**Multi-model ensembles in climate science: mathematical structures and expert judgements**

**Julie Jebeile** (Institute of Philosophy & Oeschger Center for Climate Change Research, University of Bern, Switzerland)
**Michel Crucifix** (Earth and Life Institute, Université catholique de Louvain, Belgium)

*Studies in History and Philosophy of Science Part A*, 2020, https://doi.org/10.1016/j.shpsa.2020.03.001

**Keywords**: climate model; multi-model ensemble; climate uncertainty; mathematical representation; parameterisation; expert judgement

**Abstract**. Projections of future climate change cannot rely on a single model. It has become common to rely on multiple simulations generated by Multi-Model Ensembles (MMEs), especially to quantify the uncertainty about what would constitute an adequate model structure. But, as Parker points out (2018), one of the remaining philosophically interesting questions is: "How can ensemble studies be designed so that they probe uncertainty in desired ways?" This paper offers two interpretations of what General Circulation Models (GCMs) are and how MMEs made of GCMs should be designed. In the first interpretation, models are combinations of modules and parameterisations; an MME is obtained by "plugging and playing" with interchangeable modules and parameterisations. In the second interpretation, models are aggregations of expert judgements that result from a history of epistemic decisions made by scientists about the choice of representations; an MME is a sampling of expert judgements from modelling teams. We argue that, while the two interpretations involve distinct domains from philosophy of science and social epistemology, they both could be used in a complementary manner in order to explore ways of designing better MMEs.

## 1. Introduction

The climate system is commonly modelled as an object subdivided into components, such as the atmosphere, the oceans, the land surface, the ice sheets, and the biosphere, and submitted to a number of external influences, including fluctuations in incoming solar radiation, volcanic eruptions, and human activities.[1] The dynamic range of processes involved in the climate system and the importance of living processes produce a level and a quality of complexity that are challenging to simulate. Even a General Circulation Model (GCM), involving several million lines of computer code, can only approximately represent climate dynamics based on idealisations under forcing scenarios and specific boundary conditions.

Hence, it is generally agreed that projections of future climate change should not rely on a single model. For this purpose, it has become common to rely on multiple simulations generated by Multi-Model Ensembles (MMEs). The status of these MMEs constitutes a subject of increasing philosophical attention, notably in the work of Parker (e.g. Parker 2006, 2010a, 2010b, 2013, 2018; Betz 2009; Frigg, Stainforth and Smith 2013, 2015; Katzav, Dijkstra and de Laat 2012; Lenhard and Winsberg 2010; Winsberg 2012, 2018).

[1] See Werndl (2016) for a discussion of definitions of climate and climate change.

Specifically, MMEs provide a means to estimate the uncertainty induced by choices of representation of the specific processes at work in the climate system. This form of uncertainty is termed "structural uncertainty" in the standard scientific literature (Tebaldi and Knutti 2007; Meinshausen et al. 2009; Knutti et al. 2010; Flato et al. 2013, AR5WGI chapter 9, p. 754-755).[2] Parker makes it clear that structural uncertainty is uncertainty about what would constitute an adequate model structure but cannot be uncertainty about what would constitute a *perfect* model structure, for it is explored in practice via state-of-the-art models that are imperfect. "After all, adequacy (not perfection) is all that is really needed, and it is plausible that the structures of today's models are adequate for some predictive purposes of interest" (Parker 2010b, p. 991).

Created in 1995, the Coupled Model Intercomparison Project (CMIP) is nowadays the reference framework in which GCMs are gathered into MMEs. For example, twenty-three GCMs, developed in Australia, Canada, China, France, Germany, Japan, Korea, Norway, Russia, the United Kingdom and the United States of America, composed the MME of CMIP5 built in 2013. CMIP was originally developed to enable scientists to compare model outputs in a consistent fashion, and thus identifying robust outputs, shared biases, the origins of disagreements, and which specific processes require more understanding in order to improve the models. It was later used for uncertainty quantification.

However, it is worth questioning whether MMEs such as the ones built in CMIP are well adapted to quantify climate uncertainties. According to a common critique, MMEs are "ensembles of opportunity" (Meehl et al. 2007, p. 754; Tebaldi and Knutti 2007; Knutti et al. 2010) in that their members are not designed in the first place to sample the range of uncertainty, but are rather the state-of-the-art models available at the time, provided by the modelling centres willing to participate (Parker 2010a, 2010b, 2011, 2013; Katzav and Parker 2015). Indeed, any modelling centre may in principle apply for archiving its own GCM outputs in the CMIP database, as long as the model complies with the imposed standards of CMIP.

What this criticism tells us is that, when scientists historically first saw the opportunity to take advantage of the plurality of GCMs developed all over the world, it was too late to build MMEs with adequate properties for quantifying an uncertainty range. We analyse the situation as an epistemological change: When building GCMs, scientists design and tune them to get the representations of the climate that are expected to provide the most accurate projections. From this perspective, an ensemble is "a collection of best guesses" (Parker 2013), i.e. a set of models that are roughly equally good and equally bad. However, covering the full range of uncertainty requires more than a collection of best guesses. The GCMs must jointly contribute to forming a representative sample of climate possibilities, and this sample must possess adequate properties to this end. Given the arguments that MMEs are ensembles of opportunity, we will first make clear what these adequate properties are commonly supposed to be: they include systematicity, comprehensiveness, and model independence (Section 2).

[2] Four kinds of uncertainties are presently identified in IPCC reports: first, *internal variability uncertainty,* stemming from the chaotic and spontaneously varying nature of the climate system; second, *model uncertainty*, due to omissions, idealisations and incomplete knowledge of the climate system represented by the model; third, *scenario uncertainty,* due to dependence on socioeconomic factors including future (global) (geo)political agreements to control greenhouse gases and aerosol emissions, technological advances or population movements; and finally, uncertainties in future natural radiative forcing caused by unpredicted solar and volcanic activities (Kirtman et al. 2013, AR5WGI chapter 11, section 11.3.6.2, p. 1007). Within model uncertainty, it is common to distinguish structural from parameter uncertainty: *parameter uncertainty* stems from the fact that, given a choice of structural representation, the best values of the constant parameters to be used in the model equations are unknown or ambiguously defined (Rougier 2007; Winsberg 2012; Frigg, Stainforth and Smith 2013).

The "ensemble of opportunity" criticism implicitly contains a positive message: MMEs could, in principle at least, be better designed. In other words, if we could coordinate the worldwide development of GCMs, then we might be able to design the members of MMEs to provide, for the same computing cost, a more reliable quantification of uncertainties about future climate.[3] This suggestion allows us to address what Parker (2018) introduces as one of the remaining philosophically interesting questions related to model ensembles: "How can ensemble studies be designed so that they probe uncertainty in desired ways?"

