Climate Models and Robustness Analysis – Part II: The Justificatory Challenge

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 from scientific models. 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 second part of a two-part review of MBRA, addresses the thorny issue of justifying the inferential steps taking us from the premises to the conclusions. We begin by making explicit what exactly the problem is. We then turn to a discussion of two broad families of justificatory strategies, namely top-down and bottom-up justifications. In the latter group we distinguish between the likelihood approach, independence approaches, and the explanatory approach. This discussion leads us to the sober conclusion that multi-model situations raise issues that are not yet fully understood and that the methods and approaches that MBRA has not yet reached a stage of maturity. Important questions remain open, and these will have to be addressed in future research.

Keywords

Robustness analysis, model ensemble, model agreement, independence, common structure, climate model, top-down justification, bottom-up justification, likelihood approach, independence approach, reliability independence, confirmational independence, explanatory approach.

Table of Contents

1. Introduction

The core idea of *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 many contemporary scientific contexts, the evidence in support of a hypothesis comes from scientific models. 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. A general statement of MBRA is 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*.

In Part I of this review we analysed the core notions of independence and agreement, and we discussed what it would take to establish the premises. In the current chapter, Part II of the review, we address the thorny issue of justifying the inferential steps taking us from the premises to the conclusions. We begin by making explicit what exactly the problem is (Section 2). We then turn to a discussion of two broad families of justificatory strategies, namely top-down justifications (Section 3) and bottom-up justifications (Section 4). We end by assessing the success of MBRA and highlighting alternative ways of dealing with multi-model situations (Section 5).

2. The Justificatory Challenge

There is a justificatory problem because the inferences linking the premises and the conclusions in each step of the general inference pattern of MBRA are not deductively valid: at each step it is possible for the premise to be true while the conclusion is false. 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. One might be inclined to dismiss this as a philosophical nicety. Inductive inferences are not deductively valid either, but one might say that worries about whether one can infer from the fact that the sun has risen every morning in the past that it will also rise tomorrow (Hume 1748/2007) are for exhilarated philosophical minds and can be safely set aside by practicing scientists. Whatever one's views on Hume's problem of induction, the justificatory problem for MBRA is not like that, and it is one that *should* worry practitioners. Talking about econometric models, Cartwright formulates the worry thus:

> Now here is the reasoning I do not understand: "Econometrician X used a linear form, Y a log linear, Z something else; and the results are the same anyway. Since the results are so robust, there must be some truth in them." But […] we know that at the very best one and only one of these assumptions can be right. We may look at thirty functional forms, but if God's function is number thirty-one, the first thirty do not teach us anything. […] I agree that it is a coincidence that they all find the same results. But I do not see what reason we have to assume that the correct explanation for the coincidence is that each of the instruments, despite its flaws, is nevertheless reading the outcome correctly. (1991, 154)

And worries about MBRA are not confined to the philosophical literature. Climate physicist J. Räisänen also draws attention to it:

> The risk that the uncertainty in the real world exceeds the variation between model results is obvious: even if all models agreed perfectly with each other, this would not prove that they are right. From a more physical perspective, some authors have argued that the differences between the parametrization schemes used in existing models do not cover the actual uncertainty in the representation of subgrid scale processes. (2007, 9)

So the challenge is: how can we justify the inference from facts about models to facts about the world? In the next sections, we will discuss and assess two distinct approaches that have been taken to address this justificatory challenge: top-down justifications and bottom-up justifications.

3. Top-Down Justifications

Top-down justifications aim to justify the inferential step by appeal to properties of an entire model ensemble. These approaches contrast with what we call bottom-up justifications, which are approaches that do not rely on ensemble properties and see confirmation as coming from individual models, one at a time. We discuss top-down approaches in this section; bottom-up approaches will occupy us in Section 4.

