Climate Models and Robustness Analysis – Part I: Core Concepts and Premises

Margherita Harris

London School of Economics and Political Science

Roman Frigg

London School of Economics and Political Science

Forthcoming in Gianfranco Pellegrino and Marcello Di Paola (eds.):
Handbook of the Philosophy of Climate Change. Springer.

Abstract

Robustness analysis (RA) is the prescription to consider a diverse range of evidence and only regard a hypothesis as well-supported if all the evidence agrees on it. In contexts like climate science, the evidence in support of a hypothesis often comes in the form of model results. This leads to model-based RA (MBRA), whose core notion is that a hypothesis ought to be regarded as well-supported on grounds that a sufficiently diverse set of models agrees on the hypothesis. This chapter, which is the first part of a two-part review of MBRA, begins by providing a detailed statement of the general structure of MBRA. This statement will make visible the various parts of MBRA and will structure our discussion in the remainder of the chapter. We explicate the core concepts of independence and agreement, and we discuss what they mean in the context of climate modelling. Our statement shows that MBRA is based on three premises, which concern robust properties, common structures, and so-called robust theorems. We analyse what these involve and what problems they raise in the context of climate science. In the next chapter, which is the second part of the review, we analyse how the conclusions of MBRA can be justified.

Keywords

Robustness analysis, model ensemble, climate models, model agreement, independence, truth plus error hypothesis, a priori approaches, a posteriori approaches, common structure, decompositional strategy, scrutability, robust theorems

Table of Contents

1. Introduction

In his discussion of justification, Wittgenstein (1953, §265) tells the story of an imaginary fool who buys several copies of the morning paper to assure himself that what it said was true. The fool's instinct to check his information against further sources is laudable, and yet the fool is obviously mistaken because, as Wittgenstein insists, "justification consists in appealing to something independent". To assure himself that a particular story in the morning paper was true, the fool should have diversified his evidence. In the first instance he should have checked other newspapers, making sure they were from different publishers and positioned on different parts of the political spectrum. In a next step, he should have checked news outlets other than newspapers, and he should have consulted the direct communications of news agencies. If the reports from these diverse sources all essentially said the same thing, then he would have been justified in regarding it as true. *In nuce*, *robustness analysis* (RA) is the prescription to avoid the fool's mistake by considering a diverse range of evidence and only regard a hypothesis as well-supported if all the evidence agrees on it.

The maxim to diversify evidence wherever possible is a fixture of the scientific method (Staley 2004), and RA is therefore often motivated by appeal to experimental practice. Schupbach, for instance, introduces RA with the example of Brownian motion (2018, 275-77). Brownian motion is the random motion of particles suspended in a medium. It is named after botanist Robert Brown, who first described the phenomenon in 1827 when observing pollen particles suspended in water. We nowadays regard Brownian motion as a general feature of all matter, which is due to the particles being pushed around when colliding with the molecules of the medium. To establish this conclusion, it wasn't enough to look at Brown's pollen particles. Brown himself repeated the experiment first with several different kinds of pollens and then also with various inorganic materials. Physicists then continued to vary circumstances by using different containers, different media, different light to observe the particles, and so on. The phenomenon was regarded as real only once it was detected in all these cases – that is, once it was shown to be *robust* across a wide range of physical circumstances and means of observation. Robustness is intended to rule out that the phenomenon only occurs under specific circumstances, or that it is an artefact of our means of observation. (For discussions of this case that differ from Schupbach's see, e.g., Mayo (1986), Psillos (2011), Chalmers (2011) and Hudson (2020)).

In the case of Brownian motion, the evidence for the phenomenon is experimental. Such cases are important, but they are no longer the only game in town. In numerous contemporary scientific contexts, the evidence in support of a hypothesis comes from scientific models, and, following the imperatives of RA, a hypothesis is regarded as well-supported on grounds that a sufficiently diverse set of models (in this context referred to as a *model ensemble*) agrees on it. We call this line of reasoning *model-based RA* (MBRA). Examples of MBRAs are readily to hand. It has been applied to justify results in cosmology (Gueguen 2020), economics (Kuorikoski et al. 2010), ecology (Weisberg 2006), population genetics (Plutynski 2006), environmental risk analysis (Sprenger 2012), and, indeed, climate science. Leduc *et al*. note that "[a]greements between climate change projections from several models are often interpreted as predictors of confidence" (2016, 8302) and Pirtle, Meyer, and Hamilton review the literature on climate modelling and report that "a rough survey of the contents of six leading climate journals since 1990" yields "118 articles in which the authors relied on the concept of agreement between models to inspire confidence in their results" (Pirtle et al. 2010, 353).

In general terms, models are representations of a target system, in our case the world's climate (see Frigg and Hartmann (2020) for a general discussion of scientific models, and Frigg, Thompson, and Werndl (2015) for a discussion of climate models). But systems like the world's climate are far too complex to be represented fully and truthfully in a model, which is why models always offer representations that are simplified, abstracted, idealised, and distorted in one way or another. Yet, proponents of MBRA submit that model-agreement is epistemically significant and that results that are robust across a sufficiently diverse set of models should be regarded as well-supported. Tebaldi *et al*. (2011, 1) are explicit about this when, in the context of a discussion of future climate projections, they note that "[t]he idea is that if multiple models, based on different but plausible assumptions, simplifications and parameterizations, agree on a result, we have higher confidence than if the result is based on a single model, or if models disagree on the result". So even though each model in an ensemble has its shortcomings, the fact that they all agree on a conclusion is taken to be boost our confidence in it. In Levins' by now proverbial formulation, the idea is that "our truth is the intersection of independent lies" (Levins 1966).

But why is this? That is, why should the fact that a number of models, each deficient in its own ways, agree, provide a warrant for regarding the agreed-upon proposition to be well-supported? This is the core question that a philosophical reflection on MBRA has to answer. The aim of this two-part review is to address this question, with a particular focus on climate models. This chapter, which is Part I of the review, begins by providing a detailed statement of the general structure of MBRA (Section 2). This statement will make visible the various parts of MBRA and will structure our discussion in the remainder of the chapter. We explicate the core concepts of independence and agreement (Section 3), and discuss what they mean in the context of climate modelling. Our statement in Section 2 shows that MBRA is based on three premises, which concern robust properties, common structures, and so-called robust theorems. We analyse what these involve and what problems they raise in the context of climate science (Section 4). We end with a brief conclusion (Section 5). In the next chapter, which is Part II of the review, we analyse how the conclusion of MBRA can be justified.

2. Model-Based RA

MBRA is a form of inference, and to understand how MBRA works, we have to understand what its inference pattern is. The most influential recent analysis of MBRA is Weisberg's (2006), who introduces it with the example of the Lotka-Volterra model of predator prey interaction. It is therefore helpful to briefly review Weisberg's discussion to get to a general formulation of MBRA, and then note that this formulation equally applies to climate models.

