The Coupled Model Intercomparison Project -History, Uses, and Structural Effects on Climate Research

Ludovic Touzé-Peiffer, Anouk Barberousse, Hervé Le Treut

The Coupled Model Intercomparison Project – History, Uses, and Structural Effects on Climate Research

Ludovic Touzé-Peiffer* Anouk Barberousse† Hervé Le Treut*

Article Type:

Advanced Review

Abstract

The results of the sixth phase of the Coupled Model Intercomparison Project (CMIP) are currently being analysed and will form the basis of the IPCC Sixth Assessment Report. Since its creation in the mid-1990s, CMIP has had an increasing influence on climate research. While the principle behind it has always remained the same – comparing different climate models under similar conditions – its design and motivations have evolved significantly over the phases of the project. This evolution is closely linked to the one of the IPCC, since, historically as well as today, the results of CMIP have played a major role in the panel reports. This role increased the visibility of CMIP – over time, more and more people started to be interested in CMIP and to analyze its results. Despite this success, the way CMIP is used today raises methodological issues. In fact, CMIP has promoted a particular way of doing climate research, centred on a single tool, Global Coupled Models (GCMs), and creating a gap between model developers and model users. Due to the debates regarding the interpretation of multi-model ensembles and the validation of GCMs, whether the emphasis on this particular way of studying climate is serving the progress of climate science is questionable.

*LMD/IPSL, Sorbonne Université, ENS, PSL Université, École polytechnique, Institut Polytechnique de Paris, CNRS, Paris, France

†Laboratory Sciences, Normes, Démocratie (SND), CNRS, Sorbonne Université, Paris, France

INTRODUCTION

Intercomparison projects used in climate science are based on a simple idea: run a set of numerical climate models under the same conditions and compare their results. The Atmospheric Model Intercomparison Project (AMIP), which started in 1990, was the first attempt to coordinate such an activity. Different intercomparison projects have followed: the Coupled Model Intercomparison Project (CMIP), in particular, is now considered as "one of the foundational elements of climate science" (Eyring et al., 2016, p. 1937). Since its creation in the mid-1990s, it has evolved over five phases, involving all major climate modeling groups in the world. In addition to their role in climate research, these phases have hold a central place in the reports of the Intergovernmental Panel on Climate Change (IPCC). For the expert panel, whose objective is to synthesize the state of current knowledge about climate and climate change, CMIP results were a goldmine – they provided an easy way to have a summary of what the most comprehensive climate models had to say about these questions. The figures based on their results were also perfect illustrations for the Summary for Decision Makers accompanying each IPCC report – for instance, in the fourth Assessment Report (AR4), among the 7 figures of the Summary for Decision Makers, 4 were based on CMIP results.

If CMIP has been a great opportunity for the IPCC, its role for climate research is more controversial. CMIP has indeed favored the use of one type of model, Global Coupled Models, at the expense of other simpler, less comprehensive models. Climate science is not the only area of environmental science to look at complex systems – but it is probably the only one where the most comprehensive climate models have such a supremacy (Shackley, Young, Parkinson, & Wynne, 1998). Ecologists, biologists or economists all use models of varying complexity (Gabaix & Laibson, 2008; Jørgensen, 2008). The supremacy of GCMs in climate science raises questions *a fortiori* when we look at the numerous issues surrounding the validation of GCMs and the interpretation of their results (e.g. Knutti, Furrer, Tebaldi, Cermak, & Meehl, 2010; Parker, 2011; Shackley et al., 1998). The problem gets worse when looking at the particular use of GCMs CMIP has promoted – by making the outputs of

its simulations freely available, CMIP has created a growing gap between model developers and model users. It has promoted the analyses considering GCMs as black boxes and made it extremely difficult for these analyses to be relevant for the improvement of climate models.

Our paper aims at tackling these different issues by providing an overview of the historical evolution of CMIP, what it has become today, and the effects it has had on climate research. We will start by analyzing how the project and the motivations behind it have evolved over time. We will see that part of this evolution comes from the intertwining between the history of CMIP and that of the IPCC. In a second part, we will study a set of papers based on the results of the fifth phase of CMIP (CMIP5) in order to distinguish three main uses of CMIP results. We will see that each of them raises methodological issues. Finally, borrowing a concept from Pitt (2000), we will define CMIP as an infrastructure for climate science and we will explore the various ways in which CMIP, as an infrastructure, has shaped climate research. Considering these different effects altogether, we will finally question the value of CMIP on climate research as a whole.

History of climate model intercomparison projects

AMIP and the genesis of climate model intercomparison projects

The genesis of climate model intercomparison projects lies in numerical weather forecasting. In the 1970s, intercomparison projects between numerical models used for weather predictions were pushed by the Global Atmospheric Research Program (GARP) - a program established in 1967 with the goal of coordinating weather and climate research internationally. One of the first decisions of the Joint Organizing Committee leading the program was to create a Working Group on Numerical Experimentation (WGNE), in order to promote a dialogue between modeling groups (Gates, 2015). The WGNE aimed in particular to facilitate intercomparison projects between atmospheric models used either for weather forecasting or for climate research.

At first, such intercomparison projects concerned mainly atmospheric models used for operational weather prediction. To compare different models or different versions of the same model under similar conditions, intercomparison protocols were designed and led to the publication of several papers in the 1970s (e.g. Baumhefner & Downey, 1978; D. D. Houghton & Irvine, 1976). Comparatively, fewer intercomparison projects emerged in climate research. One of the first example of juxtapositions of results from different atmospheric climate models was provided in 1975 by the report of the U.S. GARP Committee's Panel on Climatic Variation (National Academy of Sciences, 1975). This report used the simulations available at that time to deduce common diagnostics between models. Even if it did not design any protocol to compare climate models under similar conditions, it inspired further research on the subject (Gates, 1992).

The first climate model intercomparison protocols appeared in the following years with a series of stand-alone initiatives. Among others, we can cite the Intercomparison of radiation codes in climate models (ICRRM) workshop, which led to the publication of a paper comparing atmospheric radiative processes in climate models of varying complexity (Luther et al., 1988). Another example is given by the 1984 paper by Potter and Gates, in which the seasonal response of two atmospheric climate models (Potter & Gates, 1984) is compared.

Until the end of the 80s, these different initiatives remained isolated. It led the Joint Scientific Committee (JSC) of the World Climate Research Program (WCRP) to call for a systematic and comprehensive intercomparison of atmospheric climate models. Since such an intercomparison would be based on a series of expensive simulations, having access to powerful computing facilities was necessary for the emergence of the project. Under the aegis of W. Lawrence Gates, the JSC successfully convinced the U.S. Department of Energy to provide the computer facilities of the Lawrence Livermore National Laboratory (LLNL) to support the project. It led to the creation of the Program for Climate Diagnosis and Intercomparison at LLNL in 1989, with the official goal of "increasing understanding of the differences among climate models" (Gates, 1992, p. 1963). One year later, in 1990, the Atmospheric Model Intercomparison Project (AMIP) - the first major experiment of this

program - was officially endorsed by the JSC.

AMIP was designed to compare the response of atmospheric General Circulation Models (GCMs) - the models simulating the behavior of the atmosphere at a global scale - on seasonal and interannual time scales. All atmospheric GCMs could participate. The decade 1979-88 was chosen as the simulation period, and the protocol imposed to all models specific boundary counditions - more precisely standardized values for the solar constant and atmospheric CO_2 concentrations, as well as observed mean sea surface temperature and sea-ice distributions. Moreover, the outputs of the simulations had to be given in a standard format (Gates, 1992).