Orzack and Sober address the justificatory challenge and give a negative answer: MBRA cannot be justified. To reach this conclusion, they distinguish three cases (1993, 538). The first case is one in which "we know that one of a set of models [Ω] is true, but we do not know which" (*ibid.*) where, by a model being "true", they mean that the model represents T accurately in the relevant respects. This option is also discussed in the climate modelling literature, although in a probabilistic version (cf. the second qualification in Sec. 2). Baumberger *et al.* (2017), for instance, state that "[a]n inference from robustness of projections to their likely truth is legitimate if we have reasons to assume that it is likely that at least one model in the multimodel ensemble correctly projects the quantity of interest within the specified error margin" (2017, 10; see also Parker 2011, 584). Orzack and Sober's second case is that Ω is known not to contain the true model: each model in Ω is false. Their third case is that it is unknown whether Ω contains the true model.

Orzack and Sober argue that it is obvious that the second and the third options fail to support the relevant inferences. In the second case, this is because "[i]f we know that each of the models is false (each is a 'lie'), then it is unclear why the fact that R is implied by all of them is evidence that *R* is true" (*ibid.*). In the third case this is because "[i]f we do not know that one of the models is true, then it is again unclear why a joint prediction should be regarded as true" (*ibid.* 538-39).

Let us turn to the first option. The problem with the first case is different from that with the second and third cases. Orzack and Sober admit that under the assumption of the first case the inference *is* valid because a result is robust if all models in Ω agree on the result, and if Ω contains the true model, then all models agree on the truth, and therefore the robust result is true. Their worry is that this scenario is unrealistic. For one, it is far from obvious why a true model should be part of Ω to begin with, given that models typically involve simplifications and omissions. For another, even if were lucky enough to have an ensemble that contained the true model, we would rarely, if ever, be in the situation to know this to be the case.

Whether this point holds depends on other characteristics of the ensemble. Orzack and Sober discuss the case where Ω is the ensemble that scientists have *de facto* constructed. Under this assumption, their argument is hard to refute. But let us now change the scenario and assume that the ensemble is a complete ensemble Ω_c, an ensemble that contains every possible model of *T*. This is of course an entirely hypothetical scenario, but it's worth asking the question whether MBRA would be justified under this strong assumption.

There are several worries about justifying MBRA by appeal to Ω_c. The first is that it is unclear what this ensemble would be and how it should be circumscribed. What would the complete class of all models of the earth's climate be? Answering this question would amount to spelling out the mathematical form of every possible climate model and explicating how the relevant equations represent the climate. One need not have an overly pessimistic outlook on climate modelling to come to the conclusion that this an entirely unrealistic endeavour. But even if Ω_c could somehow circumscribed, it is unlikely that this would be of any use. An ensemble of *all* possible models will also contain models that are misrepresentations of the climate (see Frigg and Nguyen 2020, Ch.1, for discussion of misrepresentations). But misrepresentations will disagree with accurate representations on certain features, and these may well include *R* and *S*. So it is to be expected that *R* and *S* are not actually robust properties in Ω_c.

In response to this worry, one could change the hypothetical scenario and say that the relevant ensemble is the one that contains all models of *T* that have certain empirical credentials. Call this ensemble Ω_{ce}. Indeed, in his reply to Orzack and Sober, Levins stressed the fact that a relevant ensemble must be an ensemble of models with empirical credentials (1993, 554). Likewise, Weisberg emphasised that the models in the ensembe must have what he calls "low-level confirmation" (Weisberg 2006, 740). It is precisely the fact that models in the ensemble enjoy a degree of confirmation that makes them relevant. This raises three sets of questions: (a) how are we to understand low-level confirmation?, (b) does Ω_{ce} warrant the inferences we are interested in?, and (c) is there any chance of working with Ω_{ce} in practice?

As regards (a), Weisberg (2006) argues that what justifies the inferential step in MBRA is what he calls "low-level confirmation". Low-level confirmation essentially means that the models in the ensemble get certain basic features or properties about the system right. Writing about the ecological models that we introduced in Part I, Weisberg puts the point thus:

> In the predation case, for example, we are confident that ecological relationships can be represented with the models described by coupled differential equations. Thus when

> we discover the consequences of these models, we are confident that most of these consequences are true of any system described by the model[s]. This confidence comes from low-level confirmation, not from robustness analysis itself. Thus robustness analysis is not a nonempirical form of confirmation as Orzack and Sober suggest. It does not confirm robust theorems; it identifies hypotheses whose confirmation derives from the low-level confirmation of the mathematical framework in which they are embedded. (*ibid*., 741)

