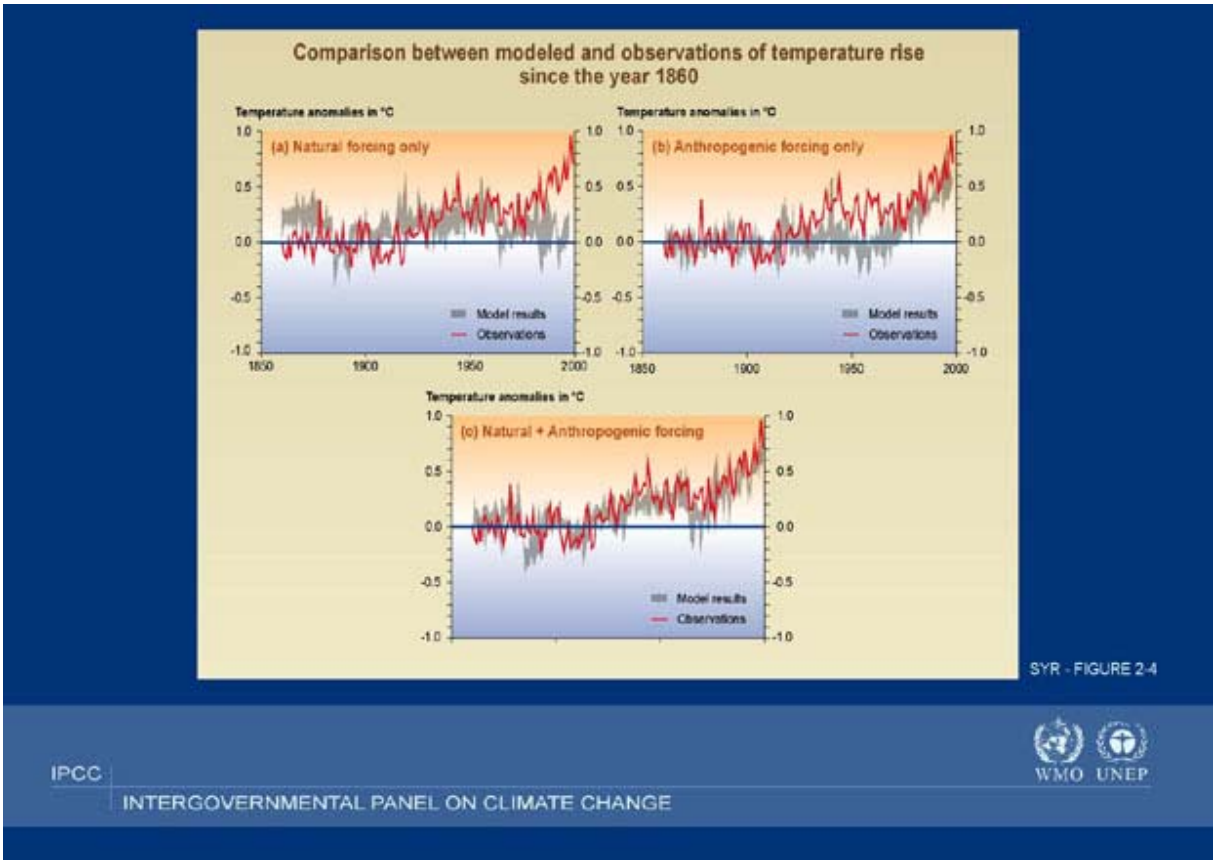
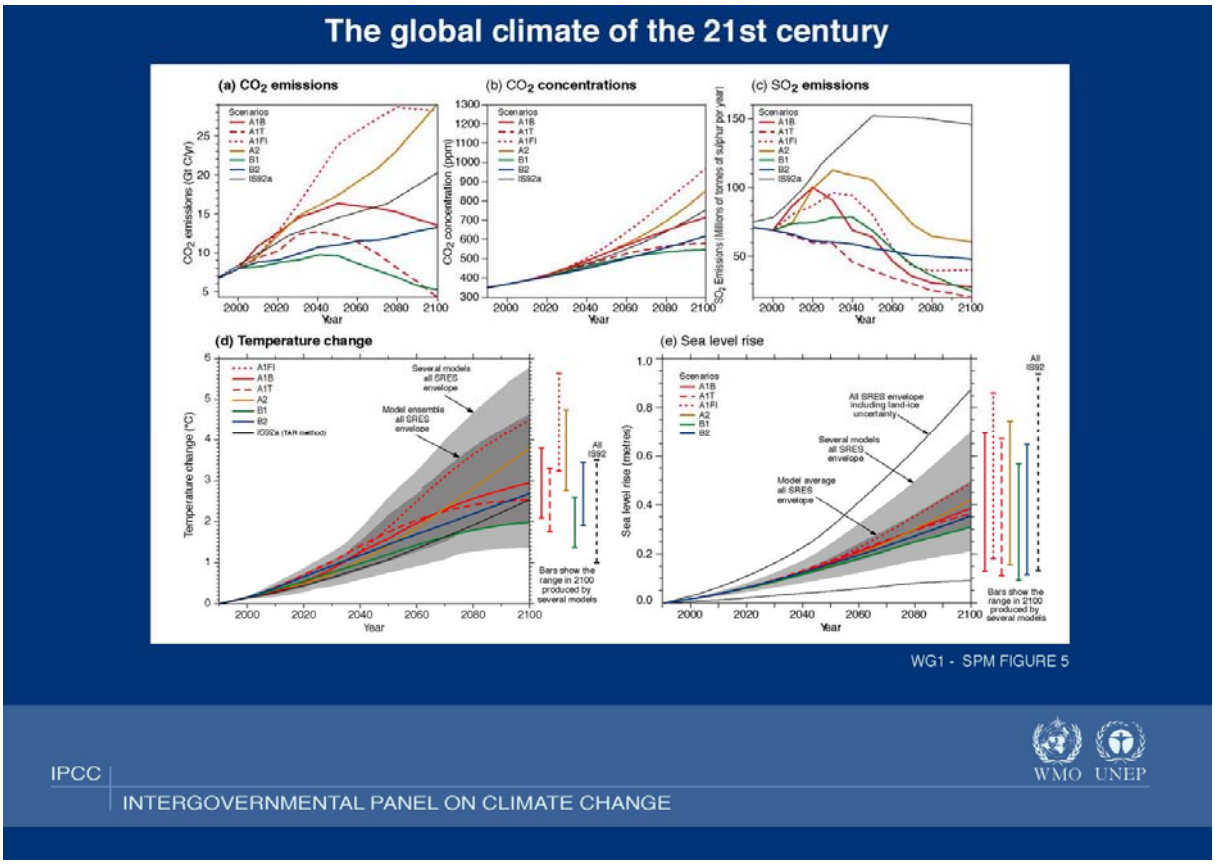


Image 2:


The global climate of the 21st century
(a) CO_2 emissions
(b) CO_2 concentrations
(c) SO_2 emissions
(d) Temperature change
(e) Sea level rise
Scenarios
A1B
A1T
A1FI
A2
B1
B2
IS92a
Year
CO_2 emissions (Gt C/y)
CO_2 concentration (ppm)
SO_2 Emissions (Millions of tonnes of sulphur per year)
Temperature change (°C)
Sea level rise (metres)
Several models all SRES envelope
Model ensemble all SRES envelope
Bars show the range in 2100 produced by several models
All SRES envelope including land-ice uncertainty
Several models all SRES envelope
Model average all SRES envelope
WG1 - SPM FIGURE 5
IPCC
INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE
WMO UNEP