There were officially two main motivations behind AMIP. The first one was to "undertake the systematic intercomparison and validation of the performance of atmospheric GCMs on seasonal and interannual time scales under as realistic conditions as possible" (Gates, 1992, p. 1963). In other words, AMIP aimed first at identifying output differences between atmospheric GCMs under the same protocol, and comparing them with observations in order to validate the performance of these models. However, the original ambition of AMIP was also "to support the in-depth diagnosis and interpretation of the model results" (Gates, 1992, p. 1963), that is to say, not only to find out what the differences between atmospheric GCMs were, but also to understand them. This second motivation led to the organisation of 26 diagnostic subprojects to analyze AMIP outputs in the years following the project (Gates et al., 1999, Appendix A).

AMIP was a major step for climate modeling - thanks to it, climate modelers had for the first time access to an institutional structure to compare and evaluate the performance of their model under similar conditions. The experiment became quickly "the most prominent international effort devoted to the diagnosis, validation, and intercomparison of global atmospheric models' ability to simulate climate" (Gates et al., 1999): by 1995, 31 modeling groups had taken part in the experiment, representing almost the entire atmospheric modeling community. This massive endorsement paved the way to subsequent intercomparison

projects, to the Coupled Model Intercomparison Project (CMIP) in particular.

The beginnings of CMIP - CMIP1, 2 and 2+

As early as the late 1960s, it had been recognized that ocean played a key role in climate, and attempts had been made to couple atmospheric and ocean GCMs (Manabe & Bryan, 1969). However, such coupled models were complex, and required a high computational capacity. Thus, it is only in the 1980s that coupled models started to be developed in more and more laboratories to represent the dynamic interactions between the atmosphere, ocean and cryosphere. As the most comprehensive models, they were sometimes seen as "potentially the most useful tools in simulating global climate, studying present-day climate fluctuations and addressing the problem of anthropogenic climate change" (G. A. Meehl, 1995).

Therefore, at the end of the 80s, even if they were still facing strong uncertainties and systematic errors, some global coupled models have been used to study the impact of an increase of anthropogenic CO_2 on climate (J. Houghton, Jenkins, & Ephraums, 1990). For instance, *the Supplementary Report to the IPCC Scientific Assessment* (J. Houghton, Callander, & Varney, 1992) compared the temperature rise associated with a transient CO_2 doubling in four different coupled models. The report recognized the limits associated with this set of simulations, but still used them to confirm its statements about the rise of temperature associated with a CO_2 increase in the atmosphere (J. Houghton et al., 1992).

As a consequence, CMIP was born at the confluence of two influences. On one side, AMIP had shown the potential of intercomparison projects to coordinate and organize research around atmospheric GCMs. Therefore, it was tempting to organize a similar project for coupled GCMs. On the other side, more simulations from coupled models were needed to make statements about anthropogenic climate change more robust.

The original structure of CMIP reflected these two motivations. In fact, CMIP was initially divided into two complementary phases (G. A. Meehl, Boer, Covey, Latif, & Stouffer, 1997):

1. CMIP1, which started in 1996, and transposed the main objectives of AMIP – measuring and understanding the ability of atmospheric GCMs to simulate current climate – to coupled models.

2. CMIP2, which compared climate changes simulated by coupled models under a 1% per year CO_2 increase. Starting in 1997, this second phase was thus directly in line with the comparison performed in *the Supplementary Report to the IPCC Scientific Assessment* (J. Houghton et al., 1992).

Due to limitations in data processing and archiving capabilities at that time, CMIP1 and CMIP2 included only a few output fields, and at a coarse temporal resolution: for example, surface temperature, precipitation, and sea level pressure were averaged over one month. This rough sample was a strong limitation for analyses based on experiments. Consequently, the Working Group on Coupled Modelling (WGCM) – a subgroup of the World Climate Research Programme (WCRP) playing a leading role in CMIP – launched a new phase, CMIP2+, in 1999 to include many more model fields, and daily data if possible. However, this new phase represented significant additional work for the modeling groups, and in the end CMIP2+ was able to collect only 12 complete sets of outputs (G. Meehl, Covey, McAvaney, Latif, & Ronald, 2005).

CMIP1 and 2 – but not CMIP2+, which was not completed in time – played a substantial role in the Third Assessment Report (TAR) of the IPCC. In particular, detailed analyses of CMIP models were presented in the chapters "Model Evaluation" (McAvaney & et al., 2001) and "Projections of Future Climate Change" (Cubasch & et al., 2001). In return, the IPCC Third Assessment Report made some recommendations about future phases of CMIP. In particular, it called for: "GCM simulations with a greater range of forcing scenarios and an increased ensemble size to assess the spread of regional predictions" (Giorgi & et al., 2001, p. 586). Taking into account these recommendations, the WGCM decided to design a new phase of CMIP.

CMIP3 – "A new era in climate change research"

A major novelty of CMIP3, compared to previous phases of CMIP, was to include climate change scenario experiments, that is to say projections of future climate change under different emission scenarios. Such simulations are of great interest for decision makers, because they can be used as a basis for choosing between different mitigation and adaptation strategies.

Scenarios had already been used in the IPCC First Assessment Report. However, the computational capacity available at that time did not make it possible to run these scenarios with GCMs; instead, very simplified models called "box-diffusion models", which were thought to give the same results as GCMs when globally averaged, had been used (J. Houghton et al., 1990).

In its subsequent reports, the IPCC continued to steer climate research around emission scenarios. In particular, in the preparation of the Third Assessment Report (TAR), it produced a Special Report on Emission Scenarios (SRES) (Nakicenovic & Swart, 2000). With its set of 40 scenarios, this report aimed to cover a wide range of assumptions about the main demographic, economic, and technological driving forces of future greenhouse gas and sulfur emissions. The idea was originally that the climate modeling community would use these scenarios in coupled model simulations which could figure in the TAR. However, the SRES was approved only in 2000 - that is to say one year before the publication of the TAR - and most modeling groups could finally run only two scenarios (A2 and B2) (Cubasch & et al., 2001; Giorgi & et al., 2001).

To avoid such coordination and timing issues in subsequent IPCC reports, it was decided that scenario experiments would henceforth be part of CMIP (G. A. Meehl et al., 2007). Thus, CMIP3 included three different climate change scenarios, corresponding respectively to the B1, A1B, and A2 scenarios of the SRES. More generally, CMIP3 was also more focused on climate change than the previous phases of CMIP - among the twelve CMIP3 experiments,

ten tested the dynamic response of climate to various CO_2 concentrations (stable or evolving with time).

A main motivation behind these experiments was to help the Fourth IPCC report (AR4) to provide "a better assessment of the state of human knowledge on climate variability and climate change from the models" (G. A. Meehl et al., 2007, p.1384). As a consequence, CMIP3 was planned early enough in order for the analyses based on model experiments to be used in the Fourth IPCC report (AR4). Providing assessments for the IPCC reports was thus a motivation at the core of CMIP3, whereas it was a side objective for previous phases of CMIP.

In CMIP3, the role of CMIP to organize and coordinate climate research – while present in the first phases of CMIP as well – also acquired a new dimension. Indeed, CMIP3 was the first phase of CMIP to give open access to all the data from its experiments. While in CMIP1, 2, and 2+, only a few modeling groups around the world have had access to the data and analyzed them, the data of CMIP3 were made accessible to any student or researcher around the world. It represented more than 30 terabytes of data. This new openness "brought global coupled climate model intercomparison and analysis to an internationally coordinated level never before achieved in the field of climate science" (WGCM, 2006). As such, it ushered "a new era in climate change research" (G. A. Meehl et al., 2007).