However, Weisberg's notion of low-level confirmation raises several questions. For a start, as Houkes and Vaesen (2012, 353) observe, Weisberg is ambiguous about the scope of his notion of low-level confirmation. Is it supposed to apply to a broad mathematical framework, say that of coupled differential equations? Or to a specific model family? Or to individual models? Secondly and relatedly, if low-level confirmation is the sort of confirmation that licenses the use of a framework to construct models of phenomena in the first place, then what are the features or properties of the framework that we compare to reality to determine when we are indeed licensed to do so? Without a clear understanding of the scope in the notion of low-level confirmation, it seems particularly hard to give an adequate answer to this question, and it is not clear what it would come down to in the case of climate models.

As regards (b), without a better understanding of the nature of low-level confirmation, it is hard to see why a set Ω_{ce} of models that enjoy a degree of low-level confirmation should in fact warrant the inferences we are interested in. Weisberg is aware of this problem and tries to address the worry by noting that low-level confirmation licenses us to believe that for all the models we have collected to undergo robustness analysis "when we discover the consequences of these models, we are confident that most of these consequences are true of *any system described by the model[s]*" (Weisberg 2006, 741, our emphasis). This shifts the target, as the systems that the models now are said to represent are fictional systems, that is systems that are unrealistic with respect to the actual target system in various respects. What low-level confirmation then tells us is that if those fictional systems happened to exist in the real world, then we would be entitled to believe that the consequences of our models are true in those systems. But this does not solve our initial problem: if we are interested in learning about properties of the actual target system and not of fictional systems, low-level confirmation of that kind offers no help.

Is there then another way we could try to define Ω_{ce}? An option might be to rely on Parker's notion of adequacy for purpose. According to Parker (2009, 2020) it is wrong to try to confirm models wholesale. Instead, what we should (and aften can) do is to confirm whether a model is adequate for purpose. Let us then call a model *A*-adequate if all claims that the model makes concerning *A* are true in the target *T*. Ω_{ce} can then be understood as the set of all A-adequate models. Although this *may* be a definable set, we now face a catch-22. On the one hand, if we are interested in *A*, then the ensemble is superfluous because we can just take one of the models to study *A*. On the other hand, if we are interested in a different aspect *B*, then it is useless to know that models are *A*-adequate because from the fact that they are *A*-adequate it does not follow that they are also *B*-adequate. So, either MBRA is superfluous, or *A*-adequacy fails to justify MBRA.

One way out might be to define Ω_{ce} as the ensemble of all *possibly* A-adequate models (i.e. this set would span *all* current scientific uncertainty about how to adequately represent the climate system for the predictive tasks at hand). Under the assumption that this ensemble includes at least one model that is A-adequate, learning that all models make the same claims concerning A would allow one to infer that those claims are true. However, what this set actually consists in and wether it is a well defined set to begin with it is not at all clear (due to concerns similar to those that we expressed earlier about Ω_c).

As regards (c), even if the problems with (a) and (b) could somehow be circumvented, this would still leave us with the question of how to work with Ω_{ce} in practice. There seem to be two options here: either we work with Ω_{ce} itself or we work with a representative sample of it. On the latter view, the representative sample plays an evidential function: it is understood as informing us about properties of all models in Ω_{ce} in much the same way in which an opinion poll with a few hundred participants is taken to inform as about the views of the entire population. However, neither of these options is realistic in the context of climate science. The first option is clearly a non-starter because there is no way to actually construct Ω_{ce}. The second option isn't viable either because, as mentioned in Part I (Section 4.1), climate model ensembles are ensembles of opportunity and they are not constructed to representatively sample 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. But even if models were intended to provide a representative random sample of Ω_{ce}, there are reasons to think that they could not actually constitute such a sample because, as Winsberg remarks, "[o]ne obvious reason to doubt [this assumption] is that all of the climate models on the market have a shared history. Some of them share code; scientists move from one lab to another and bring ideas with them; some parts of climate models (though not physically principled) are from a common toolbox of techniques, etc." (2018, 99). But then, if climate models are not constructed independently and are likely to share systematic sources of error, it really does seem unreasonable to assume that current climate model ensembles can be thought of representing anything like a random sample from Ω_{ce}. In sum, how to construct, and explore, a representative multi-model ensemble is by and large an open question.

