THE EPISTEMIC VIRTUE OF ROBUSTNESS IN CLIMATE MODELING

Parjanya Joshi

M2017CCSS007

A project report submitted in partial fulfilment of the requirements for

the Degree of Master of Arts in Climate Change and Sustainability Studies

Centre for Climate Change and Sustainability Studies

School of Habitat Studies

Tata Institute of Social Sciences

Mumbai

2019

DECLARATION

I, Parjanya Joshi, hereby declare that this dissertation entitled 'The Epistemic Virtue of Robustness in Climate Modeling' is the outcome of my own study undertaken under the guidance of Dr. Tarun Menon, Assistant Professor, Centre for Science, Technology and Society, School of Habitat Studies, Tata Institute of Social Sciences, Mumbai. It has not previously formed the basis for the award of any degree, diploma, or certificate of this Institute or of any other institute or university. I have duly acknowledged all the sources used by me in the preparation of this dissertation.

Parjanya Joshi

M2017CCSS007

15 March 2017

CERTIFICATE

This is to certify that the dissertation titled “The Epistemic Virtue of Robustness in Climate Modeling” is the record of the original work done by Mr. Parjanya Joshi under my guidance. The results of the research presented in this dissertation have not previously formed the basis for the award of any degree, diploma or certificate of this or any other university.

Date: March 15, 2019

Dr. Tarun Menon,
Assistant Professor,
Centre for Science, Technology, and Society,
School of Habitat Studies,
Tata Institute of Social Sciences, Mumbai.

Contents

DECLARATION ii
CERTIFICATE iii
List of Figures v
ACKNOWLEDGEMENT vi
Abstract: 1
Literature Review: 2
Introduction: 4
Chapter I – Robustness from Levins to Lloyd 5
Appeals to Robustness in climate science: 5
Challenges in climate modelling: 7
Robustness Arguments: 8
The LWWOL Approach: 9
Chapter II – Challenges to Robustness 17
Adequacy-for-purpose: 18
Model Agreement: What does it mean? 21
Climate Models and Inference to the Best Explanation: 25
Chapter III – Robustness post-LWWOL 29
Independence in Climate Science: 30
Probabilistic Independence: 31
Robustness and Ontic Independence: 32
Robustness Analysis as Explanatory Reasoning: 34
Beyond model agreement: Cumulative Epistemic Power 37
Robustness and Multi-Modal Evidence: Discordance and Relevance 39
Chapter IV – Robustness Analysis in Scientific Practice 43
Early Robustness Analysis in Climate Science: 43
Emergent Constraints: 46
Conclusion: 51
References: 54
Urkund Plagiarism Check Report 57

List of Figures

Figure 1: Confidence is highest with high agreement and robust evidence Source: IPCC AR5 Guidance Note on Uncertainty 5
Figure 2: Pdfs and likelihood functions for climate sensitivity based on various observational constraints.. 44

ACKNOWLEDGEMENT

I would like to thank my guide Dr. Tarun Menon for his utmost patience and for taking me under his guidance and helping me plan out and write this ambitious and unique dissertation project. I would also like to thank Dr. Tejal Kanitkar and Dr. T. Jayaraman who have always encouraged philosophical thinking, keeping a scientific temperament, and critical analysis. I would also like to thank Dr. Kamal Murari for entertaining my philosophical questions in a technical course on climate modeling since that is where I found my specific research topics. I would also like to thank my classmates, my friends, my parents, and my fiancé Rucha for their unflinching support and constant confidence in my abilities even at times when I was lacking it myself.

Abstract:

The aim of this dissertation is to comprehensively study various robustness arguments proposed in the literature from Levins to Lloyd as well as the opposition offered to them and pose enquiry into the degree of epistemic virtue that they provide to the model prediction results with respect to climate science and modeling. Another critical issue that this dissertation strives to examine is that of the actual epistemic notion that is operational when scientists and philosophers appeal to robustness. In attempting to explicate this idea, the discussion turns to arguments provided by Schupbach who completely rejects probabilistic independence in favour of explanatory reasoning, Stegenga and Menon who still see some value in probabilistic independence, and Winsberg who takes applies Schupbach's to climate science, going beyond models to involve multi-modal evidence. After an exhaustive discussion on these arguments, this dissertation attempts to provide a thorough and updated notion of robustness in climate modeling and climate science.

Literature Review:

The literature of robustness as an epistemic virtue begins in 1966 with Richard Levins who first introduced the idea with reference to population biology models. He laid the groundwork for the common causal core argument for robustness which was advanced later by Wimsatt, Weisberg, Odenbaugh, and Lloyd. According to Levins, if models that share a causal structure but run under varying assumptions and produce agreement in results, then the result is robust. Orzak and Sober (1993) wrote a direct response to Levins' notion of robustness – the inference to truth was not justified. They emphasize that the increase in confidence in the models and their predictions should come from empirical support for the models and that going simply on robustness, the models could all agree and produce a robust result that could very well be false.