In order to allow for the construction of better ensembles, we first need a characterisation of what GCMs are, and how MMEs should be designed. This is what our contribution aims to offer.

Beforehand, we make explicit why, following Parker (e.g. 2010b, 2013), MME optimisation is particularly hard to conceptualise (Section 3). We then introduce two views on how an MME should be constructed and assessed, depending on the object we assume an ensemble is a sampling of. Hence these views come with different interpretations of GCMs constituting MMEs. These interpretations are not complete or mutually exclusive. They rather highlight different aspects of GCMs.

The first interpretation is suggested by the definition of structural uncertainty: Models are combinations of modules and parameterisations. An ensemble is here obtained by "plugging and playing" with interchangeable modules and parameterisations (Section 4).

The second interpretation is suggested by practices underlying climate modelling as a social and epistemic process: Models are aggregations of expert judgements that result from a history of epistemic decisions made by scientists about the choices of representation. An ensemble is here a sampling of expert judgements from modelling teams (Section 5).

Modules and parameterisations are mathematical structures and are therefore of a different nature than expert judgements. Hence, as we will show, the two interpretations involve distinct domains from philosophy of science and social epistemology. Nevertheless, because they illuminate different aspects of GCMs, both interpretations allow us to highlight distinct problems related to MMEs and therefore are complementary to each other in our exploration of ways to design better MMEs (Section 6).

More precisely, we will argue that the first interpretation helps in properly formalising the adequate properties of MMEs (6.1) while remaining silent on the way we should define the space of model structures (6.2). We will further argue that, unlike the first interpretation (6.3), the second interpretation accounts for the fact that confidence in model projections is generated by the social and historical processes underlying model assessment (6.4) and also for the influence of non-epistemic values in choices of representation (6.5). We will finally suggest some consequences of adopting both interpretations for designing better MMEs (Section 7).

## 2. Systematicity, comprehensiveness and model independence

[3] In this paper, we put aside the recent scientific attempts at improving MMEs. For example, methods have been developed in order to go beyond the "one-model-one-vote" approach in which each model deserves the same weight within the ensemble. They assign different weights to ensemble members to get more reliable projections (Sanderson, Knutti and Caldwell 2015a, 2015b). A more recent attempt is to build subsets of models, within the main ensemble, that reach independence with regard to specific results of interest (Abramowitz, Herger, Gutmann et al. 2018; Herger, Abramowitz, Knutti, et al. 2018). These attempts deserve philosophical attention but will not be discussed here.

The definition of the adequate properties of MMEs, as well as how they should be met, remain the subject of important ongoing discussions within the scientific community. The arguments underlying the idea that MMEs are "ensembles of opportunity" convey at the same time an idea of which properties the MMEs ought to have. In this section, we present succinctly two well-known criticisms expressed in both the scientific and philosophical literature so to make these properties clear.

A first concern is that, because MMEs rely on self-selection by modelling groups, the spread of projections is "neither systematic nor comprehensive." This is described within the last IPCC report in the following terms: "because the number of models is relatively small, and the contribution of model output to public archives is voluntary, the sampling of possible futures is neither systematic nor comprehensive" (Collins et al. 2013, p. 1036).

In statistics, a sampling scheme is said to be systematic when it follows an algorithm which has been designed to confer good properties to the statistical estimator (such as asymptotic convergence, or absence of statistical bias). To use this technical definition, however, we would need to have first specified what is to be sampled. The criterion of comprehensiveness, on the other hand, suggests a deliberate effort to cover all possibilities, including the extreme ones.

That the construction of the model ensembles does not follow a systematic sampling process is clear enough. It has been recognised that "Model builders put forward various ideas based on their wisdom and experience, as well as their idiosyncratic interests and prejudices" and thereby "Model improvements are often the result of serendipity rather than systematic analysis" (Held 2005, p. 1611). An undesirable consequence of being non-systematic and non-comprehensive is that the tail of distributions is by construction subsampled (Räisänen 2007). This means that the outcomes that are considered as unlikely but nevertheless plausible are hardly covered by the range of the projections generated by the ensemble. And yet such outcomes can be important for political decision-making.

A second pitfall is that models may share design biases (e.g. Tebaldi and Knutti 2007; Knutti et al. 2010; Knutti, Masson and Gettelman 2013; Annan and Hargreaves 2017). They may all exclude some important processes, or all misrepresent a specific process the same way. It seems that this could be better avoided if models were designed more independently of each other. Today, a modelling centre may contribute to CMIP with more than one model. In that case, important pieces of computer code overlap in these different models because they come from the same lineage and, more generally, emerge from the same in-house traditions (Flato et al. 2013, AR5WGI chapter 9). Besides this particular case, the history of collaborations across modelling centres has generated a complex genealogy with shared modules (e.g. the CICE sea-ice model is found in different climate models) and model assumptions (Knutti, Masson and Gettelman 2013).

In a nutshell, the two main arguments that MMEs are ensembles of opportunity suggest that the spread of projections must in principle be systematic and comprehensive, and that the MME must be a sample of independent models. Now that we have made clear some of the adequate properties of MMEs, let us make explicit why MME optimisation is particularly hard to conceptualise.

## 3. Present state of the question: the double challenge of conceptualisation

Parker points out at several occasions (2010b p. 989, 2013 p. 220, 2018) that MME optimisation is not easy to conceptualise, and should be given philosophical priority.

The main problem, she shows, is the difficulty to sample an appropriate space of ensemble members. This becomes clear when she highlights the major difference between MMEs and Perturbed-Physics Ensembles (PPEs) (2013). While PPEs investigate parameter uncertainty, i.e., uncertainty regarding the value of parameters within a given model structure, MMEs investigate structural uncertainty with several model structures. It follows that, in PPEs, "the space of possibilities in which we are to identify plausible alternatives is clear: it is a space of numerical values", whereas, in MMEs, "the space of possibilities instead ranges over model structures—simulation algorithms to be implemented on machines with specified precision and so forth" (2010b, p. 990).

Therefore, while both ensemble approaches are computationally greedy, PPEs are better specified than MMEs. In PPEs, the "computational roadblock" is due to the high number of parameters to investigate, and yet is expected to be overcome in principle by improved computing approaches. By contrast, in MMEs, the systematic uncertainty analysis is overwhelming: "we do not want to set ourselves the task of identifying all plausibly adequate model structures, for this would seem to require that we survey all possible algorithms that might be implemented on today's computers (and as part of those algorithms, all combinations of all mathematical functions)—a truly mind-boggling task" (2010b, p. 991).