4. Bottom-Up Justifications

As noted previously, bottom-up justifications are approaches that don't attempt to mount a justification based on properties of Ω as a whole, but rather see support being built up one model at a time. In other words, bottom-up approaches make no "detour" via a complete ensemble and see members of our ensemble supporting the conclusion directly. In this section we discuss three different approaches of this kind.

4.1 The Likelihood Approach

Parker (2011, 590) considers the following Bayesian argument for why agreement across models should increase confidence in the common result *H*:

> Premise 1: Proposition *e* warrants significantly increased confidence in predictive hypothesis *H* if $p(e|H) >> p(e|{\sim}H)$.
>
> Premise 2: Take *e* to be saying that all the models in this ensemble indicate *H* to be true.

> Premise 3: The observed agreement among models is substantially more probable if H is true than if H is false; that is, $p(e|H) >> p(e|{\sim}H)$.
>
> Conclusion: e warrants significantly increased confidence in H.

The first premise follows directly from Bayes' theorem and the second one is simply a statement of robustness and hence is assumed for the sake of argument. We call this the likelihood approach because the soundness of this argument crucially depends on whether the third premise (that the likelihood of H given data e is substantially greater than the likelihood of $\sim H$ give the same data) can be plausibly justified in reference to today's climate model ensembles. Parker, herself, doesn't think so, for she worries that there are many reasons why climate models might all indicate the truth of a predictive hypothesis, despite it being false:

> First, there are climate system features and processes—some recognized and perhaps some not—that are not represented in any of today's models but that may significantly shape the extent of future climate change on space and time scales of interest. In addition, when it comes to features and processes that are represented, different models sometimes make use of similar idealizations and simplifications. Finally, errors in simulations of past climate produced by today's models have already been found to display some significant correlation (see, e.g., Knutti *et al.* (2010); Pennell (2011)). Thus, in general, the possibility should be taken seriously that a given instance of robustness in ensemble climate prediction is, as Nancy Cartwright once put it, "an artifact of the kind of assumptions we are in the habit of employing" (1991, 154). Perhaps with additional reflection and analysis, persuasive arguments for $p(e|H) >> p(e|{\sim}H)$ can be developed in some cases, but at present such arguments are not readily available (Parker 2011, 591).

Parker's worry is that models might tend to indicate the truth of a hypothesis H because they share similar idealizations and simplifications, despite H not being true, and hence it is hard to justify the assumption that $p(e|H) >> p(e|{\sim}H)$. One might suggest that a satisfactory measure of model independence could help address Parker's concern. The idea here might be the following: if we can show that models in an ensemble do not involve similar idealizations and simplifications, then we might be able to alleviate the worry that models agree merely because they are similar – and hence we might be in a better position to justify Premise 3.

But a little reflection shows that things are not as straightforward as they may seem. As Parker (2006, 363) notes, climate models in a multi-model ensemble "often incorporate conflicting assumptions about what the climate system is like". And, arguably, the more dissimilar models are, the more conflicting assumptions one might expect them to incorporate. But if this is right then one might plausibly worry that having highly dissimilar models in an ensemble merely replaces one worry with another, as far as the above argument is concerned. For although we no longer have to worry that models might agree because they share similar simplifications and idealizations, we now have to ask why models agree despite making conflicting assumptions about what the climate system is like. In other words, given that the models make conflicting assumptions about the climate system and hence "the models are […] incompatible with respect to ontology" (*ibid*., 364), why should we expect that the models are more likely to agree regarding the truth of a hypothesis on the assumption that the hypothesis is true, rather than on the assumption that the hypothesis is false? If anything, the knowledge that models agree despite making incompatible assumptions about the target system might suggest that the models are agreeing for reasons that are independent of what the climate system is like.