Wimsatt (1981) proposes the idea of robustness as reliability independence – the independence of the probability of failure of our models. Weisberg (2006) expands the common causal core idea by detailing the conditions under which a result could be robust. The idea that a result was robust if models sharing a common causal core structure under varying, heterogenous conditions because this further attributed the result to the core structure and reduced the possibility of it being an artefact of certain conditions. Lloyd argues that models should not be judged only on their fit (or lack thereof) with observed data (as was done earlier, prior to AR5). Additional evidence and support for the model should be considered which includes model-data fit, variety of empirical evidence, and independent support for model components and parameters. For Lloyd this confirms the models in question.

In response to Lloyd's claim about confirmation, Parker (2009) argues that the instances of fit and variety of support do not confirm climate models in general as Lloyd claims. Rather, they confirm their adequacy for the specific purpose at hand because models are constructed process level fidelity in mind as opposed to a system level fidelity that would be present in ideal simulations. Lloyd (2015) proposes a new, unique form of robustness – model robustness that is based not only on the convergence of outcomes of a group of models, but also on independent support for assumptions, features, parameters, and other components of the model. Lloyd argues that model robustness is a better alternative for confirmation purposes than other versions of robustness and is also a form specifically tailored to climate science and climate modeling.

Outside of the LWWOL group, Woodward (2006) classifies robustness into four types: Inferential Robustness – Inference from data to hypothesis is robust over a set of competing assumptions, Derivational Robustness – Derivation or prediction of observed data is robust over a set of assumptions, Measurement Robustness – Reduction of error by repetition in independent contexts, and Causal Robustness – Robustness attributed to underlying causal mechanisms or structural relationships. Parker (2011) uses the LWWOL common causal core and Woodward's notions of robustness to evaluate three conditions: truth, confidence, and security and investigates if robustness arguments hold for climate models. This is very specifically for the case of multi-model ensembles used for climate projection purposes. Her conclusion: Robustness does not justify – an inference from prediction to (very likely) truth, increased confidence in the hypothesis in question, and enhanced security of claims to have evidence for the hypothesis.

The most recent work in robustness was done by Lloyd (2015), Schupbach (2016), and Winsberg (2018). For Schupbach, there must be a specific notion of robustness that scientists appeal to when treating agreement as confidence building. He incorporates this requirement into a criterion he calls "RA diverse" (where RA: Robustness Analysis) and evaluates the logic of the various types of probabilistic independence (the most common and popular notion) – unconditional, reliability, conditional, and patrial probabilistic independence according to this RA diverse criteria, detailing the different concepts of independence. Finally, he proposes his own argument of explanatory reasoning or explanatory robustness which seeks elimination of alternate competing hypotheses. As mentioned above, since these two robustness arguments appeared in the literature after Parker's 2011 evaluation, that line of argument has not been advanced to evaluate these two forms.

Winsberg's paper does two things: firstly, it moves the idea of robustness beyond just models and model agreement, to encapsulate multi-modal evidence. Secondly, he applies Schupbach's notion of explanatory reasoning-based RA-diversity to climate science. Stegenga (2009) and Stegenga and Menon (2017) point out the problems that multi-modal evidence can pose to robustness arguments that are usually made in science. These problems include questions regarding the nature and necessity of independence, whether we can even know that our means of evidence are truly independent, and how to proceed when multi-modal evidence does not agree.

Introduction:

According to climate scientists across the world, the agreement of model projections is a good thing. The robustness of multiple, consistent, independent lines of high-quality evidence increases the degree of confidence in the models and their predictions (IPCC, AR5 WG1, 2010). The aim of this dissertation is to comprehensively study various robustness arguments proposed in the literature as well as the opposition offered to them and pose enquiry into the degree of epistemic virtue that they provide to the prediction results with respect to climate science and modeling. Another critical issue that this dissertation strives to examine is that of the actual epistemic notion that is operational when scientists and philosophers appeal to robustness.

The first chapter of this dissertation deals with an initial discussion of the appeals to robustness in climate science as well as an account of the evolution of robustness from Levins (1966) to Lloyd (2015). The second chapter comprises a detailed account of arguments opposing this kind of robustness. The third chapter involves a discussion on the notion that is held be an important condition for robustness arguments – independence; as well arguments against the necessity of independence which espouse that explanatory power is the operational notion in robustness; the application of explanatory robustness to climate science beyond models, and the problems regarding the independence of means of evidence. The fourth and final chapter involves a discussion and critical analysis of some experiments and scientific results from climate science to examine which, if any, sort of robustness do they appeal to.

In the end, this dissertation argues that independence might not be as irrelevant to robustness as some, such as Schupbach (2016) purport, and that it supplements good explanatory reasoning. The problem is that climate models are not independent, in fact the various modes of evidence, as we shall see, are dependent on each other for tuning, filling in of gaps in the datasets, and so on. Hence, in climate science, independence is a very difficult requirement to fulfil, and scientists, in assuming the independence of various means, give rise to different kinds of "pseudorobustness" (Stegenga and Menon, 2017) which the phenomenon that is actually operational when robustness is appealed to.