CMIP5 and 6 – the most recent phases of CMIP

The success of CMIP3 put the climate modeling community at the center of contradictory interests. More and more scientists outside the climate research community were interested in using CMIP results for their own areas of expertise. Therefore, CMIP5[1] was designed in order to satisfy not only the motivations of the climate modeling community, but also those of many different users:

[1]CMIP4 was skipped in order to make the numbering of CMIP phases in line with the numbering of IPCC reports.

> the integrated set of CMIP5 simulations attempt to address major priorities of several different communities, and incorporates some of the ideas and suggestions of many individuals and from a number of workshops and meetings. These workshops involved scientists with a wide range of interests, including climate modeling, biogeochemistry modeling, integrated assessment modeling, climate change impacts, climate analysis, climate processes, and climate observations. (Taylor, Stouffer, & Meehl, 2012, p. 486)

In order to take into account the various requests expressed, CMIP5 included more experiments than CMIP3. The analyses of the simulations conducted in these different experiments have formed the basis of the IPCC Fifth Assessment Report (WGCM, 2012, p. 6). However, the deadlines imposed by the redaction of the IPCC report have put a lot of pressure on researchers. Moreover, CMIP5 has been extremely demanding in terms of computing and time resources, and blocked other research in modeling centers (Eyring et al., 2016).

Therefore, when thinking about the design of CMIP6, there has been a common will from the modeling community to decouple CMIP experiments from the IPCC, and to reorganize CMIP towards a few precise scientific questions relevant for climate and climate change. More precisely, it was decided that CMIP6 should be centered around three main scientific questions:

> "How does the Earth system respond to forcing? What are the origins and consequences of systematic model biases? How can we assess future climate change given internal climate variability, climate predictability, and uncertainties in scenarios?" (WGCM, 2014, p. 13)

These questions reflect a will of the research community to take control of CMIP again, and use it to better understand climate processes: "Ultimately scientific progress on the most pressing problems of climate variability and change will be the best measure of the success of CMIP6" (Eyring et al., 2016, p. 1949). In particular, the importance given to scenarios in previous phases of CMIP was questioned. Indeed, although they are central in IPCC reports, scenarios are of little use in understanding climate processes themselves,

as they involve many different assumptions that make it difficult to interpret their results. Therefore, it was decided that the scenarios would not belong to the core experiments of CMIP6 – the experiments that all participating groups should perform – but to the *CMIP-Endorsed Model Intercomparison Projects (MIP)* – secondary experiments on a voluntary basis.

Uses of CMIP results and related controversies

In the previous section, we have seen that even if the principle behind CMIP has always remained the same – comparing climate models by submitting them to a common set of simulations – its design and the motivations behind it have evolved significantly over the phases of the project.

However, the evolution of CMIP design and the motivations behind it is only one part of the story. To fully grasp what CMIP has become, we also need to understand how CMIP results have been used by climate scientists. For that, we looked at peer-reviewed papers based on CMIP outputs: more precisely, we consulted a set of 280 papers based on CMIP5 results published between 2012 and 2018 in six leading climate journals[2]. In this set of publications, based on the title of the papers and their abstracts, we distinguished qualitatively three main uses of CMIP results: the exploration of future climate change and associated uncertainties, the comparison of CMIP simulations with observations and the interpretation of model results. Table 1 gives a summary of how often these different uses come in the various journals from which the papers were taken.

The three uses we distinguished are not exclusive – in particular, we counted 25 papers exploring future climate change that first compare CMIP simulations with observations in order to assess their quality. Nor are these categories exhaustive – among other uses, we noted in particular the study of paleoclimates or present climate (see Further reading). Be-

[2]Climatic Change, Climate Dynamics, Environmental Research Letters, Geophysical Research Letters, Journal of Climate, Journal of Geophysical Research.

Journal	Number of papers in the dataset	Number of papers using CMIP results mainly to: Explore future climate change and associated uncertainties	Compare CMIP simulations with observations	Interpretation of model results
Clim. Dyn.	52	21	19	22
Clim. Change	7	7	2	0
Environ. Res. Lett.	12	7	3	3
Geophys. Res. Lett.	64	28	14	29
J. Climate	92	41	26	30
J. Geophys. Res.	53	21	22	14
Total	280	125	86	98

Table 1: Summary of main uses of CMIP results for peer-reviewed papers ordered by journal. For a detailed list of all peer-reviewed papers analysed, see Further reading.

sides, there are two kinds of papers based on CMIP results we did not considered in our study. First, papers analyzing CMIP results from only one or two models. Secondly, papers not using CMIP results as such, but proposing statistical methods to analyze them or calculate the associated uncertainties.

Our attribution of uses is in part subjective, and one might adopt a different classification. However, we hold that a large portion of the literature based on CMIP results explores future climate change and associated uncertainties, compares CMIP simulations with observations and/or tries to interpret the model results. Yet, we will see in the next part that each of these uses raises methodological questions.

Exploring future climate change and the associated uncertainties

Exploring future climate change is the most widespread use of CMIP results we have identified in our set of papers. In the 125 papers tackling this issue, many different themes were

addressed: among others, we can cite regional impacts of climate change (e.g. Luomaranta et al., 2014; Penalba & Rivera, 2013; Zomer et al., 2014), decadal predictions (e.g. Gaetani & Mohino, 2013; Guemas, García-Serrano, Mariotti, Doblas-Reyes, & Caron, 2015; G. A. Meehl et al., 2014), or the consequences of climate change on one particular climate phenomenon, for instance the Asian monsoon (e.g Jayasankar, Surendran, & Rajendran, 2015; Srivastava & DelSole, 2014; Zou & Zhou, 2015), the ENSO (e.g Stevenson, 2012; Taschetto et al., 2014) or the poleward expansion of Hadley Circulation (e.g Hu, Tao, & Liu, 2013).

However, the interpretation of Multi-Model Ensembles (MME) provided by CMIP is controversial. Let's consider for instance, Figure 1, which gives the spread of the temperature and precipitation in India during the summer monsoon according to a set of CMIP5 models. We see that CMIP5 models simulate a future warming of Indian landmass at the end of 21st century by about $1.19 \pm 0.79\,°C$ for RCP2.6 and $3.99 \pm 1.27\,°C$ for RCP8.5. Concerning the change in precipitation, there is a relatively larger model spread, with a projected change of precipitation of 0.39 ± 0.79 mm/day for RCP2.6 and 0.95 ± 1.13 mm/day for RCP8.5 (Jayasankar et al., 2015). Seemingly, this range provides a measure of the uncertainty about future climate change – in this example, it seems that there is much more uncertainty about the impacts of climate change on precipitation than on temperature.