4.2 Independence Approaches

The idea that, in order for a robust result to be reliable, the various different means of access should be reliably independent goes all the way back to Levins' "independent lies" (see Section 1 of Part I) and has been elaborated by Wimsatt (1981) who argued that "we feel more confident of objects, properties, relationships and so forth that we can detect, derive, measure or observe in a variety of independent ways because the chance that we could simultaneously be wrong in each of these ways declines with the number of independent checks" (*ibid*., 196).

According to Kuorikoski *et al*. (2010), Wimsatt's view of what it takes for a robust result to be reliable is relevant for assessing the epistemic import of model robustness. This is because, according to them, when models in an ensemble include the same (realistic) core assumptions *C* about the target system, but different simplifying or "tractability" assumptions (assumptions that we know to be strictly false about the target system), it is reasonable to assume that the probability that each model has to reach the correct result is *independent* of whether or not the other models reach the correct result, since "the modeler should have no positive reason to believe that if one tractability assumption induces a certain kind of error (due to its falsity) in the result, so does another" (Kuorikoski *et al.* 2010, 562). In light of this, they argue that if the same result *R* can be derived from several models involving the same substantial assumption *C*, but different tractability assumptions, this should rationally increase our confidence in the robust theorem. Their argument can be reconstructed as follows:

> Premise 1: Models that share a common core *S* (and satisfy conditions *C*) but involve completely different tractability assumptions can be assumed to be reliably independent. That is, conditional on the hypothesis that result *R* holds (or does not hold) in any target that satisfies conditions *C*, one's confidence that a model will reach *R* is not affected by whether the other models reach *R* or not.
>
> Premise 2: If models that share a common core *S* (and satisfy conditions *C*) are reliably independent, one's confidence in the robust theorem "in any target that satisfies conditions *C*, *S* brings about *R*" should rationally increase as the number of models that agree on result *R* increases.
>
> Conclusion: When models in an ensemble that share a common core *S* (and satisfy conditions *C*) but involve completely different tractability assumptions agree on a result *R*, this should rationally increase one's confidence in the robust theorem "in any target that satisfies conditions *C*, *S* brings about *R*".

Several commentators have questioned the soundness of this argument. For an extensive assessment of those arguments see Harris' (2021b); see also Odenbaugh and Alexandrova's (2011) for some earlier objections to Kuorikoski *et al.*'s (2010) argument and Kuorikoski *et al.* (2012) for some replies. However, we set these concerns aside at this point, and address the crucial question of the relevance of this argument to most (if not all) realistic cases of model-based robustness analysis. Notice that if models in the ensemble do not differ in all or most of their tractability assumptions, then there is, according to the current approach, no reason to assume that the models are reliably independent; for if models share several tractability assumptions (which are potential sources of unreliability), then "discovering that one of the models is unreliable should often greatly increase our confidence that the other is too"

(Schupbach 2018, 283). As discussed earlier, current climate models do share many similar idealizations, simplifications, and uncertain factual assumptions. Hence the models in an ensemble of climate models simply will not be independent in the sense of Premise 1.

There are notions other than being reliably independent (in the sense of Premise 1) to which one could appeal to motivate the epistemic import of model robustness. Confirmational independence (Fitelson 2001) is one of them. Indeed, Lloyd (2009) seems to appeal to this notion of independence in her attempt to justify the epistemic import of model robustness (see Justus (2012), Vezér (2016) and Harris (2021b) for some attempts to reconstruct and evaluate Lloyd's argument). However, similar concerns arise with this notion of independence. If models share similar idealizations and simplifications, it is at best unclear why we should expect their results to be confirmationally independent regarding a hypothesis.

One might suggest that a satisfactory measure of inter-model dependencies – one which, as discussed in Part I of this review (Section 3), climate scientists are currently invested in finding – could help us address these concerns with independence approaches. The idea here could be something like the following: the more dissimilar models are from other models in an ensemble, the more reasons for believing that those models' results are reliably independent or confirmationally independent regarding a hypothesis. A problem with this idea, however, is that models' results are either reliably/confirmationally independent regarding a hypothesis, or they are not. That is, reliability independence or confirmational independence (as discussed in Kuorikoski *et al.* (2010) and Fitelson (2001)) are not matters of degree. Hence, greater dissimilarity across models, despite knowing that the models still share some idealizations, simplifications, and uncertain assumptions, does not seem to be enough to dismiss our worries about the independence approaches discussed in this section. It is worth noting, however, that there have been Bayesian attempts to introduce continuous notions of independence (see e.g. Claveau (2013) and Landes (2021)) to which one might try to appeal to motivate the epistemic import of model robustness and that might in turn make independence approaches more defensible. Hence, the identification of an adequate measure of inter-model independencies that would successfully address the concerns with independence approaches discussed in this section remains an open question.