Thus, Parker emphasises that a precondition for systematic uncertainty analysis is to specify and circumscribe a finite collection of plausibly adequate model structures, whose size would depend on the extent of our background knowledge. But then we would still have to know how to sample systematically from this collection of structures: "even if we can specify such a collection, then unless it includes only a small, finite number of model structures, so that we can simply try them all [...], we will face the further challenge of determining what it means to sample systematically from such a collection of structures; this is not at all obvious" (2010b, p. 991).

In a nutshell, state-of-the-art MMEs are ensembles of opportunity that have not been initially designed to sample structural uncertainty, but, even if we now want to design better MMEs, we would face a double challenge: specifying the space of plausibly adequate model structures and sampling systematically from this space. With this double challenge in mind, we now seize the problem of MME optimisation by offering and discussing two views on how MMEs could be designed; each of them is based on a specific interpretation of what models are.

## 4. Models as Modules & Parameterisations

In this section, we present the first interpretation of models as modules and parameterisations (4.1), and then we give reasons that justify such an interpretation (4.2 and 4.3).

### 4.1. Interpretation of models in mathematical and algorithmic terms

According to the first interpretation, GCMs are assemblages of *modules*. Modules are mathematical models that represent specific climate components, for example, the atmosphere, the oceans, sea ice, land surface, plus possibly the atmospheric chemistry, the ocean tracers, and the vegetation dynamics. These models are built out of well-confirmed physical principles as well as a number of idealisations.

The domain covered by a module is associated with a grid dividing the domain into grid cells, which supports the discretisation of the equations of fluid motion that determine the behaviour of the atmosphere and oceans. Each module also contains a number of parameterisations. For example, the module "atmosphere" can contain, among other things, parameterisations for atmospheric convection, cloud formation, atmospheric mixing by gravity waves, evapotranspiration and radiative scattering by aerosols.

Parameterisations can be considered as a sort of idealisation. They are equations representing phenomena known to occur within grid cells (convection, radiative transfer). Depending on knowledge of the process and available observations, some parameterisations are heavily constrained by theoretical foundations and measurements in laboratory experiments (radiative transfers, gravity waves), some may be based on field measurements (evapotranspiration), and others are rather based on quite simplified, idealised conceptualisations of physical processes (deep convection). The degree of idealisation in some parameterisations is such that there are generally a variety of possible formulations for a single process.

Different versions of GCMs could be obtained by interchanging modules. From this perspective, a set of GCMs can be seen as a set of combinations of modules, where, for each module type, one has combined parameterisations for each phenomenon, and has considered, for each parameterisation, the range of all physically plausible parameter values. Note that we remain silent at this stage on how to define the space of modules and parameterisations to be considered; we will come back later to this issue (Section 7.2). For the sake of the argument, let us consider that a theoretical population of GCMs can be defined as the set of all possible combinations of modules, parameterisations, and parameter values, assuming that "all possible combinations" are not infinite.

The MME is a statistical sample of that population. Statistical theory seems indeed to provide the formal framework to establish what a good MME is. This is suggested in AR5 by Collins and co-authors when they write that "the difficulty in producing quantitative estimates of uncertainty based on multiple model output originates in their peculiarities as a *statistical* sample, neither random nor systematic" (2013, p. 1040, emphasis added).

Within the framework of statistical theory, building a comprehensive and systematic ensemble seems to require an automatisation of the sampling of the modules and parameterisations. As we have here defined the population of GCMs as an enumerable set, we can foresee the possibility of sampling this population automatically. We can imagine an algorithmic system that probes the stock of available modules and parameterisations in an automated way, designed so as to satisfy the requirements of systematicity and comprehensiveness, which creates MMEs by "plugging and playing" with modules and parameterisations.

## 4.2. Justification by common definition of structural uncertainty

The first interpretation we have provided is suggested by the very definition of structural uncertainty. Structural uncertainty refers to the uncertainty induced by choices of representation of the specific processes at work in the climate system, which includes the form of parameterisations (Palmer 2005).

From the assumption that the range of available parameterisations and modules is an expression of uncertainty about the representation of these processes, it follows that the variance of the population of models probes this uncertainty.

Structural uncertainty is indeed explored via MMEs, in that idealisations vary from model to model "in terms of the fundamental numeric and algorithmic structures, forms and values of parameterisations, and number and kinds of coupled processes included" (Collins et al. 2013, AR5WGI chapter 12, p. 1039). The diversity of models thus stems from differences in the choice of a numerical scheme, the choices of modules, and the choices of parameterisations.

As it happens, the first interpretation considers GCMs as collections of modules and parameterisations that may differ from a member of the MME to another. Thus, this interpretation is justified by the very idea that sampling modules and parameterisations is expected to quantify the structural uncertainty, and that statistics provide a formal framework for expressing this uncertainty with probabilities (Tebaldi and Knutti 2007; Meinshausen et al. 2009; Knutti et al. 2010; Flato et al. 2013, AR5WGI chapter 9, p. 754-755).

## 4.3. Justification by methodological aspects

Some aspects of the methodology commonly followed in climate modelling support the first interpretation.

First of all, the first interpretation of GCMs as sets of modules and parameterisations is compatible with the way, in practice, a GCM is built. It is composed of subroutines that are often arranged in different files and embedded within a main program; the main program calls the subroutines, passing relevant fluxes and state variables to them. This way of organising the computer code into subroutines mirrors the common representation of the climate system as an entity which can be decomposed into sub-components, including the atmosphere, the oceans, the ice sheets, which exchange heat, water, and momentum. In this representation, the dynamics of the atmosphere, in turn, emerges from the interactions between different objects: cloud formation, precipitation, radiative transfer, heat exchanges, and chemical reactions.

Furthermore, the first interpretation also comes with the possibility of developing the different modules by distinct teams. Such a division of labour is common in building GCMs. For instance, the ocean model NEMO is developed in France as a stand-alone model of ocean dynamics and is used for research activities and forecasting services in ocean sciences (NEMO 2019). However, it is also coupled with the LIM model of sea-ice dynamics developed in Belgium, and then is coupled with the IFS model developed at the European Center for Middle-Range Weather Forecast in Reading, where it is also used as a stand-alone model (ECMWF 2019). The ECEARTH model is the combination of these different modules, interfaced by a technical piece of software called a coupler, which is designed to allow these different assemblages (ENES 2015).

Given that IFS is not the only possible software for simulating atmosphere dynamics, one can imagine a collection of alternatives to ECEARTH by plugging in successive alternatives to IFS. Within IFS, different parameterisations of the moist convection scheme are available; likewise, different parameterisations are available for the heat and momentum exchanges with soil, vegetation, snow and mountains. These possibilities provide further opportunities to augment the collection of possible assemblages.