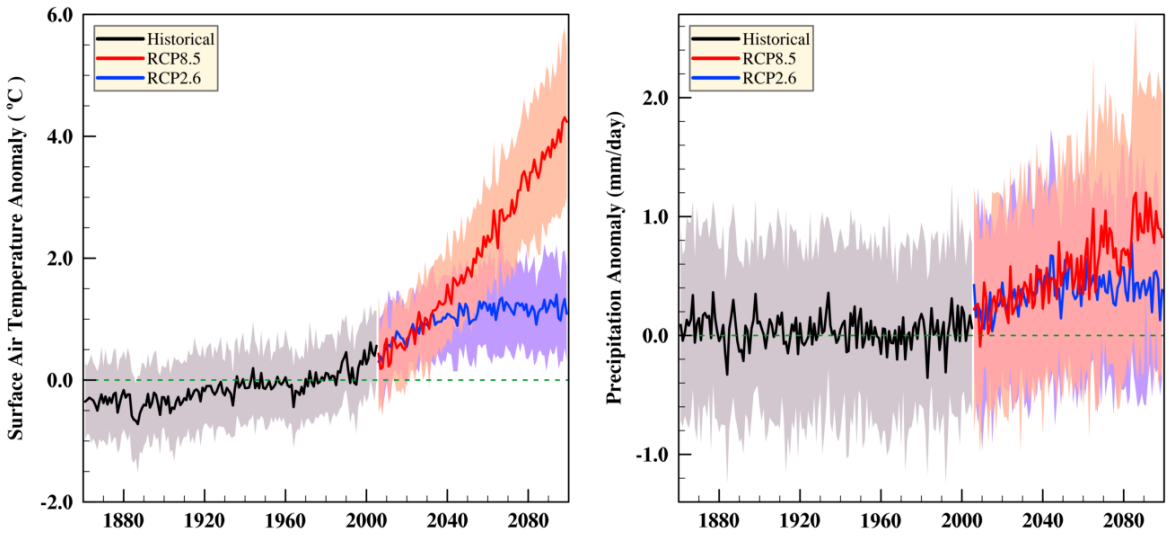

Figure 1: JJAS anomalies with respect to the base period (1961–1990) in CMIP5 historical simulations and 21st century projections of (a) surface air temperature and (b) precipitation for India. The intermodel spread is shown in shades (gray for historical runs, purple for RCP2.6 simulations, and light orange for RCP8.5 simulations). The solid lines represent the multi-model average for historical (black), RCP2.6 (blue), and RCP8.5 (red). Courtesy Jayasankar et al. (2015).

Nevertheless, such a statement raises epistemological issues. First, since the participation in CMIP is made on a voluntary basis, the set of models participating in CMIP has not been designed to span the uncertainty about future climate change. The distribution of CMIP multi-model ensembles is in a large part arbitrary. It led Tebaldi and Knutti (2007) to describe multi-model datasets as "ensembles of opportunity". Since these ensembles of opportunity do not systematically explore the uncertainty about future climate change, some scientists considers them only as a lower bound of this uncertainty (e.g Stainforth, Downing, Washington, Lopez, & New, 2007). How low is this bound, however, remains unknown.

As a corollary, it is difficult to interpret the agreement between models participating in CMIP. As documented by Pirtle, Meyer, and Hemilton (2010), in many scientific papers, confidence in model projections about future climate change is justified by the agreement between GCMs in multi-model ensembles. For instance, in Figure 1, all models agree that

the surface air temperature anomalies in India with respect to the base period (1961-1990) are below $6°C$ under the business as usual scenario (RCP8.5). Yet, if we consider the spread in CMIP projections only as a lower bound of the uncertainty, we would not conclude with confidence that this result stands as well for the real world.

A less ambitious goal would be to use CMIP simulations to have a "best guess" of future climate change. A natural candidate, chosen in many studies, is the average of the projections, as in Jayasankar et al. (2015) (see Figure 1, where the average is represented in solid lines). However, in some situations, the rough average masks the signal entirely. For precipitation for instance, when temperatures rise, many climate models predict large-scale drying in the subtropics and moistening at high latitudes , but at slightly different locations. As Knutti et al. (2010) showed, the problem is that when models are averaged, they tend to cancel each other out. Thus, at some latitudes, all models show significant drying over a relatively large fraction of the land surface, but the average do not.

More generally, the relevance of the gross arithmetic mean for climate projections – what Knutti (2010) called "model democracy" – has often been questioned for two main reasons: 1. Climate models are not independent of each other – on the contrary, as Masson and Knutti (2011) and Knutti, Masson, and Gettelman (2013) have shown, they share pieces of code and common ideas. 2. Climate models do not have the same degree of agreement with observations: some perform better than others. These disparities between models motivated the work of Knutti et al. (2017) who proposed, as an alternative to the gross arithmetic mean, a weighting scheme that takes into account both the large differences in model performance and the interdependencies between models. Whether similar weighting schemes should be used more frequently remains controversial[3].

Taking CMIP results as a lower bound of the uncertainty or using them as a "best guess" are not the only ways to interpret CMIP multi-model ensembles. Both climate scientists and

[3]For a detailed discussion of this issue and other challenges in combining projections from multiple climate models, see Knutti et al. (2010).

philosophers of science (see for instance Parker, 2011; Tebaldi & Knutti, 2007; Winsberg, 2018) have proposed other alternatives, but none of them have reached a consensus yet. Although CMIP multi-model ensembles are often used to explore future climate change and associated uncertainties, how to interpret thus remains an open question.

Comparing CMIP simulations with observations

Another common use of CMIP results consist in comparing them with observations in order to assess the performance of the corresponding GCMs (86 papers in our set). A variety of observations can be used, such as station data, satellite data, proxy data or reanalysis data. However, this measure of performance of CMIP models raises another bunch of analytical problems.

The first one is related to the fact that, in complex climate models, many parameters are poorly constrained by observations and are adjusted in order to satisfy key climate metrics. Hourdin et al. (2017) provide an excellent overview of the tuning of parameters in GCMs and the epistemological problems it poses. As they explain, the agreement with observations can be improved by changing parameters not directly relevant for the problem at stake. For instance, they show that the global top-of-atmosphere energy balance can be adjusted by changing a parameter controlling the fall velocity of ice crystals, a parameter which, at first glance, seems far from the issue at hand. It is thus possible to get the right result for the wrong reasons – the agreement with observations may result from compensating errors and do not necessary prove that processes are well represented in the model. As noted by Tebaldi and Knutti (2007), Frisch (2019), and others, this limitation is particularly strong if the same datasets are used to tune a model and to evaluate its performance.

Another issue concerns observations themselves. In our set of papers, the large majority of studies do not compare CMIP simulations with model-independent observations, but only with reanalysis. Since reanalysis models are based on numerical methods, assumptions, and parameterizations similar to those of real climate models, they all exhibit biases of various kinds (Edwards, 2010). Even if the analysis is continually corrected by available obser-

vational data, reanalysis models thus transmit part of their biases to reanalysis products. Therefore, an agreement between climate models and reanalysis data might just illustrate a common bias in both climate and reanalysis models. This problem gets worse for processes for which we have few observations, such as the hydrological cycle (Tebaldi & Knutti, 2007) – the reanalysis is then loosely constrained by observations, but mainly model-derived.

As a consequence, although the agreement between models and observations is valuable to identify systematic biases in climate models, it is not a guarantee of their reliability. Due to the number of parameters in GCMs and to the interdependencies between GCMs and observations, the agreement between GCMs and observations validates GCMs only in a very weak sense.

Interpreting model results

Originally, the official motivations behind AMIP were formulated as such: "The basic purpose of AMIP is to undertake the systematic intercomparison and validation of the performance of atmospheric GCMs on seasonal and interannual time scales under as realistic conditions as possible, and to support the in-depth diagnosis and interpretation of the model results." (Gates, 1992, p. 1963). The hope was that this "in-depth diagnosis and interpretation of the model results" would help to identify the causes of success and failures of participating climate models, and thus to improve their performance. In particular, Gates had identified the parameterization of convection and precipitation as an outstanding modeling problem and called for further analysis of AMIP results to reduce errors related to it (Gates et al., 1999).