4.3 The Explanatory Approach

Schupbach (2018) has offered an explanatory Bayesian account of RA to defend the epistemic import of model robustness. He agrees with the critical points made previously and submits that models in an ensemble can rarely (if ever) be assumed to be reliably independent or confirmationally independent regarding a hypothesis. However, he argues that this is not a problem, because models do not have to be independent in any relevant sense for their consensus to be epistemically significant.

According to Schupbach's explanatory account of RA, using an additional means to detect the same result can incrementally confirm an explanatory hypothesis concerning the target as long as its detection is able to rule out a competing explanation for that result left standing by the previous detections. These means of detections (which Schupbach calls *RA diverse*) do not have to be independent from one another in any relevant sense for this to happen. To illustrate why this is, Schupbach considers the example of Brownian motion mentioned in Part I of this review (Section 1). When Brown first observed the curious motion of sample of pollen granules suspended in water, Einstein's molecular explanation for this motion was not the only viable one. The motion might have been due to currents or evaporation of the water, or a sexual drive inherent in pollen etc. But according to Schupbach, each new detection of this motion (using different materials or different media or different means of suspending the particles etc.) was able to rule out a competing explanation for this motion not yet ruled out by previous means of detection, and in so doing could incrementally confirm Einstein's molecular explanation. Schupbach argues that distinct means of detection do not have to be independent in any sense discussed in the previous section for them to be able to rule out competing explanations for the robust result. Consider, for instance, the competing explanation that the motion is exclusively due to a sexual drive inherent in pollen. By additionally detecting this motion using an inorganic material (instead of granules of pollen), one can rule out this competing explanation. However, these two means of detections are not reliably independent since "both could be misleading us due to the way particles are being suspended, due to the use of the same medium, due to the use of the same environmental conditions surrounding the apparatus, and so on" (Schupbach 2018, 283) - and for similar reasons they are also not confirmationally independent regarding Einstein's molecular explanation.

Schupbach suggests that his explanatory account of RA can also shed light on the epistemic import of model robustness, in particular. By considering multiple models that "may be quite similar apart from some modest differences in their simplifying assumptions" (Schupbach 2018, 289) and observing that they all reach the same result, one is able to rule out competing explanations for that result, thereby incrementally confirming the target explanation for that result. In his recent book, Winsberg (2018) argues that Shupbach's explanatory account of RA can shed light on the epistemic import of model robustness in climate science (albeit with some qualifications):

> Whether or not an ensemble of models is a good candidate for lending strong support for a hypothesis via RA depends almost entirely on the extent to which the set of models suffices for ruling out competing hypotheses. This means that just because the set of procedures we have that detect H are RA-diverse does not imply that we should have confidence in H. RA-diversity only implies CEP [cumulative epistemic power], i.e. it only implies that you are headed down the road to acceptance as you increase the size of the set of procedures. Once we know that a set is RA-diverse the question of whether it is large enough to warrant acceptance of H, whether it is sufficiently RA-diverse, is a further question. And the answer to that further question will always be a matter of judgment, context, considerations of inductive risk, etc. (Winsberg 2018, 194)

Winsberg's argument for the epistemic import of model robustness in climate science has been well received in the literature. According to O'Loughlin (2021, 36), "Winsberg (2018) convincingly argues that [Schupbach's account] can be applied to climate models." In reviews of Winsberg's book, Lusk (2019) writes that "Winsberg's argument is a convincing reconceptualization of robustness analysis in climate science" and Knüsel (2020, 116) that "Winsberg [. . .] makes a novel, convincing suggestion for when multiple sources of evidence in favor of a hypothesis are meaningful in climate science". However, despite this positive reception, there are some reasons to be wary of Schupbach's explanatory account's ability to shed light on the epistemic import of model robustness in climate science.