A number of investigators encourage developing a collection of climate models by combining alternative modules. Kalnay et al. (1989) advocated “rules for interchange of physical parameterisations” supported by methods of “plug compatibility” (p. 620). Similarly, one can read that the so-called Community Climate Model (CCSM) was built on software engineering process ensuring its “modularity” and “extensibility” (Drake, Jones and Carr 2005). The “Modular Earth Submodel System” relies on a “universal coupler” allowing the user to easily control which components are being plugged on a “standard interface” (Jöckel et al. 2005).

That said, we will later argue that, even if the modules and parameterisations look individually decent, the practice of plugging them together does not always generate an acceptable GCM, because it may lead to unexpected and undesirable effects, and also because modularity in GCMs is fuzzy (Section 6.2).

Now that we have offered a first interpretation of MME members as sets of modules and parameterisations, let us turn to the second interpretation. After this, we will argue that both interpretations, while involving distinct philosophical domains, are complementary in reflecting on how to design better MMEs.

## 5. Models as Aggregations of Expert Judgements

In this section, we present the second interpretation of models as aggregations of expert judgements (5.1) and then we give reasons that justify such an interpretation (5.2).

### 5.1. Interpretation of models in terms of expert judgements

According to the second interpretation, models are the products of histories of epistemic decisions made by scientists about the choice of representations. This interpretation is suggested by the social dynamics underlying scientific practices in climate modelling.

Throughout the construction of GCMs, modellers make epistemic decisions about the representations underlying the models. Such decisions include the choices of the numerical scheme, the modules and the parameterisations that the models contain. Models pass multiple quality control procedures during which assessments are made on their acceptability.

Besides the direct choices of representation, the criteria of acceptability of internal and external consistency are also subject to judgement. Internal consistency includes the absence of numerical errors (bugs) and compatibility between different scientific assumptions. External consistency refers to consistency of our knowledge of the state of dynamics of the climate system. The collection of expert judgements made throughout the history of a model is effectively encoded in the code of the GCM, along with boundary conditions, parameters, and even minute details such as compiler options.

The decisions underlying the choice of representations involve expert judgements from modelling teams. Expert judgements are based on objective knowledge but also contain subjective components that depend on the experience of the experts. On this view, building an ensemble is about sampling expert judgements, historically made by modelling teams. In other words, it is about sampling expert judgements from different research institutes. They are algorithmically formalised and encoded within the software including the computer code, the standard boundary conditions and the parameter values.

### 5.2. Justification by practices

The second interpretation finds its justification in the scientific practices underlying climate modelling.

First of all, models are not built from scratch. The representations they contain often stem from older versions that have been embedded and tested within previous models. It is not uncommon that representations are revised and reused from one generation of models to another. In other words, the genesis of the representations that a model contains can precede the building of any particular model. Models belong to families and genealogies of models (Knutti, Masson and Gettelman 2013). They succeed each other and, for each new generation, scientists make decisions on which representations will be retained from one model to the next and how they will be improved.

Second, when developing a climate model, climate scientists may choose to couple certain pre-existing components for practical reasons (e.g. some code has already been tested on the local high-performance computer; there is local knowledge about the module in question). However, certain combinations may appear more scientifically consistent than others. For example, it may be judged better to use ocean and atmospheric modules that implement a similar numerical scheme. In other words, scientists usually make judgements when choosing to combine modules and parameterisations.

Third, and perhaps more crucially, there is an important phase of work between the decision to obtain a GCM by combining modules, and the final step of producing simulations with the GCM to contribute to the MME ensemble. This phase includes testing, bug tracking, and tuning. It routinely involves a team of scientists and technicians in a process that requires frequent decisions, until the GCM is judged to be ready for producing the experiments that will be lodged in the CMIP database. This process shows that the team of scientists takes ownership of that specific combination of the model, and, through the testing and tuning phases, injects information, which contributes to bringing the present-day simulation into a state that they judge acceptable. This injected information is, for instance, about which level of tuning is tolerable, which compiler optimisations are acceptable, how much testing should be made, and which aspects of the climate system should be taken care of. This information is based on choices that partly reflect the identity of the developer team, and partly pertain to epistemic values shared across the community of modellers.

This point of view outlines the fact that an MME member is more than a member of a population obtained by the systematic collection of possible assemblages. The testing, bug tracking and tuning process confer upon it a specific status owing to experts taking ownership and responsibility of the version of the model that they release.

Now that we have introduced two interpretations of what GCMs are and how they should be sampled, we want to argue that both interpretations are complementary for reflecting on ways to design better MMEs.

## 6. Complementary interpretations

An important distinction between the two interpretations is that, while the second interpretation is more descriptively realistic with regard to the social dynamics of climate research and modelling, the first interpretation offers merely a synchronic view of model building. In the first interpretation, models are seen as mathematical representations with no consideration of the way they are actually built by scientists in practice.

That said, as we want to argue, both interpretations are complementary from a philosophical point of view. The two interpretations involve distinct philosophical domains:

On one hand, the first interpretation appeals to the part of philosophy of science that studies mathematical models as representations of target phenomena and discusses the rational norms and principles under which these representations, albeit idealised, produce genuine knowledge, and reflects on the current methodologies for validating them.

On the other hand, the second interpretation pertains to the part of social epistemology that, given that knowledge is produced collectively by agents, studies the division of scientific labour. This part examines, for instance, under which conditions the collaborative practices in scientific research, based on epistemic dependence and trust between knowers, can legitimately produce knowledge. The second interpretation pertains also to the part of social epistemology that studies how the aggregation of judgements can be normatively justified.

The complexity of the problem is such that we cannot prefer one interpretation to the other. Both are required to study how MMEs can be better designed than they are today. In the end, we contend that both interpretations are relevant and offer complementary insights. We will show that they act as complementary lenses that could help in recognising, addressing and conceptually framing the various aspects of the issues raised by MMEs.

## 6.1. Formalisation of the adequate properties of MMEs

The question arises whether the given interpretations solve the double challenge of conceptualisation (Section 3). Here we will argue that the first interpretation gives half the answer since it is the appropriate framework for conceptualising the norms of systematicity, comprehensiveness and independence that MMEs need to satisfy. This is an important virtue of the first interpretation over the second.

Let us assume that we have defined a set of modules and parameterisations that delineates formally a countable set of possibilities; we will immediately discuss this assumption (6.2). This set of possibilities delimits in turn the domain in which inferences are being made and specific questions can be answered.

Then, to answer a question, for example "what is climate sensitivity?", one will attach, to every combination of modules and parameters, the climate sensitivity obtained by running the model following a well-specified protocol. If we choose to adopt a Bayesian framework, then we need to define a likelihood function, which will effectively rate every individual member of the population, based on a well-defined set of observations. With these assumptions at hand, it is the statistician's job to sample the population such as to deliver an estimate of the probability distribution function of climate sensitivity and to answer the question.