Consider the fish in the Adriatic Sea. They can be sorted into a population of prey and a population of predators, which have sizes V and P, respectively. Since predators eat prey, the time evolution of the population sizes are related. Their relation can be represented in a model that is based on two coupled first-order differential equations, the so-called Lotka-Volterra equations: $\dot{V} = rV - (aV)P$ and $\dot{P} = b(aV)P - mP$, where *r* is the birth rate of the prey population, *m* is the death rate of the predator population, and *a* and *b* are linear response parameters. This is the Lotka-Volterra model. An analysis of the model shows that the populations in it have the so-called *Volterra property*, namely that a general biocide (the uniform reduction of all species) favours prey in the sense that after introducing the biocide the number of preys grows and the number of predators shrinks (Weisberg 2006, 735). Further analysis shows that the model has a feature known as *negative coupling*, namely that increasing the number of predators decreases the number of preys and increasing the number of preys increases the abundance of predators. Finally, one can show that in this model the following holds: if the system is negatively coupled, then it has the Volterra property.

The question now is whether these features of the Lotka-Volterra model are also present in real populations of fish in the Adriatic Sea. That is, can we assert that the predator and prey populations in the Adriatic Sea have the Volterra property and negative coupling, and that in that population the former is brought about by the latter? This is where MBRA enters the scene. Weisberg (2006), and later Weisberg and Reisman (2008), consider alternative models: a family of models whose equations are the same as in the original Lotka-Volterra model but where the parameters have different values; then a model whose equations are the Lotka-Volterra equations with a density term added, and finally an individual-based model which represents individual organisms and their behaviours rather than describing the populations at an aggregate level. The result of this modelling exercise is a model ensemble consisting of four models (setting aside the protracted but ultimately inconsequential question of how one counts models that differ only in their parameter values). One can then show that all these alternative models still have the Volterra property and negative coupling, and hence that in the class of models that make up the ensemble the following holds: if the system is negatively coupled, it has the Volterra Property. Hence, these features are *robust* in the ensemble in the sense that all models in the ensemble have them. This is Weisberg's motivation for calling the Volterra property the *robust property R*; for saying that negative coupling is the *common structure S* of the models in the ensemble, and for dubbing the proposition that (ceteris paribus) *S* brings about *R* a *robust theorem*.

The punchline of MBRA is that the fact that these features hold in all models in the ensemble is taken to warrant the belief that all (or at the very least some) of those features also hold in the *target population*, the fish in the Adriatic Sea, because the models in the ensemble are sufficiently independent. In this instance, according to Weisberg, MBRA is taken to establish the truth of the proposition "under conditions *C*, negative coupling brings about the Volterra property". The proviso "under conditions *C*" is added to make explicit the fact that this regularity is supposed to hold only *ceteris paribus*. If, for instance, the prey population catches a disease that greatly reduces the number of preys, then the robust theorem may cease to hold.

The general inference pattern of MBRA can be summarised as follows (Frigg 2022, Sec 15.3):

> Assume we have a model ensemble Ω consisting of sufficiently independent models that all represent target system *T*.
>
> **Step 1: Robust property**
> Premise 1 – *Ensemble-Robust-Property*: all models in Ω have property *R*. This property is called the "robust property".
> Conclusion 1 – *Target-Robust-Property*: *T* has *R*.
>
> **Step 2: Common structure**
> Premise 2 – *Ensemble-Common-Structure*: all models in Ω have structure *S*. This structure is called the "common structure".
> Conclusion 2 – *Target-Common-Structure*: *T* has structure *S*.
>
> **Step 3: Robust Theorem**

> Premise 3 – *Model-Robustness-Theorem*: in all models in Ω it is the case that, under conditions *C*, *S* brings about *R*. This proposition is called the "robust theorem".
> Conclusion 3 – *Target-Robustness-Theorem*: under conditions *C*, *S* brings about *R* in *T.*

Lloyd (2010, Sec. 5) discusses MBRA and points out that it is a suitable way of looking at climate models. She considers an ensemble of 14 climate models and then notes that the property of having increasing global mean surface temperature is robust in the ensemble because it is exhibited by all models. This is the robust property *R*. The models also all have a common core which consists of physical principles which describe the interaction of increasing greenhouse gas emissions with the earth's energy balance. This is the common structure *S*. Furthermore, an investigation of the models shows that an increasing concentration of greenhouse gases brings about increasing global mean surface temperatures. This is the robust theorem. Lloyd does not comment on the conditions *C*, but presumably these are conditions that rule out major interference from other factors (such as a large-scale volcanic eruption).

Let us now add some qualifications to the above general scheme. First, at this point, nothing is assumed about the ensemble Ω beyond the fact that its models are sufficiently independent. Specifically, it's not assumed that Ω is large or complete (in some relevant sense), or that the models in it are well-confirmed. We note, however, that there is an interesting connection between the choice of Ω and the formulation of the conditions *C* in the robust theorem, because the conditions under which the theorem holds will depend on what is covered by the models in Ω. In essence, the larger the spectrum of scenarios that are covered by the models in Ω, the less restrictive *C* will be. This is a point that we think is often not sufficiently stressed in discussions surrounding the epistemic import of robustness analysis, especially when it comes to the conclusion in Step 3. Indeed, it is often assumed that one can talk about the conclusion of Step 3 without any reference to the conditions *C* under which the theorem is supposed to hold - see e.g. Kuorikoski *et al.* (2010) and Schupbach (2018). But without such clarification the empirical content of the robust theorem remains unspecified.

Second, rather than construing the argument as one that establishes conclusions in a categorical way, one can see it as increasing our confidence in the propositions in the conclusions. Indeed, this is how the core idea of MBRA has been summarised in the quotes by Leduc *et al.* and Tebaldi *et al.* in the previous section, and it is also how Baumberger *et al.* (2017) and Parker (2011) formulate the approach. On this reading, the conclusions in the three steps don't make the categorical statement that *T* has *R*, but instead establish that a statement is well supported (or at least that it is better supported than it was prior to having carried out the MBRA). The conclusion of Step 1, for instance, could then be the statement that we have increased confidence that *T* has *R*. The issues we discuss in what follows are independent of whether conclusions are formulated categorically or in terms of increased confidence. We use the categorical formulation for simplicity; readers can always substitute the confidence formulation if this is their preference.

Third, Step 3 ensures that there is a genuine connection between *S* and *R*, and that it is not just a coincidence that systems that have *S* also have *R*. However, we recognise that the notion "bringing about" is vague, and deliberately so. What notion exactly is appealed to will depend both on the context (negative coupling will bring about the Volterra property in a way that is different from how the exposure of a human body to high levels of radiation brings about cancer) and on one's philosophical commitments (such as one's views on causation and laws of nature). At a general level, MBRA need not commit to a particular notion of bringing about, and a relevant notion can be introduced in specific cases.