The first is a fundamental concern. As Harris (2021a) argues, there is an important difference between empirically driven RAs and model-based RAs, which may affect the applicability of Schupbach's account to the latter. In empirically driven RAs, the various detections of a result (e.g. Brownian motion) are physical measurement processes taking place in the actual world, and the hypothesis that we want to confirm (i.e. Einstein's molecular explanation), which also concerns the actual world, is a possible explanation of these detections. By contrast, in model-based RAs the various distinct detections of a result are all operations in "model land" and since the hypothesis that we want to confirm (e.g. a particular climate hypothesis) concerns the actual world the hypothesis is not a possible explanation for these detections. This does not necessarily imply that Schupbach's account is not applicable to model-based RAs in general, but it does nonetheless show that any attempt to successfully apply it will have to acknowledge this difference and show that it can be applied in spite of it. This point is not addressed in Winsberg's use of Schupbach's account of RA diversity in the context of climate model ensembles, and so questions about the applicability of this account in the context of climate remain.

The second concern is of a practical nature. As O'Loughlin (2021, 37) remarks "because climate scientists may engage in robustness inferences that are not focused solely on pinning down the value of a climate variable and that do not include the elimination of competitor hypotheses, we should be critical of the notion that [RA diversity] applies generally across all cases of RA in climate science". Indeed, this is an understatement. Most current multi-model ensembles in climate science are not intended to rule out specific explanations for a result, nor is it clear how current ensembles could be used in this way. It is also an open question how climate model ensembles that would serve this purpose would have to be designed in practice. Furthermore, it is hard to reconcile climate scientists' current efforts to find a measure of inter-model dependencies (which stem from the view that *independence* is what matters for choosing the best model ensembles) with an approach that sees the role of multi-model ensembles to be that of eliminating specific competing explanations for a result.

The third is an epistemic concern. Under Schupbach's account the extent to which the target hypothesis is confirmed is partly determined by how plausible the rival hypothesis is prior to elimination (2018, 293-96). Hence, in the case of model robustness, the extent to which the target hypothesis will be confirmed would have to partly depend on the agent's knowledge and beliefs about the derivational relationships in a family of models. This gives rise to two worries. First, knowledge and beliefs about the derivational relationships in a family of models can vary substantially from agent to agent. Hence, although non-omniscient agents might agree that some models' results are RA diverse, they might nonetheless strongly disagree about the extent to which this should confirm a hypothesis. Hence, the extent to which a target hypothesis is confirmed is bound to be highly contextual. Second, the extent to which a target hypothesis is confirmed seems also very difficult to assess within a given context, since it requires agents to assess their own knowledge and beliefs about the various derivational relationships in a family of models: evidently not an easy task. Although Winsberg acknowledges that "[o]nce we know that a set is RA-diverse the question of whether it is large enough to warrant acceptance of H, whether it is sufficiently RA-diverse, is a further question" there is here an implicit and questionable assumption that scientists are in fact able to assess the extent to which the target hypothesis is confirmed by an RA-diverse set of models in the first place.

5. Conclusion

In this chapter we have reviewed different ways in which the inferences drawn in MBRA can be justified. The sober conclusion can only be that multi-model situations raise issues that are not yet fully understood and that MBRA has not yet reached a stage of maturity. Important questions remain open, and these will have to be addressed in future research. This marks a juncture where two options are available. The first option is to tackle the issues we have discussed head on with the aim of formulating a version of MBRA that does not suffer from the difficulties discussed. The other option is to abandon MBRA and explore alternatives. Alternatives that one might explore at this point are perspectivism (Giere 2006; Massimi 2022), modal modelling (Katzav 2014; Sjölin Wirling and Grüne-Yanoff 2021; Massimi 2022) and a programme focussed on managing severe uncertainty (Bradley and Steele 2015; Roussos et al. 2021), but there may well be others.

References

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.

Bradley, R., & Steele, K. (2015). Making Climate Decisions. *Philosophy Compass*, *10*, 799-810.

Cartwright, N. (1991). Replicability, Reproducibility, and Robustness: Comments on Harry Collins. *History of Political Economy*, *23*, 143-155.