In our set of papers, we counted 98 papers attempting to interpret model results. Yet, none of them look at the details of the parameterizations involved. Most of these studies stay at the level of model outputs; by analyzing correlations between the outputs of each model, they exhibit causal relationships between them. However, where this causality comes from in the details of the corresponding models is not addressed. Kent, Chadwick, and Rowell (2015) provide an example of such a study; their objective is to understand uncertainties in

future projections of seasonal tropical precipitation. They investigate correlations between precipitation, global mean temperature, pattern in sea-surface temperature and a few other variables. The influence of the precipitation schemes used by the different models is left completely out of the picture. Another example is provided by studies who try to assess the effect of one or several feedbacks under CO2 increase (e.g. Long & Collins, 2013; Qu & Hall, 2014). These studies diagnose and interpret climate feedbacks and their effect solely on the basis of simulation results, but never discuss the underlying parameterizations.

According to Lenhard and Winsberg (2010), CMIP analyses stay at the level of model results because climate models exhibit a form of "confirmation holism". This concept is traditionally defended in philosophy of science as the idea that a single hypothesis can never be tested in isolation, but that such tests inevitably depend on other theories or hypotheses. As Pierre Duhem – one of the first to formulate this theory – writes it:

> In sum, the physicist can never subject an isolated hypothesis to experimental test, but only a whole group of hypotheses; when the experiment is in disagreement with his predictions, what he learns is that at least one of the hypotheses constituting this group is unacceptable and ought to be modified; but the experiment does not designate which one should be changed. (Duhem, 1954, p. 187)

As Lenhard and Winsberg (2010) explain, climate models face a particularly strong form of confirmation holism because there is at the same time a high modularity in their development and lots of interactions between their different modules. As they put it: "The complexity of interaction between the modules of the simulation is so severe that it becomes impossible to independently assess the merits or shortcomings of each submodel." (Lenhard & Winsberg, 2010) To illustrate their point, Lenhard and Winsberg (2010) take the example of AMIP and other intercomparison projects, and point out the difficulties these intercomparison projects have historically had to diagnose the causes of the successes and failures of climate models. In their view, because of these difficulties, model convergence is unlikely: in the foreseeable future, there will continue to be a plurality of models making divergent pre-

dictions. Policy makers should therefore accept this plurality, and not wait for a unanimous voice from the climate modeling community.

While we agree with some of the conclusions from Lenhard and Winsberg (2010), we would qualify their claim that climate models themselves face a strong form of confirmation holism. Indeed, a climate model is not just the sum of the code (and the assumptions behind it) and the results it gives for a particular set of simulations. On the contrary, it is a dynamical entity with which it is possible to interact. When climate scientists want to study a climate model, they can initialize it with various conditions, change the parameters of the model, compare the results of simulations with data from observations or high resolution simulations, use simplified or idealized versions of the model or other models, etc. Thanks to these various interactions, climate scientists can acquire a knowledge about the behaviour of a climate model, what it is doing and why.

This knowledge is most of the time collective, because it results from collaboration in research laboratories between different individuals working on separate but complementary aspects of the same climate model.[4]. However, if this knowledge is collective, it stays usually at the level of one research team working on one model. Indeed, due to the complexity of the models involved in CMIP, acquiring knowledge about the behaviour of a climate model takes time and scientists generally focus their efforts on one particular model. In fact, in the literature, we can find many studies investigating the link between the results of a model and its parameterizations (e.g. Hourdin et al., 2013; Notz, Haumann, Haak, Jungclaus, & Marotzke, 2013). But most of these studies are done for one model only – when many models are studied, as in CMIP multi-model ensembles, the details of the parameterizations involved are almost never taken into account. There are, of course, some rare exceptions: in particular, studies comparing radiation codes in different climate models, such as Oreopoulos et al. (2012) and Pincus et al. (2015), where the authors analyze not only the model results, but also the corresponding parameterizations and the assumptions they make[5]. Based on the

[4]This collective aspect of knowledge is not a specificity of climate modeling. On the contrary, as underlined by Hardwig (1985), nowadays, epistemic dependence among scientists is pervasive.

[5]This is due to the fact that the representation of radiation in GCMs is a very different problem from

set of papers we have studied, we claim nevertheless that the large majority of papers using CMIP results do not look at the details of the parameterizations involved. In other words, our analysis suggests that CMIP has promoted analyses which do not discuss the content of GCMs, but consider them as black boxes. Hence, a strong form of confirmation holism does exist: however, unlike Lenhard and Winsberg (2010), we argue that it is not so much GCMs *per se*, but GCMs as they are used in CMIP that are facing it.

In this section, we have examined how CMIP results have been used in scientific papers. We will see in the next section that this gives only a partial view of CMIP influence on climate research – beyond analyses based on their results, intercomparison projects have indeed had structural effects on climate research.

Structural effects of CMIP on climate research

CMIP should not be reduced to a set of simulations performed every 5 or 6 years by the main modeling groups in the world. All the social interactions, questions, and scientific coordination it creates have to be taken into account. CMIP is indeed a conglomerate of technical tools, common objectives, dedicated workshops, and so on – what we will call an infrastructure for climate research.

Let's define more precisely what we mean by this. For that, we will rely on the concept of technological infrastructure introduced and studied by the contemporary philosopher Joseph Pitt. Pitt defines a technological infrastructure as "an historically determined set of mutually supporting artifacts and structures that enable human activity and provide the means for its development" (Pitt, 2000, 129). As far as science is concerned, Pitt explains that his definition does not include only shared machines or techniques, but also all the work relations among scientists which makes the doing of science possible. There is therefore a social

the other parameterizations, and much better grounded. For more details about what makes radiation parameterizations unique, see Pincus and Stevens (2013), Pincus et al. (2015), and Pincus, Mlawer, and Delamere (2019).

component in Pitt's definition of the "technological infrastructure of science". There is also an historical and historiographical aspect behind this concept – it can be used to follow and understand advances of science:

> the mechanism that makes the discoveries of science possible and scientific change mandatory is the technological infrastructure within which that science operates, and that to understand why a science worked the way it did, and why it works the way it does, you need to understand its context, which happens to include in important ways its technological infrastructure. (Pitt, 2000, p. 132)

As an infrastructure, CMIP has shaped climate science in many ways. First, the existence of CMIP has focused the effort of the scientific community on the construction of GCMs. Twenty years ago, Shackley et al. (1998) already underlined that GCMs were commonly considered as the "best climate models". We argue that CMIP has reinforced this trend. In almost all major climate modeling groups in the world, each phase of CMIP has appeared as an international *rendez-vous*. To have a voice in it, every modeling group had to have the most recent version of its GCM ready. It created a pressure around the development of GCM, at the expense of other tools used to study climate.

This supremacy of GCMs is questionable, when considering the many issues exposed in section 2 regarding the validation of GCMs and the interpretation of their results. When modeling complex processes, there is always a trade-off to find between the inclusion of perceived complexity and the understanding of dominant processes, interactions, feedbacks, and uncertainties (Shackley et al., 1998). Yet, for GCMs, there is no trade-off: the balance is clearly in favor of the inclusion of perceived complexity. As a consequence, some climate scientists have argued that hierarchies of models of varying complexity would bring more insights about the climate system and should be more frequently used in climate science (e.g. Bony et al., 2013; Maher et al., 2019). Examples of commendable initiatives to compare GCMs with simpler models already exist, such as: 1. RCMIP (Reduced Complexity Model Intercomparison Project), a unique CMIP6 sub-experiment, which provides a standard protocol for comparing simple models and emulators to the latest CMIP results 2. The use of

single-column versions of GCMs compared to explicit high-resolution simulations to build and test parameterizations (e.g. C. Rochetin, Hourdin, Couvreux, & Jam, 2010; N. Rochetin, Couvreux, Grandpeix, & Rio, 2014).