Fourth, the above scheme, which involves three steps, is what one might call a complete template of MBRA. Depending on the problem at hand and one's research question, only parts of the scheme may be of interest. For instance, one may only want to establish that a system has property *R* while not being interested in identifying a structure *S* and connecting that structure to *R* in a robust theorem. If so, then one only carries out Step 1. An example of such case is a study by Seager *et al*. (2007) in which they aim to establish through model agreement that the Southwestern United States will experience increased aridity and drought over the next one hundred years. In another scenario it may be the case that it is known empirically that *T* has *R*, which makes Step 1 obsolete and a robustness analysis will focus on Steps 2 and 3. An example of such case is Vicedo-Cabrera *et al*.'s (2021) attribution study which used pairs of factual–counterfactual ensemble runs of daily mean temperature between 1991 and 2018 from ten general circulation models, to conclude that 37% of the heat wave deaths across 43 countries from 1991–2018 were attributable to human induced climate change. Here the heat wave deaths are taken as given and one tries to show that they are (partly) attributable to the common structure of climate physics plus increases in greenhouse gases. In yet other cases one may just be interested in Step 3 and aim to establish the connection between *S* and *R*, while leaving it open whether a particular system has either of the properties. An example of such case is Kuorikoski *et al*.'s (2010) application of MBRA to a family models in geographical economics in which they aim to establish the following robust theorem: "*Ceteris paribus*, if firms benefit from economies of scale, goods are costly to transport, and there are both immobile and mobile activities, spatial agglomeration occurs when economies of scale are high, market power is strong, and transportation costs are low" (2010, 557).

MBRA raises three issues, and these will be the subject matter of our two-part review. The first issue concerns the proper articulation of the basic notions in MBRA. In particular, what does it mean for models in an ensemble to be sufficiently independent and what does it mean for models to agree? We discuss that in Section 3.

The second issue concerns how the premises are established. How do we show that all models in the ensemble have the property *R*, and all models have a common structure *S*, and that *S* brings about *R*? For MBRA to get off the ground, we need to know that the premises are true, or at the very least have evidence to support them. We discuss the problem of how to establish the premises in Section 4.

The third issue is the validity of the inferences drawn. The transition from the premises to the conclusion amounts to a transition from what Smith calls *model-land* (2007, 135) to the real world: the premises make assertions about the model ensemble and the conclusion concerns the target system. It is obvious that none of the inferences are deductively valid: in each step it is possible for the premise to be true while the conclusion is false. What, then, justifies the inference from the premises to the conclusions? This is a thorny issue, and we turn to it in the second part of this two-part review.

3. Articulating Core Concepts

MBRA is based on two core notions: independence and agreement. In this section we look at different articulations of these notions in the context of climate modelling.

3.1 Independence

As we have seen, a crucial ingredient in MBRA is independence. So the first question we have to answer is: what does it mean for models to be independent? There are at least three different interpretations of independence that have been discussed in the climate literature, and the distinction between them is significant in many respects. Broadly they can be characterized as follows:

1. Under the first interpretation, "the assumption of independence is equivalent to the interpretation that each model approximates the real world with some random error" (Knutti et al. 2010, 2745). This is often referred to as the *truth plus error hypothesis* (or the *truth plus error paradigm*).

2. Under the second interpretation, the degree of independence is determined by the amount of divergence between models' outputs independent of observations (Abramowitz and Gupta 2008), or by the degree of correlation of observed model errors (Bishop and Abramowitz 2013; Sanderson et al. 2015). Approaches of this kind are referred to as *a posteriori approaches*.

3. Under the third interpretation, the degree of independence is determined by the degree of shared formulation in the models. Hence, under this conception of independence models are classified "based on the independence of their structure" (Abramowitz 2010). Approaches of this kind are referred to as *a priori approaches.*

Notice that under the first interpretation, independence is not a matter of degrees. In other words, under the first interpretation models are either independent or they are not. By contrast, under the second and third interpretations, models can be more or less independent, and what we are interested in is the extent to which they are. We now discuss each of these interpretations in turn.

Under the first interpretation, what it means for models to be independent is that their errors are independent and identically distributed (typically assumed to be normally distributed with zero mean). As Knutti *et al.* (2010, 2745) note, many Bayesian methods that are used to interpret the results from multi-model ensembles rely on the assumption that the truth plus error hypothesis is true, and according to Leduc *et al.* (2016, 8302) "the truth-plus-error paradigm remains the most widely used technique for processing multimodel ensemble". However, as Annan and Hargreaves (2017) point out, if the truth plus error hypothesis were true, it would have remarkable consequences:

> Although it has not generally been explicitly stated, even a small ensemble of samples drawn from such a distribution would be an incredibly powerful tool. If we could sample models from such a distribution, then we could generate arbitrarily precise statements about the climate, including future climate changes, merely by proceeding with the model-building process indefinitely and taking the ensemble mean. This would obviate the need both for computational advances and also for any additional understanding of how to best simulate the climate system. (2017, 212-13)

As an example of the impact of the error plus truth hypothesis, consider the figure below taken from Knutti *et al.* (2010), showing various probability density functions obtained using a Bayesian method developed by Furrer *et al.* (2007), which relies on the truth plus error assumption. The graph shows that the uncertainty in the true value of the temperature change (i.e. the width of the probability density function) decreases substantially as the number of models included in the ensemble increase from 4 to 21 models.

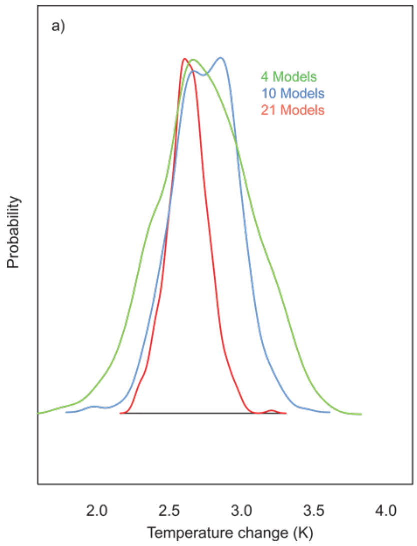

Figure 1: Probability density functions for annual global temperature change (for a particular period and under a specific scenario) obtained by Knutti *et al*. (2010) with Bayesian method developed by Furrer *et al*. (2007), for 4, 10, and 21 models.