Claveau, F. (2013). The Independence Condition in the Variety-of-Evidence Thesis. *Philosophy of Science*, *80*(1), 94-118.

Fitelson, B. (2001). A Bayesian Account of Independent Evidence with Applications. *Philosophy of Science*, *68*(3), S123-S140.

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

Frigg, R., & Nguyen, J. (2020). *Modelling nature: An opinionated introduction to scientific representation*: Springer.

Giere, R. N. (2006). *Scientific perspectivism*. Chicago and London: University of Chicago Press.

Harris, M. (2021a). Conceptualizing uncertainty: the IPCC, model robustness and the weight of evidence. *PhD thesis*, London School of Economics and Political Science.

Harris, M. (2021b). The epistemic value of independent lies: false analogies and equivocations. *Synthese*, *199*(5-6), 14577-14597.

Houkes, W., & Vaesen, K. (2012). Robust! Handle with Care. *Philosophy of Science*, *79*(3), 345-364.

Hume, D. (1748/2007). *An enquiry concerning human understanding*. Oxford: Oxford University Press.

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.

Knüsel, B. (2020). Philosophy and Climate Science. *Ethics, Policy & Environment*, *23*(1), 114-117.

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.

Kuorikoski, J., Lehtinen, A., & Marchionni, C. (2012). Robustness analysis disclaimer: please read the manual before use! *Biology and Philosophy*, *27*, 891–902.

Landes, J. (2021). The variety of evidence thesis and its independence of degrees of independence. *Synthese*, *198*(11), 10611-10641.

Levins, R. (1993). A Response to Orzack and Sober: Formal Analysis and the Fluidity of Science. *The Quarterly Review of Biology*, *68*(4), 547–555.

Lloyd, E. A. (2009). Varieties of Support and Confirmation of Climate Models. *Aristotelian Society Supplementary Volume*, *83*(1), 213-232.

Lusk, G. (2019). Eric Winsberg's Philosophy and Climate Science. *BJPS Review of Books*. URL=<http://www.thebsps.org/reviewofbooks/lusk-on-winsburg/>.

Massimi, M. (2022). *Perspectival Realism*. New York: Oxford University Press.

O'Loughlin, R. (2021). Robustness Reasoning in Climate Model Comparisons. *Studies in History and Philosophy of Science*, *85*, 34-43.

Odenbaugh, J., & Alexandrova, A. (2011). Buyer beware: robustness analyses in economics and biology. *Biology and Philosophy*, *26*, 757–771.

Orzack, S. H., & Sober, E. (1993). A Critical Assessment of Levins's The Strategy of Model Building in Population Biology (1966). *The Quarterly Review of Biology*, *68*(2), 533-546.

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

Parker, W. S. (2009). Confirmation and adequacy-for-purpose in climate modelling. *Aristotelian Society Supplementary Volume*, *83*(1), 233-249.

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

Parker, W. S. (2020). Model evaluation: An adequacy-for-purpose view. *Philosophy of Science*, *87*(3), 457-477.

Pennell, C., & Reichler, T. (2011). On the Effective Number of Climate Models. *Journal of Climate*, *24*(9), 2358-2367.

Räisänen, J. (2007). How reliable are climate models? *Tellus A: Dynamic Meteorology and Oceanography*, *59*(1), 2-29.

Roussos, J., Bradley, R., & Frigg, R. (2021). Making confident decisions with model ensembles *Philosophy of Science*, *88*(3), 439-460.

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

Sjölin Wirling, Y., & Grüne-Yanoff, T. (2021). The epistemology of modal modeling. *Philosophy Compass*, *e12775*, 1-11.

Vezér, M.A. 2016. Computer models and the evidence of anthropogenic climate change: An epistemology of variety-of-evidence inferences and robustness analysis. *Studies in History and Philosophy of Science*, *56*, 95-102.

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

Wimsatt, W. C. (1981). Robustness, Reliability, and Overdetermination. In M. B. Brewer, & B. E. Collins (Eds.), *Scientific Inquiry and the Social Sciences: A Volume in Honor of Donald T. Campbell*. San Francisco: Lexington Books, 123-162.

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