Despite these efforts, voices have raised in the climate modeling community to protest that the pace of progress was to slow and that GCMs as they were used were not the appropriate scientific response to the challenges posed by global warming. Different alternatives in the practice of climate modeling have been proposed (e.g. Hurrell et al., 2009; T. N. Palmer, 2012; Shukla et al., 2009) – Katzav and Parker (2015) provide an overview and critical examination of these different approaches. More recently, T. Palmer and Stevens (2019) have argued for a new strategy based on higher resolution models and a new approach to parameterizations with stochastic modeling. Though in different ways, all of these proposals offer alternatives to the CMIP dogma.

In addition to focusing the climate community on GCMs, another main effect of CMIP on climate science is to have promoted connections between climate laboratories. It has helped them to talk to each other and share common references. As the historian of science Paul Edwards puts it:

> By permitting regular, direct, and meaningful comparisons of the models with one another and with standardized data sets, [climate model intercomparison projects] have helped to transform climate modeling from a craft activity of individual laboratories into a more modular and standardized collective activity involving virtually all of the world's climate modeling groups; in theoretical terms, they linked a set of isolated systems and created a network. (Edwards, 2010, p. 350)

In practice, these connections were made through different means:

1. First, CMIP made its data freely available from all over the world thanks to a distributed structure, the *Earth System Grid Federation*. It has also supported the development of technical tools to convert the output data of various modeling group into a

standardized format, making therefore their exchange easier.

2. Second, CMIP has spread some common scientific approaches in the climate modeling community, thereby facilitating the comparison of climate models with one another. We can cite for instance abrupt 4 times CO_2 experiments – simulations in which the CO_2 concentration in the atmosphere is immediately and abruptly quadrupled from its pre-industrial values – or transient simulations in which the CO_2 concentration is increased gradually at a rate of 1% per year. Thanks to CMIP, these two kinds of simulations have become standard tools to investigate future climate change with GCMs.

3. Last but not least, dedicated workshops, special issues in scientific papers, etc. have fostered the interest of climate scientists around CMIP results and led to various collaborations between modeling groups.

A side effect of the open-access to CMIP data is the creation of a growing gap between model developers and model users. Before AMIP and CMIP, the results of a GCM simulation were usually analyzed by the few people who had been involved in the development of the corresponding GCM (e.g. Manabe & Wetherald, 1975; Washington, Semtner, Meehl, Knight, & Mayer, 1980). Model users had therefore a critical view of the strengths and weaknesses of the climate model they analyzed, because they had contributed to develop it. When data from intercomparison projects were made freely available, GCMs started to be analyzed by people who had not participated in their development. It resulted in a loss of understanding of climate model results, and increased the tendency to use GCMs only as black boxes (see 2.3).

Nevertheless, in CMIP6 overview paper, the authors claim that CMIP has favored "scientific progress on the most pressing problems of climate variability and change" (Eyring et al., 2016, p. 1949). Is it true? As an infrastructure, did CMIP succeed in making climate research more effective?

Here, we think it is important to distinguish between effectiveness and efficiency. Something is effective if it is adequate to achieve an objective. In contrast, something is efficient if it works in the best possible way. In other words, being effective is about doing the right things, while being efficient is about doing things right. A process can thus be efficient – for instance, if it is fast or cheap – but ineffective – if it is not well suited to the objective we want to achieve.

CMIP has certainly helped climate research to be more efficient. Indeed, it has pooled many time-consuming activities at the level of the research community, and therefore facilitated the work of climate scientists. For a single laboratory, building a simulation protocol is indeed a time-consuming and costly effort. Boundary conditions, forcings and the associated databases, output parameters and their format, as well as versions of the models involved have to be carefully defined. Thanks to CMIP, climate laboratories have shared this burden. The use of standardized format for the output data has also simplified a lot the analyses based on the results of simulations, and in particular, the comparison with data from observations. As Eyring et al. (2016) explain:

> A key to the success of CMIP and one of the motivations for incorporating a wide variety of coordinated modeling activities under a single framework in a specific phase of CMIP (now CMIP6) is the desire to reduce duplication of effort, minimize operational and computational burdens, and establish common practices in producing and analysing large amounts of model output. (Eyring et al., 2016)

We thus agree that the production and the analyses of GCMs simulations have been made more efficient thanks to CMIP. But did it make climate research more effective? CMIP has focused the effort of the climate research community on one specific tool, GCMs, and has promoted the interpretation of this specific tool as a black box. Was it and is it still the most effective way to help climate scientists better understand climate variability and change? The debate stays open.

Conclusion

When looking at the history of CMIP, it seems that this project has known a growing success. It has promoted a coordination at an international level never before achieved in climate science. With all the major modeling groups in the world participating, CMIP has been massively endorsed by the research community. In addition, its results have played a key role in IPCC reports. However, while the analyses of CMIP6 are currently being undertaken, it might be time to pause and reflect about what CMIP has become and what it has brought to climate research.

First, CMIP has focused the attention of the climate research community on GCMs. As the ethnographer of climate modeling Simon Shackley observed, GCMs are often considered as the best models in climate science (Shackley et al., 1998) and the most useful one to predict future climate change. However, this view has been challenged by Bony et al. (2013); Shackley et al. (1998), and others. According to them, though valuable, GCMs should not be seen as the 'panacea" of climate science (Bony et al., 2013). Different models are useful for different purposes and simpler models can also provide valuable insights for understanding climate processes. They should therefore be more frequently used in complementary of more complex GCMs.

Another issue comes from the fact that CMIP promoted the particular use of GCMs as black boxes. Since CMIP data are freely available, anybody can analyse them. As a consequence, the large majority of CMIP analyses are conducted by scientists who have not been involved in the development of the corresponding GCMs, and have therefore a poor knowledge of the content of the climate models at hand. In fact, CMIP has created two distinct communities – model users and model developers – with few interactions between them. The existence of these two communities is an issue for the interpretation of CMIP results. In particular, CMIP results are almost never used to guide model improvement, whereas it was one of the main objective for AMIP.

We argue that there has not been enough explicit debate on the value of CMIP for climate research and policy guidance. We strongly encourage scientists to examine the consequences both for science and society of the particular form of research they have entered with CMIP. In the context of the challenges posed by climate change, given the limited means and computing resources available, there should be more discussions on the goals, epistemology and policy context of this tool at the core of climate research. Hopefully, our historical and epistemological perspectives about what CMIP was and what it is now will not close this debate but open it.

References

Baumhefner, D., & Downey, P. (1978). Forecast intercomparisons from three numerical weather prediction models. *Monthly Weather Review*, *106*(9), 1245-1279.

Bony, S., Stevens, B., Held, I., Mitchell, J., Dufresne, J.-L., Emmanuel, K., ... Senior, C. (2013). Carbon dioxide and climate : Perspectives on a scientific assessment. In G. Asrar & J. Hurrell (Eds.), *Climate science for serving society, research, modeling and prediction priorities* (pp. 391–414). Springer.

Cubasch, U., & et al. (2001). *Climate change 2001: The scientific basis* (J. Houghton et al., Eds.). Cambridge University Press.

Duhem, P. (1954). *The aim and structure of physical theory.* Princeton University Press.