The approach for designing such an ensemble of experiments can then be formalised in the language of experimental design with computer experiments (see, e.g. Santner, Williams and Notz 2003). Classically, the statistician, guided by the climate scientist, will formulate assumptions about the interdependency of different models of the population. One mathematical way of formalising the problem is to attach a large list of numbers to every combination of modules, parameterisations, and parameters. The list will contain integers, which indicate the choice of a particular module or parameterisation, and real numbers, which specify the values of parameters attached to this parameterisation. The set of all the lists can then be seen as a formal, mathematical space, which mirrors the population of models.

The problem is then to sample this space efficiently. To this end, one will typically assume that a given climate model will simulate similar climates if run with two very similar parameter values. One may also assume that two climate models with the same cloud scheme should, *a priori,* have similar climate sensitivities. It is on the basis of these different assumptions (which may be revised *en route*, as the result of simulations is obtained) that statisticians solve the problem of optimal design (given a budget of experiments, selecting which ones will be a priori most informative), and avoid leaving unexplored areas (space-filling design).

In other words, with the first interpretation, the demand of IPCC authors to have systematic and comprehensive sampling can be given a technical meaning, which can be, in principle, applied once the mathematical space hosting the population of models is defined.

By contrast, the second interpretation does not support a statistical formalisation of such properties. This is why MMEs are said to be "ensembles of opportunity". The second interpretation could still be a fruitful framework to discuss the properties of systematicity, comprehensiveness and independence. For instance, Parker (2011, pp. 591-593) draws a parallel between the role of independence between models in our confidence in robust climate-modelling results and the independence between voters in a jury's final choice following generalisations of Condorcet's Jury Theorem. Thus she studies the independence between models through the lens of the theories of decision and social choice that are more naturally connected with the second interpretation.

## 6.2. Specification of the space of model structures

The question still remains on how to specify and circumscribe the space of model structures. The first interpretation is blind to the way scientists create modules and parameterisations, which involves creative steps and judgements as model output and new data become available. On the other hand, this view is compatible with a very large collection of model structures. In principle, we could imagine a list of all possible modules and parameterisations. But the approach just described in 6.1 would be unworkable given the computation cost of the GCMs.

In fact, acknowledged by Murphy et al., (2007), "It is not clear how to define a space of possible model configurations of which [today's multimodel ensemble] members are a sample" (Parker 2010b, 2011). What is the criterion of selection? Should we "identify a collection of plausibly adequate structures among which we can expect to find at least one that is actually adequate" as Parker suggests (2010b, p. 991)?

We would rather suggest selecting, among the available model structures, the ones based on contradictory, yet complementary, representations that presumably might not converge in their projections. For example, if the community finds equally valid the finite difference, the spectral, and the finite element schemes, it may require the MME to cover these three numerical schemes. If it finds no reason *a priori* to favour the sea ice module over another, it could also include both of them. If it disapproves that many GCMs in an MME use the same coupler, it may ask to use alternative formulations.

However, today, the existing modules and parameterisations might be considered as but a small fraction of the plausible ones. In this sense, one may require scientists not to overlook manners of representing the different aspects of the climate system in the future. The pragmatic answer to this objection is to consider that the existing modules and parameterisations may be enough because they represent our current state of best knowledge about the climate system that, in turn, can represent a legitimate Bayesian prior. Our justification stands in the way confidence in GCMs is built in practice. As we will show, confidence in GCMs cannot be gained in combinations of simply plugged modules (6.3) but is actually found within the social and historical processes of modelling that the second interpretation mirrors (6.4).

## 6.3. Why the modular view of GCMs fails

The first interpretation offers a modular view in which GCMs can be generated by "plugging and playing" with modules and parameterisations. But we will argue that one cannot be confident *a priori* that any resulting combination will produce an acceptable GCM. There are two reasons: the rules that define what a good model is, and how it should be calibrated, are not and cannot be decided unambiguously; and modularity is fuzzy.

It is impossible in practice to decide everything in advance so that GCMs can then be constituted automatically. Here is the main argument. Sampling modules and parameterisations in an automated fashion with the help of an algorithmic system would *in principle* generate a tremendous base of possible GCMs, from which an MME with the expected properties could be optimally designed. But practice has shown that the mere assemblage of plausible modules with plausible parameterisations does not necessarily produce an acceptable GCM; an acceptable model is able, by definition, to provide reliable information about the dynamics and evolution of the climate. Consequently, the rejection of some combinations of modules and parameterisations is at some point necessary. This cannot be performed by the algorithmic system, since the rules of rejection cannot all be set *a priori*: it has to be performed *a posteriori* by scientists instead, thus involving expert judgements.

The main reason why the assemblage of modules and parameterisations does not always lead to an acceptable GCM is because coupling large non-linear dynamic modules can result in numerical instabilities and deviations from the "true" climate[4]. Thus, some combinations of modules and parameterisations inevitably turn wrong, but such effects are difficult to anticipate. For instance, simulations with a GCM freshly assembled from reputable ocean and atmosphere modules often produce disappointing results. In other words, the set of *acceptable* GCMs is a subset of the space generated by the combination of acceptable parameterisations.

Furthermore, which criteria need to be retained in order to decide which GCM is acceptable is a modelling decision. Such a decision cannot be made *a priori* because it depends on the observations at hand and the criteria used to assess whether the observations are sufficiently well reproduced. As no model can reproduce all observations, these criteria can legitimately depend on the questions addressed with the model.

If GCM acceptance criteria were defined in advance, assuming a reference set of observations and a standardised metric quantifying the distance between the simulation and the observations, then, one could imagine some algorithmic procedure for sampling the set of *acceptable* GCMs. In practice this would, however, require consensus among modellers about which observations to use, which distance metric to adopt, and which experiments to run with the GCM candidate to generate the test data needed for deciding its acceptability. Such an approach would be transparent, but it would also be restrictive, because it would impose one view about what makes a GCM acceptable. Therefore, it may miss its objective as a "comprehensive" assessment of climate change uncertainties.

In other words, expert judgements reduce and constrain the range of uncertainty. In this way, the range of uncertainty cannot be seen as merely a space of possibilities generated by a set of parameterisations; it is constrained by expert judgements that determine what is a valid model, considering both the design and the output of this model.

We can now further de-idealise the modular view: as Lenhard and Winsberg argue (2010), modularity in GCMs is actually fuzzy. The reason is precisely that GCMs are products of their own contingent histories. As a result, modules are not autonomous nor interchangeable, and their contributions do not just add up linearly. It follows that in no way can they provide acceptable GCMs just by being plugged to each other in an automated fashion.