Despite the fact that many Bayesian methods that are used to interpret the results derived from multi-model ensembles rely on the truth plus error hypothesis, there are reasons to doubt that it is applicable to actual model ensembles. The problem is that models' errors are often correlated, and that the mean of an ensemble does not converge to the truth when the number of models in an ensemble increases, as one would expect if the error plus truth hypothesis were true. Knutti *et al*. (2010), for instance, show that the errors of the models' results in the CMIP3 are strongly correlated and that the mean of the CMIP3 does not asymptotically converge to observations. But crucially, the knowledge that climate models often share many simplifications, limitations and assumptions should already provide enough of a reason to suspect that this assumption is not appropriate in the first place, as many have noted (see, e.g., Knutti *et al*. (2010) and Bishop and Abramowitz (2013, 886)).

Under the second interpretation, independence is measured in terms of the amount of divergence between models' outputs. Specifically, "the proximity of GCMs results or of their errors is used to quantify a posteriori their interdependencies" (Boé 2018, 2772). A posteriori approaches can be further divided into ones that assume that the level of dependence reflects only the amount of divergence of model outputs independent of observations, and ones that also take agreement of model outputs with observations into account.

An instance of the former is Abramowitz and Gupta's (2008) measure of independence, which assesses independence solely on the divergence of outputs independent of observations: the closer the models' outputs under similar input and initial conditions, the more dependent they are. Annan and Hargreaves (2017, 213) object that "this approach has the potential weakness that models that agree because they are all accurate will be discounted, relative to much worse models, without any allowance being made for their good performance relative to reality". However, Abramowitz and Gupta (2008, 3-4) note that "to choose the best model ensemble, we must consider both the independence *and* performance of potential ensemble members" (our emphasis) and that "choosing model weights for an ensemble is then a process of deciding on a performance measure (or aggregation of performance measures) and then using a weight description that values performance and independence in an appropriate ratio." So this does seem to be a possible reply to Annan and Hargreaves's concern. However, there nonetheless remains a question about how performance relative to reality is integrated into Abramowitz and Gupta's approach, which, as stated, only takes divergence of outputs into account.

Abramowitz *et al.* present a stronger objection when they argue that "inter-model distances alone in the absence of observational data are an incomplete proxy for model independence" (2019, 95). According to them, when models perform well (i.e. model-observation distances are relatively small) they should not be considered dependent even if their outputs are similar (since their outputs are all close to observations). Essentially, this is because "an ideal definition of model dependence would only include variability in process representations that are not tightly observationally constrained" (*ibid*., 94). However, if models' outputs are similar merely because the models perform well, then it seems unlikely that the similarities across models' outputs reflect similarities in the sections of models' representation that are not tightly constrained by observations. If this is right, then measures of independence that rely on inter-model distances alone in the absence of observational data are clearly inadequate.

To address this concern, other a posteriori approaches assume that the level of dependence is a function of the level of model error covariance or error correlation (Collins *et al.* 2011; Bishop and Abramowitz 2013). As Abramowitz *et al.* (2019, 95) note, these approaches have "the advantage that 'error' only reflects deviations from an observational product (rather than similarity in model outputs per se)" and hence it is perhaps more reasonable to assume that "differences in the structure of error between models are likely to reflect differences in the sections of model representation that are not tightly constrained by observations".

However, despite this advantage, these more sophisticated a posteriori approaches are not immune to general concerns that apply to all a posteriori measures of independence. Indeed, as Pirtle *et al*. (2010, 354) remark, *all* a posteriori measures of independence "essentially treat models as black boxes, ignoring the causal reasons for disagreement between models. It is possible that two models could agree with respect to outputs despite their having different causal assumptions, but such a result, using this approach, would falsely indicate model 'dependence,' because these models would yield the same output despite the fact that they make different and possibly conflicting claims about the underlying mechanisms". Similarly, Annan and Hargreaves (2017, 218) worry that "[p]airwise similarity between model outputs may arise through convergence of different approaches to understanding the climate system, and not merely through copying of ideas, and this would not indicate any dependence as defined here. […] We do not believe that coincidentally similar behaviour should be penalised by downweighing of these models, as it may represent a true 'emergent constraint' on system behaviour". And Abramowitz *et al*. (2019, 98) worry about the sensitivity of a posteriori dependence measures "to the choice of variable, constraining observational data set, metric, time period and the region chosen".

A posteriori approaches are seen as pragmatic approaches to quantifying inter-model dependencies, and the underlying hope is that the proximity of models' results or model error correlations are good proxy measures for model interdependencies (i.e. the similarities in the way the models represent the world and its causal structure). However, the objections discussed cast doubt on whether a posteriori approaches to quantifying inter-model dependencies are really fit for purpose. In light of this, some scientists argue that inter-model dependencies should be assessed using a priori approaches instead, where "the independence of models is judged a priori, based only on the knowledge of their codes, and not of their results" (Boé 2018, 2772).

However, a priori approaches are still very much in their infancy. A very basic a priori approach is the "institutional democracy" proposed by Leduc *et al.* (2016). In contrast to the popular but highly criticised "one model, one vote" approach (Knutti 2010), under institutional democracy models that come from the same institution (i.e. the same modelling centre) are assumed to be fully dependent if they lead to equivalent projections and hence are considered as a single model when their signals are statistically indistinguishable (but not otherwise). The motivation behind this approach is that "[c]limate models developed within a given research group or institution are prone to share structural similarities" (Leduc *et al.* 2016, 8301) and hence institutional democracy could be used as a proxy for quantifying inter-model dependences. However, many have found the institutional democracy approach unsatisfactory, since models can have significant similarities despite not being from the same modelling centre and hence "deciding whether or not two GCMs are independent based on their institutions is just a first step. A better knowledge of how code similarity impacts GCMs['] results is needed to go forward" (Boé 2018, 2772).

Annan and Hargreaves (2017) propose a general account of independence that is determined a priori in terms of the *anticipated* outputs of the models. According to them, two models should be considered independent if a researcher's subjective belief about a possible outcome of one of the models in the ensemble is not affected by learning an output of the other model. However, this assessment of independence is extremely subjective, and they only show how it is supposed to work in cases where all the researcher knows is the model's institution.

Boé (2018) has recently proposed using the number of shared components by GCMs as a proxy for model independence, where each GCM is characterized by its four key components: atmosphere, ocean, land surface, and sea ice models. However, Boé acknowledges that this approach "is still crude and has some limits" (*ibid*., 2777). For a start, determining whether two components are different is not a trivial exercise and is bound to be rather subjective. Indeed, Boé relies on the version numbers of the GCMs' components to determine whether two components are different, but as Abramowitz *et al.* note "it is unlikely that the approach to version numbering is consistent across modelling centres, meaning that two components might be very different even if they share a major version number, or vice versa" (2019, 94). Furthermore, as Boé (2018, 2777) notes, different versions of a component "often share identical parameterization schemes and are therefore themselves not independent". Another issue that Boé points out is that "the impact of tuning is not considered. Some components may be considered 'identical' in this work but use different parameters, which may be a source of important differences. A better documentation of tuning in GCMs would be necessary to go further" (Boé 2018, 2777). Finally, as Abramowitz *et al.* (2019, 100) remark, "Boé's approach quickly becomes difficult and time consuming for large ensembles such as CMIP, given the lack of transparency regarding precisely what constitutes different models and the role of tuning". And furthermore "shared history as it pertains to dependence should only include process representations that are not tightly observationally constrained (so that Navier–Stokes equations might not represent dependent process treatment, for example)" (*ibid*., 100), which might further complicate things.