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming.* The MIT Press.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. doi: https://doi.org/10.5194/gmd-9-1937-2016

Frisch, M. (2019). Calibration, validation, and confirmation. In C. Beisbart & N. J. Saam (Eds.), *Computer simulation validation: Fundamental concepts, methodological frameworks, and philosophical perspectives* (pp. 981–1004). Springer International Publishing. doi: 10.1007/978-3-319-70766-2_41

Gabaix, X., & Laibson, D. (2008). The seven properties of good models. In *The foundations of positive and normative economics.* Oxford University Press. doi: 10.1093/acprof:oso/9780195328318.003.0012

Gaetani, M., & Mohino, E. (2013). Decadal prediction of the Sahelian precipitation in CMIP5 simulations. *Journal of Climate*, *26*(19), 7708-7719. doi: 10.1175/JCLI-D-12-00635.1

Gates, W. L. (1992). An AMS continuing series: Global change–AMIP: The atmospheric model intercomparison project. *Bulletin of the American Meteorological Society*, *73*(12), 1962-1970.

Gates, W. L. (2015). *Comments of the history of the working group on numerical experimentation.* (http://www.wmo.int/pages/prog/arep/wwrp/new/Presentations_wgne30_March_2015.html)

Gates, W. L., Boyle, J. S., Covey, C., Dease, C. G., Doutriaux, C. M., Drach, R. S., ... Williams, D. N. (1999). An overview of the results of the atmospheric model intercomparison project (AMIP I). *Bulletin of the American Meteorological Society*, *80*(1), 29-56.

Giorgi, F., & et al. (2001). *Climate change 2001: The scientific basis* (J. Houghton et al., Eds.). Cambridge University Press.

Guemas, V., García-Serrano, J., Mariotti, A., Doblas-Reyes, F., & Caron, L.-P. (2015). Prospects for decadal climate prediction in the Mediterranean region. *Quarterly Journal of the Royal Meteorological Society*, *141*(687), 580-597. doi: 10.1002/qj.2379

Hardwig, J. (1985). Epistemic dependence. *Journal of Philosophy*, *82*(7), 335–349. doi: jphil198582747

Houghton, D. D., & Irvine, W. S. (1976). A case study comparison of the performance of operational prediction models used in the United States. *Monthly Weather Review*, *104*(7), 817-827.

Houghton, J., Callander, B., & Varney, S. (Eds.). (1992). *Climate change 1992: the supplementary report to the IPCC scientific assessment.* Cambridge University Press.

Houghton, J., Jenkins, G., & Ephraums, J. (Eds.). (1990). *IPCC first assessment report 1990.* Cambridge University Press.

Hourdin, F., Grandpeix, J.-Y., Rio, C., Bony, S., Jam, A., Cheruy, F., ... Roehrig, R.

(2013, May 01). LMDZ5B: the atmospheric component of the IPSL climate model with revisited parameterizations for clouds and convection. *Climate Dynamics*, *40*(9), 2193–2222. doi: 10.1007/s00382-012-1343-y

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., ... Williamson, D. (2017). The art and science of climate model tuning. *Bulletin of the American Meteorological Society*, *98*(3), 589-602.

Hu, Y., Tao, L., & Liu, J. (2013). Poleward expansion of the Hadley circulation in CMIP5 simulations. *Advances in Atmospheric Sciences*, *30*(3), 790–795. doi: 10.1007/s00376-012-2187-4

Hurrell, J., Meehl, G. A., Bader, D., Delworth, T. L., Kirtman, B., & Wielicki, B. (2009). A unified modeling approach to climate system prediction. *Bulletin of the American Meteorological Society*, *90*(12), 1819-1832. doi: 10.1175/2009BAMS2752.1

Jayasankar, C. B., Surendran, S., & Rajendran, K. (2015). Robust signals of future projections of indian summer monsoon rainfall by IPCC AR5 climate models: Role of seasonal cycle and interannual variability. *Geophysical Research Letters*, *42*(9), 3513-3520. doi: 10.1002/2015GL063659

Jørgensen, S. E. (2008). Overview of the model types available for development of ecological models. *Ecological Modelling*, *215*(1), 3 - 9. doi: https://doi.org/10.1016/j.ecolmodel.2008.02.041

Katzav, J., & Parker, W. S. (2015). The future of climate modeling. *Climatic Change*, *132*(4), 475–487. doi: 10.1007/s10584-015-1435-x

Kent, C., Chadwick, R., & Rowell, D. P. (2015). Understanding uncertainties in future projections of seasonal tropical precipitation. *Journal of Climate*, *28*(11), 4390-4413. doi: 10.1175/JCLI-D-14-00613.1

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

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., & Meehl, G. A. (2010). Challenges in combining projections from multiple climate models. *Journal of Climate*, *23*(10), 2739-2758.

Knutti, R., Masson, D., & Gettelman, A. (2013). Climate model genealogy: Generation CMIP5 and how we got there. *Geophysical Research Letters*, *40*(6), 1194-1199.

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., & Eyring, V. (2017). A climate model projection weighting scheme accounting for performance and interdependence. *Geophysical Research Letters*, *44*(4), 1909-1918. doi: 10.1002/2016GL072012

Lenhard, J., & Winsberg, E. (2010). Holism, entrenchment, and the future of climate model pluralism. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *41*(3), 253–262.

Long, D. J., & Collins, M. (2013). Quantifying global climate feedbacks, responses and forcing under abrupt and gradual CO2 forcing. *Climate Dynamics*, *41*(9), 2471–2479. doi: 10.1007/s00382-013-1677-0

Luomaranta, A., Ruosteenoja, K., Jylhä, K., Gregow, H., Haapala, J., & Laaksonen, A. (2014). Multimodel estimates of the changes in the baltic sea ice cover during the present century. *Tellus A: Dynamic Meteorology and Oceanography*, *66*(1), 22617. doi: 10.3402/tellusa.v66.22617

Luther, F. M., Ellingson, R. G., Fouquart, Y., Fels, S., Scott, N. A., & Wiscombe, W. J. (1988). Intercomparison of radiation codes in climate models (ICRCCM): Longwave clear-sky results—a workshop summary. *Bulletin of the American Meteorological Society*, *69*(1), 40-48.

Maher, P., Gerber, E. P., Medeiros, B., Merlis, T. M., Sherwood, S., Sheshadri, A., ... Zurita-Gotor, P. (2019). Model hierarchies for understanding atmospheric circulation. *Reviews of Geophysics*, *57*(2), 250-280. doi: 10.1029/2018RG000607

Manabe, S., & Bryan, K. (1969). Climate calculations with a combined ocean-atmosphere model. *Journal of the Atmospheric Sciences*, *26*, 786-789.

Manabe, S., & Wetherald, R. T. (1975). The effects of doubling the CO2 concentration on the climate of a general circulation model. *Journal of the Atmospheric Sciences*, *32*(1), 3-15. doi: 10.1175/1520-0469(1975)032⟨0003:TEODTC⟩2.0.CO;2

Masson, D., & Knutti, R. (2011). Climate model genealogy. *Geophysical Research Letters*, *38*(8). doi: 10.1029/2011GL046864

McAvaney, & et al. (2001). *Climate change 2001: The scientific basis* (J. Houghton et al., Eds.). Cambridge University Press.

Meehl, G., Covey, C., McAvaney, B., Latif, M., & Ronald, S. (2005). Overview of the coupled

model intercomparison project. *Bulletin of the American Meteorological Society*(86), 89-93.

Meehl, G. A. (1995). Global coupled general circulation models. *Bulletin of the American Meteorological Society*. (Meeting Summary)

Meehl, G. A., Boer, G. J., Covey, C., Latif, M., & Stouffer, R. J. (1997). Intercomparison makes for a better climate model. *Eos, Transactions American Geophysical Union*, *78*(41), 445–451.

Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J. F. B., ... Taylor, K. E. (2007). The WCRP CMIP3 multimodel dataset: A new era in climate change research. *Bulletin of the American Meteorological Society*, *88*(9), 1383-1394.

Meehl, G. A., Goddard, L., Boer, G., Burgman, R., Branstator, G., Cassou, C., ... Yeager, S. (2014). Decadal climate prediction: An update from the trenches. *Bulletin of the American Meteorological Society*, *95*(2), 243-267. doi: 10.1175/BAMS-D-12-00241.1

Nakicenovic, N., & Swart, R. (2000). *Emissions scenarios - special report of the intergovernmental panel on climate change.* Cambridge University Press.

National Academy of Sciences. (1975). Survey of the climate simulation capability of global circulation models. *Understanding Climatic Change*, 196–239.

Notz, D., Haumann, F. A., Haak, H., Jungclaus, J. H., & Marotzke, J. (2013, 6). Arctic sea-ice evolution as modeled by Max Planck Institute for Meteorology's earth system model. *Journal of Advances in Modeling Earth Systems*, *5*(2), 173–194. doi: 10.1002/jame.20016

Oreopoulos, L., Mlawer, E., Delamere, J., Shippert, T., Cole, J., Fomin, B., ... Rossow, W. B. (2012). The continual intercomparison of radiation codes: Results from phase I. *Journal of Geophysical Research: Atmospheres*, *117*(D6). doi: 10.1029/2011JD016821

Palmer, T., & Stevens, B. (2019). The scientific challenge of understanding and estimating climate change. *Proceedings of the National Academy of Sciences*, *116*(49), 24390–24395. doi: 10.1073/pnas.1906691116

Palmer, T. N. (2012). Towards the probabilistic Earth-system simulator: a vision for the future of climate and weather prediction. *Quarterly Journal of the Royal Meteorological Society*, *138*(665), 841-861. doi: 10.1002/qj.1923

Parker, W. S. (2011). When climate models agree: The significance of robust model predictions. *Philosophy of Science*, *78*(4), 579–600.

Penalba, O., & Rivera, J. (2013, 09). Future changes in drought characteristics over Southern South America projected by a CMIP5 multi-model ensemble. *American Journal of Climate Change*, *2*, 173-182. doi: 10.4236/ajcc.2013.23017

Pincus, R., Mlawer, E. J., & Delamere, J. S. (2019). Balancing accuracy, efficiency, and flexibility in radiation calculations for dynamical models. *Journal of Advances in Modeling Earth Systems*, *11*(10), 3074-3089. doi: 10.1029/2019MS001621

Pincus, R., Mlawer, E. J., Oreopoulos, L., Ackerman, A. S., Baek, S., Brath, M., ... Schwarzkopf, D. M. (2015). Radiative flux and forcing parameterization error in aerosol-free clear skies. *Geophysical Research Letters*, *42*(13), 5485-5492. doi: 10.1002/2015GL064291

Pincus, R., & Stevens, B. (2013). Paths to accuracy for radiation parameterizations in atmospheric models. *Journal of Advances in Modeling Earth Systems*, *5*(2), 225-233. doi: 10.1002/jame.20027

Pirtle, Z., Meyer, R., & Hemilton, A. (2010). What does it mean when climate models agree? A case for assessing independence among general circulation models. *Environmental Science and Policy*.

Pitt, J. (2000). *Thinking about technology: Foundations of the philosophy of technology* (N. Y. S. B. Press, Ed.).

Potter, G. L., & Gates, W. L. (1984). A preliminary intercomparison of the seasonal response of two atmospheric climate models. *Monthly Weather Review*, *112*(5), 909-917.

Qu, X., & Hall, A. (2014, Jan 01). On the persistent spread in snow-albedo feedback. *Climate Dynamics*, *42*(1), 69–81. doi: 10.1007/s00382-013-1774-0

Rochetin, C., Hourdin, F., Couvreux, F., & Jam, A. (2010). Resolved versus parametrized boundary-layer plumes. part II: Continuous formulations of mixing rates for mass-flux schemes. *Boundary-Layer Meteorology*, *135*(3), 469-483. doi: 10.1007/s10546-010-9478-z

Rochetin, N., Couvreux, F., Grandpeix, J.-Y., & Rio, C. (2014). Deep convection triggering by boundary layer thermals. part I: Les analysis and stochastic triggering formulation.

Journal of the Atmospheric Sciences, *71*(2), 496-514. doi: 10.1175/JAS-D-12-0336.1

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*, 159-205.

Shukla, J., Hagedorn, R., Miller, M., Palmer, T. N., Hoskins, B., Kinter, J., ... 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. doi: 10.1175/2008BAMS2759.1

Srivastava, A. K., & DelSole, T. (2014). Robust forced response in South Asian summer monsoon in a future climate. *Journal of Climate*, *27*(20), 7849-7860. doi: 10.1175/JCLI-D-13-00599.1

Stainforth, D. A., Downing, T. E., Washington, R., Lopez, A., & New, M. (2007). Issues in the interpretation of climate model ensembles to inform decisions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *365*(1857), 2163-2177. doi: 10.1098/rsta.2007.2073

Stevenson, S. L. (2012). Significant changes to ENSO strength and impacts in the twenty-first century: Results from CMIP5. *Geophysical Research Letters*, *39*(17). doi: 10.1029/2012GL052759

Taschetto, A. S., Gupta, A. S., Jourdain, N. C., Santoso, A., Ummenhofer, C. C., & England, M. H. (2014). Cold tongue and warm pool ENSO events in CMIP5: Mean state and future projections. *Journal of Climate*, *27*(8), 2861-2885. doi: 10.1175/JCLI-D-13-00437.1

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*(4), 485-498.

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. doi: 10.1098/rsta.2007.2076

Washington, W. M., Semtner, A. J., Meehl, G. A., Knight, D. J., & Mayer, T. A. (1980). A general circulation experiment with a coupled atmosphere, ocean and sea ice model.

Journal of Physical Oceanography, *10*(12), 1887-1908. doi: 10.1175/1520-0485(1980)010⟨1887:AGCEWA⟩2.0.CO;2

WGCM. (2006). Report of the tenth session of the JSC/CLIVAR working group on coupled modelling [WCRP Informal Report No.5/2007].

WGCM. (2012). Report of the sixteenth session of the working group on coupled modelling [WCRP Report No.2/2013].

WGCM. (2014). Report of the eighteenth session of the working group on coupled modelling [WCRP Report No.2/2015].

Winsberg, E. (2018). *Philosophy and climate science*. Cambridge University Press. doi: 10.1017/9781108164290

Zomer, R. J., Trabucco, A., Metzger, M. J., Wang, M., Oli, K. P., & Xu, J. (2014). Projected climate change impacts on spatial distribution of bioclimatic zones and ecoregions within the Kailash sacred landscape of China, India, Nepal. *Climatic Change*, *125*(3), 445–460. doi: 10.1007/s10584-014-1176-2

Zou, L., & Zhou, T. (2015, 06). Asian summer monsoon onset in simulations and CMIP5 projections using four Chinese climate models. *Advances in Atmospheric Sciences*, *32*. doi: 10.1007/s00376-014-4053-z

Acknowledgements

The authors would like to thank Sandrine Bony, Julie Jebeile, Nicolas Rochetin, and Romain Roehrig for helpful comments and suggestions.

Further reading

The set of papers based on CMIP5 results analysed in section 2 can be found in the supplementary materials.