During the runtime of the simulation, fuzzy modularity is due to that modules interact continuously to each other, exchanging data: "data are continuously exchanged between all modules during the runtime of the simulation. 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." (Lenhard and Winsberg 2010, p. 256). An important consequence is that the "net-effect" of a module or a parameterisation can be tested only by the overall outcome of all the modules and all the parameterisations that compose the model (p. 256).

[4] The first generation of models associating an atmosphere module with an ocean module (e.g. Cubash et al. 1992) included an artificial flux, called, "freshwater-flux correction" calibrated to stabilise the ocean circulation into an acceptable state. Current models no longer include a freshwater flux correction, but tuning an ocean-atmosphere model such that the ocean behaves well remains a delicate exercise.

In model development, fuzzy modularity is due to the generatively entrenched nature of methodological choices made in climate modeling. Lenhard and Winsberg (2010) — and also Winsberg (2012, p. 127-129) and Parker and Winsberg (2018) — claim that the historical nature of climate model optimisation can be grasped by the concept of generative entrenchment; the latter has been introduced by Wimsatt (2007) for characterising adaptive design functions. On this view, model optimisation is considered as a layered process that consists in sequentially adding and assembling modules by adapting them to each other with the final aim of increasing the overall performance of the model. This can be seen as an evolutionary process where modules are adjusted, on the basis of what is there already, following pragmatic software-engineering measures (Lenhard and Winsberg 2010, p. 257).

To sum up, the modular view of GCMs falls short in accounting for important modelling limitations: "plugging and playing" with modules and parameterisations would not automatically yield acceptable GCMs, and, in practice, they are not interchangeable entities since modularity in GCMs is fuzzy.

### 6.4. Confidence in models generated by the social processes

Unlike the first interpretation, the second interpretation leads us to recognise that the social and historical processes underlying model assessment warrant the reliability of models. Such processes generate more confidence than the mere selection among the combinations of the possible modules and parameterisations.

Consider HadGEM3, the flagship climate modelled used by the UK Met Office Hadley Centre. On its presentation sheet, it is presented as "the third generation of HadGEM configurations [that] includes the NEMO ocean model and CICE sea-ice model components" (USGCRP 2019). Attention is thus drawn on the modules (NEMO and CICE) but the presentation sheet also focuses on its lineage: this is the third generation of a family, which itself is a successor to the HadCM family.

An important warrant of GCMs is therefore that model assessment is based on a history of tests and epistemic decisions based on expert judgements. Models have a quality that is historically gained. While, following the first interpretation, HadGEM3 would be considered as just one model among many other models, it has actually withstood various rigorous "pass or fail" tests. This partially explains why it has a particular value and brings more confidence in the eyes of climate scientists.

The lineage is also an important warrant of GCMs in that it expresses an in-house tradition. In the HadGEM3 example, one can see that the UK Met Office has coordinated the efforts of its employees to deliver a climate model satisfying what this institution considers to be good practices in the development of a climate model. This includes quality assurance policies, coding standards, the practice of tuning (selecting a best set of parameters) and in-house policy in response to different events such as the discovery of bugs or the availability of new observations. At the Hadley Center, climate scientists autonomously exert expert judgement in the development of parameterisations, in the use of observations, in the management of international collaborations, but the cycle of model development and release is governed strategically.

Following that perspective, an institute such as the Hadley Centre can be seen as a *model-generating agent*: it produces a model that, it considers, satisfies acceptable criteria of internal and external consistency.

### 6.5. Significance of values in climate models

Because expert judgements are situated and contain the standpoints of their owners, sampling expert judgements is also about sampling non-epistemic values and other research contingencies that may influence scientists. Thus, a significant contribution that the second interpretation offers is to enable us to consider the influence of non-epistemic values in climate modelling.

Non-epistemic values are contextual values of a social, economic, political, cultural or ethical kind. They determine purposes and priorities in representing some particular aspect of the climate system (Intemann 2015; Parker and Winsberg 2018). They therefore depend on the modelling centre, and thus, models may differ from one centre to another depending on the ethical values the research group respectively has. As Intemann (2015) illustrates, if one is ethically concerned with how to adapt to worst-case scenarios, models should be built so as to capture extreme weather events. If one feels moral obligations to protect future generations, models should be designed to support longer runs (e.g. 200 years versus 100 years). On the other hand, scientists may prioritise certain kinds of representation given the regional interests in knowing the climate future of their home country. Thus, regional concerns about climate change may also create diversity. One could think that, for example, Indian centres would have more interests in the monsoon, European centres in storms, Canadian or Russian centres in permafrost and sea ice, Dutch centres in rising waters.

Other research contingencies may also be at stake. The choices of representations may depend on the peculiar specialities of research centres, and specific competencies and favourite subjects of their members.

Recent philosophical attention has been paid to whether these non-epistemic values contravene the "value-free" ideal of climate science and lead to "wishful thinking" (Intemann 2015). "Wishful thinking" is "based on what we wish the model would predict rather than decisions about what will make the model more accurate or accountable to the "way the world really is" (Brown 2013)" (Intemann 2015, p. 221).

Following this argument, one could be worried about a form of social bias among experts which makes them oversensitive to certain aspects of the problem at the expense of others, in a way that produces undesirable effects on their collective judgement. To cite one concrete example, James et al. (2018) explain how the lack of presence of African scientists in climate model development teams affects the model development and validation process in a way that impoverishes climate scientists' assessment of the future climate in Africa. That said, some values may be desirable (Intemann 2015; Schroeder 2017), and diversity of standpoints and associated values may be important in an intergovernmental context in helping to avoid undesirable biases.

In the second interpretation, building an MME is the process of sampling a population of expert judgements from research institutes. The advantage of this account is that it offers a perspective for criticising the representativeness of the resulting sample and maybe, when desirable, to correct it, e.g., in excluding a model because some purposes and priorities are insufficient or lacking, such as the goal of simulating African regions. These considerations may be surprising from the mere perspective of the philosophy of science dealing with scientific representations. Nevertheless, they are consistent with the dominant literature on values in science.

## 7. Practical upshot

Each interpretation offers a partial view for designing better MMEs. Both are complementary, and adopting them simultaneously can offer a more comprehensive picture of the epistemological issues raised by ensemble optimization. Then, once we recognise this, it is worth questioning how differently scientists should operate. What are the consequences of looking at MMEs from both perspectives for the practice of ensemble modelling?

While the first interpretation seems widely (implicitly) shared within the scientific community, the second interpretation is certainly a less common way of considering MMEs. From this new perspective, scientists can see the MME as an elicitation of experts. The GCMs in the MME could then be treated as opinions of experts. Hence, one could in principle defer the problem of generating a population of models by sampling a population of institutes, which would all deliver their best model simulations. The simulations could even be associated with a measure of uncertainty, which could be obtained by sampling what they consider to be reasonable parameter ranges as in Murphy et al. (2004). Given that the actual population of research centres actually exists, the models may be collected comprehensively.