Overall, although a priori approaches to measure inter-model dependencies may intuitively seem more promising, they clearly face considerable challenges. Indeed, there is currently no scientific consensus on how to measure inter-model dependencies.

3.2 Agreement

Agreement seems to be a simple concept, but on closer examination it turns out that different and often incongruent notions of agreement are at work in climate modelling. We follow common usage and take "agreement" and "robustness" to be interchangeable (Parker 2011, 580). We note, however, that some climate scientists (e.g. Pirtle *et al*. (2010)) assume that for a model result to be robust the models not only have to agree on it, but they also have to be sufficiently independent from one another. We don't follow this usage in the remainder of this chapter. Even if one were willing to set aside the challenges (discussed in the previous subsection) that arise when we try to identify an adequate account of model independence, a discussion of the epistemic import of robustness analysis becomes significantly more difficult if we muddy the waters by including some notion of epistemic significance in the *definition* of robustness.

A key consideration when formulating a definition of robustness in climate science is how many models must agree on a result for it to be deemed robust. While the most straightforward notion of robustness would require all models to agree, this notion is not universally accepted. Almazroui *et al*. (2016, 164), for instance, define an increase (or decrease) in the projected signal "to be robust if at least 66% of the models agree in the direction of change". By contrast, Screen and Blackport (2019, 11410) define "a robust response as being when nine or greater (of the eleven) models depict individual responses of the same sign as the ensemble-mean response". In yet other cases, robustness is assumed to be a non-categorial notion, where "the greater the number of models in agreement, the greater the robustness" (Field *et al.* 2012, 131).

This draws attention to an important question that arises when defining robustness in climate science: is robustness best thought of as a categorical notion? If so, how many models have to agree on a result for it to be considered robust? If not, how should we understand the notion of robustness in a non-categorical way, and is robustness comparable across different model ensembles and different results?

Another important aspect of a definition of robustness in climate science has to do with what it takes for models to agree in the first place. One approach might be to consider a particular range of possible results and ask: do all (or most) of the models agree on this range of possible results (where the range in question can be determined even after looking at the models' individual results)? If the question is answered affirmatively, then that range is considered robust. This is how Baumberger *et al*. (2017, 9) seem to understand robustness when they write that "[a] model projection is robust if all or most models in the ensemble agree regarding the projection. If all models in an ensemble show more than a 4˚C increase in global mean surface temperature by 2100 when run under a certain forcing scenario, this projection is robust".

Alternatively, one might define model robustness without reference to a range of possible results. Ukkola *et al*., for instance, regard "projections as 'robust' when the magnitude of the multi-model mean future change exceeded the inter-model standard deviation of the change" (2020, 4). This is a very different understanding of robustness, and its application is limited to answering the question of whether a particular change (e.g. an increase in mean precipitation or an increase in the frequency and duration of seasonal meteorological drought) is robust. Under this approach, the question is no longer whether a *range* is robust, but rather whether the mean of the models' results is sufficiently large (i.e. larger than the inter-model standard deviation). This implies that even if most of the models show more than a 4˚C increase in global mean surface temperature by 2100 when run under a certain forcing scenario, if the mean of the models' results is not greater than the inter-model standard deviation, then the models' results would not be considered robust. Notice further than under this approach, the greater the magnitude of the multi-model mean future change, the more spread out the models' projections are allowed to be before the change is deemed not to be robust.

The rationale behind this definition of robustness is not entirely clear and gives rise to several questions. To begin with, why should we focus on the multi-model mean to determine whether model projections are robust? In other words, why is the multi-model mean the relevant variable here, rather than, say, an unequally weighted mean or another variable altogether? Indeed, if models are dependent, then it is unclear why the multi-model mean is a meaningful variable to consider in the first place. At the beginning of this section, we counselled against including a notion of epistemic significance in the definition of robustness, and it is important to note that the focus on the multi-model mean is a choice that disagrees with this maxim by at least implicitly relying on a model democracy approach, and that this is certainly not the only choice available. Furthermore, why should this mean be greater than one inter-model standard deviation rather than, say, two standard deviations? We ask these questions to highlight the fact that this approach to defining model robustness relies on several not obviously natural choices.

Tebaldi *et al.* (2011, 1) make a distinction between a lack of signal and a lack of information due to model disagreement and "categorize three levels of multi-model agreement: 1) the majority of models agree that future changes will be statistically significant and of the same sign 2) the majority of models show significant change but in opposite directions and 3) most of the models show no significant change. The basic idea is that testing for model agreement is only meaningful if the models are producing significant changes, i.e., changes outside of internal variability" (*ibid*., 4). This categorization clearly has profound consequences for how one determines whether a result is robust. For instance, as Tebaldi *et al.* remark, "in contrast to popular belief, model agreement of future precipitation change is greater than currently thought. Only few places in the world show significant changes of opposite sign in different models" (*ibid*., 4). It is also worth noting that, under this approach, the extent of the assessed agreement may be affected by the choice of method used to assesses the natural variability of the system (since this might affect whether one deems a change in the model to be statistically significant).

Each of these three approaches define robustness in a different way, and hence could lead to a different assessment of robustness in any given case. The fact that scientists have not agreed on a definition of robustness adds a layer of complexity to investigations into the epistemic import of model robustness in climate science.

4. Establishing the Premises

In this section, we turn to the question of how to establish the three premises of the general inference pattern of MBRA presented in Section 2.

4.1 Premise of Step 1: Finding the Robust Property

To justify the premise of Step 1, one must establish that all models in an ensemble of models Ω have property *R*. But what kind of ensemble of models should one consider? There are three kinds of ensembles in climate science that are worth discussing: perturbed parameter ensembles, multi-model ensembles, and initial condition ensembles.

These three kinds of model ensembles are explored through different techniques. Studying a perturbed parameter ensemble requires us to vary the parameters in the model and check whether, and if so how, the desired results change. This is simple in theory, but it is often difficult to do in practice. The number of parameters may be large, and equations may not be solvable analytically. In such cases scientists have to resort to computer simulations and run multiple versions of the same model, where each version incorporates a different set of parameter values. But no amount of simulation results can explore the full range of parameter values, and there will always be gaps. These gaps are particularly significant if models are large and computationally costly to explore. Contemporary climate models, for instance, have hundreds of parameters and yet the available computational infrastructure only allows scientists to make a comparatively small number of runs, which results in large parts of the parameter space remaining unexplored. For instance,- HadCM3, a global climate model on which the UK's official climate policies were based until a few years ago, has hundreds of parameters (leading to billions of combinations of values), and yet the results communicated to policy makers were based on less then 300 model runs, only 17 of which were runs of the full model (for a discussion of this case, see Frigg, Smith and Stainforth's 2013; 2015). Understanding how changes in model parameters affect the model result of interest in the face of difficulties like these has turned into a scientific discipline in its own right, namely *sensitivity analysis*. Philosophical discussions of sensitivity analysis can be found in Bokulich and Oreskes' (2017, Sec. 41.6) and Raerinne's (2013, Sec. 2); its place in the broader edifice of RA is discussed in Justus' (2012, 801) and Weisberg and Reisman's (2008, 115). For a technical discussion see Saltelli, Tarantola, Campolongo, and Ratto's (2004).

Things get even more complex when we turn to multi-model ensembles. The purpose of such ensembles is to evaluate whether a result is robust under structural changes to the model. This involves changing the substantial modelling assumptions and the mathematical structure of the model. Such stability is required because if a model is idealised and it turns out that a result vanishes when idealisations are removed or changed, then the result is not epistemically significant. Fletcher (2020) traces the demand for stability back to Duhem and Maxwell, and then discusses topological notions of stability in dynamical systems; for further discussions of that kind of stability, see Frigg, Bradley, Du, and Smith's (2014) and Frigg and Smith's (2022). Making good on this intuition is a challenging task. Unlike in the case of perturbed parameter ensembles, where the problem is to establish results about a well-defined ensemble, the problem now is how to define the ensemble to begin with. In the example discussed in Section 2, Weisberg and Reiman considered a small multi-model ensemble consisting of three models and then studied each model individually. But what justifies this choice? Why these three models? Why not an ensemble of four, or five models, or an ensemble with a larger, or even infinite, number of models? In climate science, multi-model ensembles often include considerably more than three models. Phase 6 of the Coupled Model Intercomparison

Projection (CMIP6), for instance, includes over 100 models (PCMDI 2022). However, no matter how large ensembles may be, they are nonetheless “ensembles of opportunity” (Tebaldi and Knutti 2007; Parker 2013) because the selection of models to be included in these ensembles is neither systematic nor standardized, and models are not constructed to sample the existing uncertainty. Rather, what models are included in any given multi-model ensemble will ultimately depend on contingent factors such as what state-of-the-art models are currently available and whether a modelling group is willing and able to do the requested simulations.

The third type of model ensembles, initial condition ensembles, is one in which the initial conditions of a model are perturbed, and what is studied is how the result responds to this perturbation. In this context there is an important distinction between predictions and projections that is worth mentioning, for it affects the interpretation of an initial ensemble. Predictions are claims about the actual evolution of the climate system based on *current* initial conditions. Whereas projections are claims about the response of the climate system to external forcing scenarios based on possible initial conditions where the system has at least partially adjusted to the external forcings a t_0 (where t_0 is some point of time during the pre-industrial period). Hence, whereas for prediction an initial ensemble is interpreted as estimates of the *actual* initial conditions, for projections an initial ensemble is interpreted differently, namely as *potential* initial conditions at pre-industrial times. In the climate literature it is often assumed that when it comes to projections, initial-condition uncertainty is not very important. For instance, Tebaldi *et al*. (2007) state that “[i]nitial condition uncertainty is most relevant for the shortest time scales. Weather is chaotic, and predictions are sensitive to the value of observations used to initialize numerical models […] Long-term projections of climate change are typically averaged over decades and often across several ensemble members, and are thus largely insensitive to small variations in initial conditions”. Indeed, it is common practice to only consider very few initial states per model (often only one to five, and rarely more than ten). However, initial uncertainty might be more important. Werndl (2020) has recently argued that there is little if any justification for the claim that projections are independent of the details of the initial ensemble, and research does suggest a much larger number of initial conditions are needed to reliably estimate projections (see for instance, Daron and Stainforth (2013) and Deser *et al*. (2012)).

4.2 Premise of Step 2: Finding the Common Structure

The task of establishing that the models in the ensemble have a common structure (as required by the premise of Step 2) can be broken down into two sub-tasks: (a) say what it means for there to be such a structure and (b) state how we find it.

Let us begin with (a). One way in which models in an ensemble can have a common structure S is it being the case that every model M_i in the ensemble can be decomposed into a core and a set of idealisations: $M_i = S \,\&\, I_i$, where i is an index that ranges over all the models in the ensemble. The crucial aspect here is that while idealisations are particular to each model (hence the index for the idealisations I_i), the structure S must be common to all models. Rice calls this the “decompositional strategy” and argues that it is a dead end: “many of our best scientific models cannot be decomposed in the ways required by the decompositional strategy” (Rice 2019, 180). The contributions of S to a model’s output cannot be isolated from the contributions of the contributions of I_i because the two are inextricably intertwined and they collaborate to produce the model’s output. The idealisations are introduced to render the basic mathematical frameworks applicable, and they often distort difference-making features. Hence, there is no such thing as the contribution of the idealisation that can be isolated from the result of the core (*ibid*. 189-95).

This is a serious worry and those who wish to perform an RA on a given ensemble will have to argue that the models at stake do not face the issue Rice describes. But even if this were possible and decomposition would not be an in-principle limitation, there still remain practical obstacles, which bring us to (b): saying how to find the common structure. Few ensembles will consist of models whose structure naturally decomposes into a core and idealisations, and different models may even be formulated in different mathematical frameworks. It is then a challenge to find a core structure that they all have in common. Weisberg and Reisman's (2008) ensemble is one case in point. The models use different formalisms and isolating negative coupling as the common structure involved much more than just watching out for shared elements in the mathematical formulation of the models. Weisberg (2006, 738) recognises this difficulty and notes that "[s]uch cases are much harder to describe in general, relying as they do on the theorist's ability to judge relevantly similar structures". Even if one has faith in theorists' ability to do so, certain cases may present insurmountable obstacles. Justus (2012, 802-03) discusses the case of climate models and points out that these large computational structures are opaque, and that the sheer number and complexity of equations involved undercuts any attempt to duplicate the kind of analysis that Weisberg and Reisman were able to carry out on the relatively simple models of the predator-prey system.