One can indeed find, in the model development process, an analogy with the build-up of expertise on a specific subject. Academic knowledge is present in the basic equations concerning fluid dynamics and their parameterisations, and “experience” is encoded throughout the long process of parameterisation choices, model adjustment and tuning, which is nurtured by observations and the personal experience of individual researchers. Models therefore answer questions on the basis of encoded academic knowledge and experience. Rougier and Goldstein endorse this point of view when they write that “Insofar as the simulator is the outcome of many judgements, its distribution is subjective” (2014, p. 105).

Concretely, if a coordination of MMEs were to be established worldwide, at a collective level, it could potentially address two important criticisms behind the argument that MMEs are ensembles of opportunity, somewhat differently as they are today. The analogy with expert elicitation could help us to spell out the terms of this coordination.

First of all, one should see that there is not always reason to worry about the fact that different models share codes. For some subject matters, it is not considered to be a problem that different experts share knowledge and factsheets. It is indeed important to distinguish between different types of dependencies among models and that not all equally undermine the trust we can place in consensus projections. Some dependencies, for example, may simply reflect a prior consensus on the fundamental physical equations governing the climate system. By contrast, model idealizations, parameterisations and calibrations are often “in-house” traditional recipes of centres whose transmission from a model to another may be epistemically harmful. Rather than speaking of independence, which can easily be taken as a criterion involving information theory, one could prefer to value the concept of scientific autonomy, which emphasises the ability of scientific teams to take their own decisions, based on their own epistemic and non-epistemic values.

To pursue the analogy, one could be worried about a form of conformism or common biases among experts. In particular, it seems that scientists favour models whose projections are more in the centre of the “consensus range” than models that provide projections far outside this range. As noticed by Knutti, “Although this is hard to confirm or reject, there may even be an element of ‘social anchoring’ and a tendency towards consensus” (Knutti 2010, p. 397). This worry is getting particularly serious if one aims to design a set of complementary GCMs (rather than a mere collection of best guesses). Conformism can be found more particularly in the practice of tuning. Examples include tuning a model such that its climate sensitivity falls within the range of other models (i.e. the “consensus range”), using similar targets for calibration (e.g. sea-ice cover, thermohaline circulation depth, monsoon area and intensity), or calibrating a model over similar historical cases. Here, one could imagine that scientists would have to resist to this tendency towards conformism.

Finally, if one adopts the two interpretations simultaneously, sampling models appears to be at the centre of a trade-off between, on one side, a coordination of worldwide model development, that should aim at diversity among modules and parameterisations (following the first interpretation), and, on the other side, scientific autonomy, particularly in the tuning of models that may well be the place of highest risk of conformism (as suggested by the second interpretation).

## 8. Conclusion

Projections of future climate change cannot rely on a single model. It has become common to rely on multiple simulations generated by Multi-Model Ensembles (MMEs), especially to quantify the uncertainty about what would constitute an adequate model structure.

We offered two interpretations of what GCMs are and how MMEs should be designed. In the first interpretation, models are combinations of modules and parameterisations; an MME is obtained by plugging presumably interchangeable modules and parameterisations. In the second interpretation, models are aggregations of expert judgements that result from a history of epistemic decisions made by scientists about the choices of representation; an MME is a sampling of expert judgements from modelling teams.

Modules and parameterisations can be seen either as mathematical structures or as historical expert judgements. These two interpretations therefore instantiate objects of a different nature and involve distinct epistemologies. We nevertheless argued that they may be used in a complementary manner in order to explore ways to design better MMEs. While the first interpretation helps in properly formalising the adequate properties of MMEs, the second accounts for the fact that confidence in model projections also hinges upon trust in the social and historical processes underlying model assessment, and the influence of non-epistemic values in choices of representation. We finally suggest the consequences of viewing MMEs as elicitations of experts. Following Winsberg (2018, last chapter), we think that the recent developments in social epistemology are beneficial to study the objects of climate science. We hope that adopting the two perspectives we have made explicit could help in framing other epistemic issues as well.

## Acknowledgments

We thank Joel Katzav and Charles Pence, as well as the two anonymous reviewers for their precious comments on this paper.
This work was supported by “MOVE-IN Louvain” Incoming Post-doctoral Fellowship, cofunded by the Marie Curie Actions of the European Commission, and by the Swiss National Science Foundation project PP00P1_170460 “The Epistemology of Climate Change”.

## References

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R. and Schmidt, G. A. (2018) Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing, *Earth System Dynamics Discussions*, pp. 1-20.

Annan, J. D. and Hargreaves, J. C. (2017) On the meaning of independence in climate science. *Earth Syst. Dynam.*, 8, 211–224.

Betz, G. (2009) Underdetermination, Model-Ensembles and Surprises: On the Epistemology of Scenario-Analysis in Climatology, *Journal for General Philosophy of Science*, 40(1): 3–21.

Brown, M. J. (2013) Values in Science beyond Underdetermination and Inductive Risk. *Philosophy of Science* 80(5):829-39.

Collins, M., R. Knutti, J. Arblaster, J.-L. Dufresne, T. Fichefet, P. Friedlingstein, X. Gao, W.J. Gutowski, T. Johns, G. Krinner, M. Shongwe, C. Tebaldi, A.J. Weaver and M. Wehner (2013) Long-term Climate Change: Projections, Commitments and Irreversibility. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Cubasch U., K. Hasselmann, H. Höck, E. Maier-Reimer, U. Mikolajewicz, B. D. Santer and R. Sausen (1992), Time-dependent greenhouse warming computations with a coupled ocean-atmosphere model, Climate Dynamics, (8) 55–69 doi:10.1007/bf00209163

Drake J. B., P. W. Jones and G. R. Carr (2005) Overview of the Software Design of the Community Climate System Model, *The International Journal of High Performance Computing Applications*, (19) 177–186.

ECMWF (European Centre for Medium-Range Weather Forecasts). “Modelling and prediction. Land.” Accessed September 22, 2019. https://www.ecmwf.int/en/research/modelling-and-prediction/land

ENES (European Network for Earth System Modelling). “Earth System Models and Modelling groups, EC-EARTH, The EC-Earth ESM, V3.” Last modified September 15, 2015. https://portal.enes.org/models/earthsystem-models/ec-earth-1/ec-earth

Flato, G., J. Marotzke, B. Abiodun, P. Braconnot, S.C. Chou, W. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring, C. Forest, P. Gleckler, E. Guilyardi, C. Jakob, V. Kattsov, C. Reason and M. Rummukainen (2013) Evaluation of Climate Models. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Frigg, R., Stainforth, D. A. and Smith, L. A. (2013) The Myopia of Imperfect Climate Models: The Case of UKCP09, *Philosophy of Science* 80(5), 886–897

Frigg, R., Stainforth, D A. and Smith, L. A. (2015) An Assessment of the Foundational Assumptions in High-Resolution Climate Projections: The Case of UKCP09. *Synthese*, Volume 192, Issue 12, pp 3979–4008.