This brings us to the issue of scrutability. According to Justus (2012, 802) determining that the models in an ensemble have a common structure requires that the models' structure and dynamics be "scrutable". One could interpret this as the claim that the models have to be totally scrutable: we have to know and understand every detail of their structure. But is this necessary? Can "partial scrutability" be enough? For instance, we can certainly scrutinise climate models to the point that we know that all of them have increasing concentration of greenhouse gases. Could this suffice to identify a common structure? Lloyd (2015, 62), for instance, seems to think so when she writes that "despite this model variation, all the models in this model family share a core representation of greenhouse gases (GHG) as a radiative cause. *We can consider this the common causal core shared by this entire GHG model-type under consideration*" (our emphasis). As we will see shortly, this may be requiring too little by way of scrutability. But then a question naturally arises: if we don't think we have to go all the way to full scrutability, how much scrutability do we need to be able to identify a common structure in a model ensemble?

According to Katzav (2014, 230) "some false assumptions that are shared by all the GCMs in question play a crucial role – the models would not run without them – in generating model successes". But then it seems clear that partial scrutability of the kind required by Lloyd in her identification of a common structure is not enough. For if Katzav is right, we would, at the very least, need enough scrutability to identify as part of the common structure the false assumptions that are responsible for the models' successes. This, in turn, gives rise to two concerns. First, as Katzav remarks, this may not always be possible for "[w]e will often have good reason to conclude that some shared climate model assumptions are wrong without being able to identify which are wrong" (*ibid*., 230). Second, even if we had enough scrutability to determine all the false assumptions that should be identified as part of the common structure, the very fact that the common structure would include some false assumptions is problematic. For in such cases, we would know that the conclusion of Step 2 of the general inference pattern of MBRA presented in Section 2 is false, and hence it would no longer be clear what the epistemic role of Step 2 is.

In sum, when it comes to the premise of Step 2, there remain question marks about the identification of a common structure both in principle and in practice.

4.3 Premise of Step 3: Understanding Robust Theorems

As we have noted when discussing the Volterra Principle, the formulation of the relevant conditions of the robust theorem, which is an important part of Step 3, is a formidable problem. The problem is linked to the issues concerning the construction of multi-model ensembles. If we fully understood the structure of the models in the ensemble, and if we could show that in all these models *S* brings about *R*, then we would probably have at least some idea about what goes into *C*. But since a full characterisation of models in the ensemble often remains elusive, it is unsurprising that formulating the relevant criteria remains a hard nut to crack.

But even if we had an idea of the content of *C*, there would be the further worry that *C* might include unrealistic assumptions about the target system. Suppose, for instance, that we didn't know that the Volterra principle is insensitive to density dependence. In this case we could not assume that the Volterra principle concerns any system with density dependence, no matter how small. But a Volterra principle which concerns only predator-prey systems with no density dependence at all is not a theorem about the actual world, because any real system is bound to have some density dependence. To give another example, if we don't know whether or not the Volterra principle is sensitive to predators' and preys' responses to seasonal fluctuations, we cannot assume that the Volterra principle concerns predator-prey systems in a seasonal environment. And yet virtually all real-world biological populations live in a seasonal environment. Arguably, this might be the very worry that underlies Weisberg's recommendation to collect "a sufficiently diverse set of models so that the discovery of a robust property does not depend in an arbitrary way on the set of models analyzed" (2006, 737). But this is insufficient to put the worry to rest. For no matter how diverse the set may be, if all the models in the ensemble involve a particular unrealistic assumption about the target system despite differing in many other respects, there is no justification for not including that unrealistic assumption in *C*. But if *C* ends up including even just one unrealistic assumption about the target system, the robust theorems end up being inapplicable to the actual world, and this renders them useless for explaining or predicting real-world phenomena.

5. Conclusion

In this chapter we have introduced the structure of MBRA and analysed its core concepts and premises. We have seen that the articulation of these concepts raises important questions, and that establishing the premises is a formidable task. But the most serious challenge still lies ahead: justifying the inferential step that takes us from the premises to the conclusions. None of the three inferential steps that occur in MBRA are deductively valid. But the use of a deductively invalid inference pattern needs a justification, assuring us that at least in the instances in which we use it, the conclusions we draw are nevertheless correct. This is the topic of Part II of this review.

References

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

Abramowitz, G., & Gupta, H. (2008). Toward a model space and model independence metric. *Geophysical Research Letters*, *35*(5), doi:10.1029/2007gl032834.

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., et al. (2019). ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing. *Earth System Dynamics*, *10*(1), 91-105.

Almazroui, M., Saeed, F., Islam, M. N., & Alkhalaf, A. (2016). Assessing the robustness and uncertainties of projected changes in temperature and precipitation in AR4 Global Climate Models over the Arabian Peninsula. *Atmospheric Research*, *182*, 163-175.

Annan, J. D., & Hargreaves, J. C. (2017). On the meaning of independence in climate science. *Earth System Dynamics*, *8*(1), 211-224.

Baumberger, C., Knutti, R., & Hirsch Hadorn, G. (2017). Building confidence in climate model projections: an analysis of inferences from fit. *WIREs Climate Change*, *8*(3), e454.

Bishop, C. H., & Abramowitz, G. (2013). Climate model dependence and the replicate Earth paradigm. *Climate Dynamics*, *41*(3-4), 885-900.

Boé, J. (2018). Interdependency in Multimodel Climate Projections: Component Replication and Result Similarity. *Geophysical Research Letters*, *45*(6), 2771-2779.

Bokulich, A., & Oreskes, N. (2017). Models in Geosciences. In L. Magnani, & T. Bertolotti (Eds.), *Springer Handbook of Model-Based Science*. Dordrecht, Heidelberg, London and New York: Springer, 891-911.

Chalmers, A. (2011). Drawing philosophical lessons from Perrin's experiments on Brownian motion: A response to van Fraassen. *The British Journal for the Philosophy of Science*, *62*(4), 711-732.

Collins, M., Booth, B. B. B., Bhaskaran, B., Harris, G. R., Murphy, J. M., Sexton, D. M. H., et al. (2011). Climate model errors, feedbacks and forcings: a comparison of perturbed physics and multi-model ensembles. *Climate Dynamics*, *36*(9-10), 1737-1766.

Daron, J. D., & Stainforth, D. A. (2013). On predicting climate under climate change. *Environmental Research Letters*, *8*(3), 034021.

Deser, C., Phillips, A., Bourdette, V., & Teng, H. (2012). Uncertainty in climate change projections: the role of internal variability. *Climate Dynamics*, *38*(3-4), 527-546.

Field, C. B., Barros, V., Stocker, T. F., & Dahe, Q. (2012). *Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change*: Cambridge University Press.

Fletcher, S. C. (2020). The Principle of Stability. *Philosopher's Imprint*, *20*(3), 1-22.

Frigg, R. (2022). *Models and Theories. A Philosophical Inquiry*. London: Routledge.