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

Herger, N., Angélil, O., Abramowitz, G., Donat, M. et al. (2018) Calibrating climate model ensembles for assessing extremes in a changing climate. *Journal of Geophysical Research: Atmospheres*, 123, 5988–6004.

Intemann, K. (2015) Distinguishing between legitimate and illegitimate values in climate modeling. *European Journal for Philosophy of Science*, 5, 217–232.

James, R.,Washington, R., Abiodun, B., Kay, G., Mutemi, J., Pokam, W. Hart, N., Artan, G. and Senior, C. (2018) Evaluating climate models with an African lens. *Bulletin American Meteorological Society*.

Jöckel P., R. Sander, A. Kerkweg, H. Tost and J. Lelieveld (2005) Technical Note: The Modular Earth Submodel System (MESSy)—a new approach towards Earth System Modeling, Atmospheric Chemistry and Physics, (5) 433–444.

Kalnay E., M. Kanamitsu, J. Pfaendtner, J. Sela, J. Stackpole, J. Tuccillo, M. Suarez, L. Umscheid and D. Williamson (1989) Rules for Interchange of Physical Parameterizations, *Bulletin of the American Meteorological Society*, (70) 620–622.

Katzav, J., Dijkstra, H. A., (Jos) de Laat, A.T.J. (2012) Assessing climate model projections: State of the art and philosophical reflections. *Studies in History and Philosophy of Modern Physics* 43:258–276.

Katzav, J. and Parker, W. S. (2015) The future of climate modeling. *Climatic Change* 132:475–487.

Kirtman, B., S.B. Power, J.A. Adedoyin, G.J. Boer, R. Bojariu, I. Camilloni, F.J. Doblas-Reyes, A.M. Fiore, M. Kimoto, G.A. Meehl, M. Prather, A. Sarr, C. Sch.r, R. Sutton, G.J. van Oldenborgh, G. Vecchi and H.J. Wang, (2013) Near-term Climate Change: Projections and Predictability. In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

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

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J. and Meehl, G. A. (2010) Challenges in combining projections from multiple climate models. *J. Clim.*, 23, 2739–2758.

Knutti, R., Masson, D. and Gettelman, A. (2013) Climate model genealogy: Generation CMIP5 and how we got there, *Geophys. Res. Lett.*, 40, 1194–1199, doi:10.1002/grl.50256.

Lenhard, J., and Winsberg, E. (2010). Holism, Entrenchment, and the Future of Climate Model Pluralism. *Studies in History and Philosophy of Science* Part B, 41(3):253–262.

Meehl, G. A., Stocker, T. F., Collins, W. D., Friedlingstein, P., Gaye, A. T., Gregory, J. M., et al. (2007) *Global climate projections*. In Solomon et al. (Eds.) (pp. 747–845).

Meinshausen M., N. Meinshausen, W. Hare, S. C. B. Raper, K. Frieler, R. Knutti, D. J. Frame and M. R. Allen (2009) Greenhouse-gas emission targets for limiting global warming to 2°C, *Nature*, (458) 1158–1162.

Murphy, J. M., Sexton, D. M. H., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., Stainforth, D. A. (2004) Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature* 430, pages 768–772.

Murphy, J. M., Booth, B. B. B., Collins, M, Harris, G. R., Sexton, D. M. H. And Webb, M. J. (2007) A Methodology for Probabilistic Predictions of Regional Climate Change from Perturbed Physics Ensembles. *Philosophical Transactions of the Royal Society A* 365:1993–2028.

NEMO Community ocean model for multifarious space and time scales. "About NEMO." Accessed September 22, 2019. https://www.nemo-ocean.eu/

Palmer T. (2005), Global warming in a nonlinear climate - Can we be sure?, *Europhysics News*, (36) 42-46.

Parker, W. S. (2006) Understanding pluralism in climate modelling, *Foundations of science*, 11, pp. 349-368.

Parker, W. S. (2010a) Predicting Weather and Climate: Uncertainty, Ensembles and Probability. *Studies in History and Philosophy of Modern Physics* 41, 263-272.

Parker, W. S. (2010b) Whose Probabilities? Predicting Climate Change with Ensembles of Models. Proceedings of PSA08. *Philosophy of Science* 77(5), 985-997.

Parker, W. S. (2011) When Climate Models Agree: The Significance of Robust Model Predictions. *Philosophy of Science*, Vol. 78, No. 4, pp. 579-600

Parker, W. S. (2013) Ensemble modeling, uncertainty and robust predictions. *WIREs Climate Change*, 4:213–223.

Parker, W. S. (2018) Climate Science, *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/sum2018/entries/climate-science/.

Parker, W. S. and Winsberg, E. (2018) Values and evidence: how models make a difference. *European Journal for Philosophy of Science* 8(1): 125-142.

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

Rougier, J. (2007) Probabilistic inference for future climate using an ensemble of climate model evaluations, *Climatic Change*, (81) 247-264.

Rougier J. and Goldstein, M. (2014) Climate Simulators and Climate Projections, *Annual Review of Statistics and Its Application*, (1) 103–123.

Sanderson, B. M., Knutti, R., and Caldwell, P. (2015a) Addressing interdependence in a multimodel ensemble by interpolation of model properties*, J. Climate*, 28, 5150–5170.

Sanderson, B. M., Knutti, R., and Caldwell, P. (2015b) A representative democracy to reduce interdependence in a multimodel ensemble, *J. Climate*, 28(13), 5171–5194.

Santner T. J., B. J. Williams and W. I. Notz (2003) The Design and Analysis of Computer Experiments, Springer Texts in Mathematics, Springer.

Schroeder, S. A. (2017) Using Democratic Values in Science: An Objection and (Partial) Response, *Philosophy of Science* 84 (5): 1044-1054

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

USGCRP (U.S. Global Change Research Program). Global Change Information System. “Hadley Centre Global Environment Model version 3.” Accessed September 22, 2019. https://data.globalchange.gov/model/hadgem3

Werndl, C. (2016) On Defining Climate and Climate Change. *The British Journal for the Philosophy of Science* 67 (2), 337-364.

Wimsatt, W. (2007). *Re-engineering philosophy for limited beings*. Piecewise approximations to reality. Cambridge, MA and London, England: Harvard University Press.

Winsberg, E. (2012) Values and Uncertainties in the Predictions of Global Climate Models. *Kennedy Institute of Ethics Journal* 22 (2), 111-137.

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