Frigg, R., Bradley, S., Du, H., & Smith, L. A. (2014). The adventures of Laplace's demon and his apprentices. *Philosophy of Science*, *81*(1), 31–59.

Frigg, R., & Hartmann, S. (2020). Models in Science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, URL = <https://plato.stanford.edu/archives/spr2020/entries/models-science/>.

Frigg, R. & Smith, L. A. (2022). An ineffective antidote for hawkmoths. *European Journal for Philosophy of Science*, *12*, Article 33.

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

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

Frigg, R., Thompson, E., & Werndl, C. (2015). Philosophy of Climate Science Part II: Modelling Climate Change. *Philosophy Compass*, *10*, 965-977.

Furrer, R., Sain, S. R., Nychka, D., & Meehl, G. A. (2007). Multivariate Bayesian analysis of atmosphere–ocean general circulation models. *Environmental and Ecological Statistics*, *14*(3), 249-266.

Gueguen, M. (2020). On Robustness in Cosmological Simulations. *Philosophy of Science*, *87*(5), 1197-1208.

Hudson, R. (2020). The Reality of Jean Perrin's Atoms and Molecules. *The British Journal for the Philosophy of Science*, *71*(1), 33-58.

Justus, J. (2012). The Elusive Basis of Inferential Robustness. *Philosophy of Science*, *79*(5), 795-807.

Katzav, J. (2014). The epistemology of climate models and some of its implications for climate science and the philosophy of science. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *46*, 228-238.

Knutti, R. (2010). The end of model democracy? *Climatic Change*, *102*(3-4), 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.

Kuorikoski, J., Lehtinen, A., & Marchionni, C. (2010). Economic Modelling as Robustness Analysis. *The British Journal for the Philosophy of Science*, *61*(3), 541-567.

Leduc, M., Laprise, R., De Elía, R., & Šeparović, L. (2016). Is Institutional Democracy a Good Proxy for Model Independence? *Journal of Climate, 29*(23), 8301-8316.

Levins, R. (1966). The Strategy of Model Building in Population Biology. *American Scientist*, *54*(4), 421–431.

Lloyd, E. A. (2010). Confirmation and Robustness of Climate Models. *Philosophy of Science*, *77*(5), 971-984.

Lloyd, E. A. (2015). Model robustness as a confirmatory virtue: The case of climate science. *Studies in History and Philosophy of Science*, *49*, 58-68.

Mayo, D. G. (1986). Cartwright, causality, and coincidence. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *1986* (Vol. 1): Philosophy of Science Association, 42-58.

Parker, W. S. (2011). When Climate Models Agree: The Significance of Robust Model Predictions. *Philosophy of Science*, *78*(4), 579-600.

Parker, W. S. (2013). Ensemble modeling, uncertainty and robust predictions. *Wiley Interdisciplinary Reviews: Climate Change*, *4*(3), 213-223.

PCMDI (2022). ESGF CMIP6 Data Holdings. https://pcmdi.llnl.gov/CMIP6/ArchiveStatistics/esgf_data_holdings/. Accessed 24 June 2022.

Pirtle, Z., Meyer, R., & Hamilton, A. (2010). What does it mean when climate models agree? A case for assessing independence among general circulation models. *Environmental Science & Policy*, *13*(5), 351-361.

Plutynski, A. (2006). Strategies of Model Building in Population Genetics. *Philosophy of Science*, *73*(5), 755-764.

Psillos, S. (2011). Moving Molecules Above the Scientific Horizon: On Perrin's Case for Realism. *Journal for General Philosophy of Science*, *42*(2), 339-363.

Raerinne, J. (2013). Robustness and sensitivity of biological models. *Philosophical Studies*, *166*(2), 285-303.

Rice, C. (2019). Models don't decompose that way: a holistic view of idealized models. *The British Journal for the Philosophy of Science*, *70*(1), 179–208.

Saltelli, A., Tarantola, S., Campolongo, F., & Ratto, M. (2004). *Sensitivity Analysis in Practice. A Guide to Assessing Scientific Models*. Chichester: John Wiley & Sons Ltd.

Sanderson, B. M., Knutti, R., & Caldwell, P. (2015). Addressing Interdependency in a Multimodel Ensemble by Interpolation of Model Properties. *Journal of Climate*, *28*(13), 5150-5170.

Seager, R., Ting, M., Held, I., Kushnir, Y., Lu, J., Vecchi, G., Huang, H.-P., Harnik, N., Leetmaa, A., Lau, N.-C., Li, C, Velezan, J., & Naik, N. (2007). Model Projections of an Imminent Transition to a More Arid Climate in Southwestern North America. *Science*, *316*(5828), 1181-1184.

Schupbach, J. N. (2018). Robustness Analysis as Explanatory Reasoning. *The British Journal for the Philosophy of Science*, *69*(1), 275–300.

Screen, J. A., & Blackport, R. (2019). How Robust is the Atmospheric Response to Projected Arctic Sea Ice Loss Across Climate Models? *Geophysical Research Letters*, *46*(20), 11406-11415.

Smith, L. (2007). *Chaos: a very short introduction*. Oxford: Oxford University Press.

Sprenger, J. (2012). Environmental Risk Analysis: Robustness Is Essential for Precaution. *Philosophy of Science*, *79*(5), 881-892.

Staley, K. W. (2004). *The Evidence for the Top Quark: Objectivity and Bias in Collaborative Experimentation*. Cambridge: Cambridge University Press.

Tebaldi, C., Arblaster, J. M., & Knutti, R. (2011). Mapping model agreement on future climate projections. *Geophysical Research Letters*, *38*(23), doi:10.1029/2011gl049863.

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.

Ukkola, A. M., De Kauwe, M. G., Roderick, M. L., Abramowitz, G., & Pitman, A. J. (2020). Robust Future Changes in Meteorological Drought in CMIP6 Projections Despite Uncertainty in Precipitation. *Geophysical Research Letters*, *47*(11), doi:10.1029/2020gl087820.

Vicedo-Cabrera, A. M., Scovronick, N., Sera, F., Royé, D., Schneider, R., Tobias, A., et al. (2021). The burden of heat-related mortality attributable to recent human-induced climate change. *Nature Climate Change*, *11*(6), 492-500.

Weisberg, M. (2006). Robustness Analysis. *Philosophy of Science, 73*(5), 730-742.

Weisberg, M., & Reisman, K. (2008). The Robust Volterra Principle. *Philosophy of Science*, *75*(1), 106-131.

Werndl, C. (2020). Initial-condition dependence and initial-condition uncertainty in climate science. *The British Journal for the Philosophy of Science 70*(4), 953-976.

Wittgenstein, L. (1953). *Philosophical Investigations* (G. E. M. Anscombe, Trans., 3rd ed.). Oxford: Blackwell.