Chapter I – Robustness from Levins to Lloyd

Appeals to Robustness in climate science:

Climate models, since they were first used in 1990 to establish a scientific basis for climate change in the First Assessment Report (AR1) of the Intergovernmental on Climate Change (IPCC), have become our most dependable resource for the prediction of future levels of warming, sea level rise, and socio-economic impacts resulting out of them. They have also grown in size, scope, purpose, predictive accuracy, and computational power; from simple zero-dimensional models such as radiative-convective models to complex, gridded, parametrized Generalized Circulation Models (GCMs) such as Earth System Models (ESMs).

ESMs are highly complex simulations of the Earth's climate system, simulating the various physical processes that occur on the surface and how they affect the state of the system. The fact that climate models have grown and developed so much since 1990 has increased the confidence of scientists and policymakers in the resulting output. It is not just the growth and increase in ability of the models that has predicated confidence, but also the agreement of different model groups on the results. In the latest of the IPCC's assessment reports, the Fifth Assessment Report (AR5), there is a guidance note on uncertainty which carries a figure depicting a rubric to be used to assess confidence in the certainty of a finding using agreement and robust evidence.

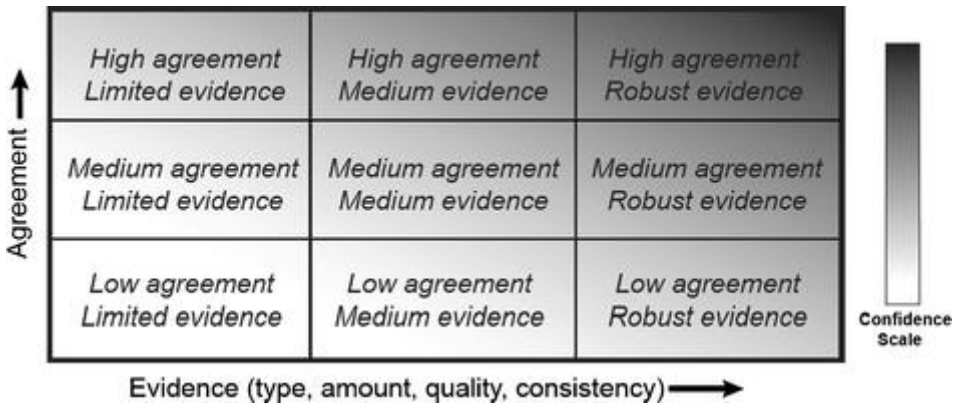

Figure 1: Confidence is highest with high agreement and robust evidence
Source: IPCC AR5 Guidance Note on Uncertainty

The "Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties" (hereby referred to as the Guidance Note on Uncertainty or simply the Guidance Note) sets the standard for assessing the results of climate models, the uncertainties in these results, and confidence in assessing climate hypotheses that

are made based on these results. The confidence regarding the hypotheses comes from the amount, quality, and consistency of the evidence, as well as the degree of agreement of evidence. More specifically, the note states that "generally, evidence is most robust when there are multiple, consistent, independent lines of high-quality evidence" (Mastandrea, et al. 2010).

Scientists are accordingly asked to communicate the results and uncertainties in a corresponding manner, providing measures of confidence in terms of robustness of evidence as well as a probabilistic expression of uncertainty. The logic behind this seems to stem from the notion of independence and variety of climate models that are used to simulate the climate. These models are run in ensembles, wherein multiple instances of the same model run under different conditions. They also include parametrizations and idealizations of processes that cannot be simulated because they are sub-grid (since simulation occurs at a minimum grid size), and in addition to this, they are subject to uncertainties. Hence, an agreement of results despite the diversity and independence is seen as a good thing, and as increasing confidence in the notion that climate models are accurately simulating and representing the target climate system.

In 2010, Pirtle et al. observed that for 118 authors in climate science literature since 1990, model result agreement inspired increased confidence in the results and hence in the simulating and representing ability of the models (Pirtle et al., 2010). Further, their research indicated that it was not just agreement between model results, but the emphasis was, in fact on agreement between independent sources of evidence. As Pirtle et al. note, the IPCC's Fourth Assessment Report (AR4) used the results of 23 different GCM as evidence of reported claims but did not study or discuss the details of the kinds and degrees of difference between models, nor did it provide any clarification on what kinds or degrees of independence could be appealed to, to make model agreement significant (Pirtle et al., 2010).

Pirtle et al. provide an elementary list of the different dimensions along which independence can be established, which include: basic physics, parameters, scope and idealizations, observed data, numerical coding/computer processes, and history and sociology (Pirtle et al, 2010). This list of dimensions is a hierarchy of degrees of independence. The implication here is that within each of these dimensions, there are different ways in which models can be diverse. The notion of independence presented here intuitively seems to be that the more diverse the models i.e. the more the dimensions they differ across, the agreement is proportionally valuable and significant.

Challenges in climate modelling:

Since the predictions of the models are for future years (e.g. 2100), there is no way to actually know the truth or falsity of the predictions of the various model ensembles until there is observed data to compare with at that point in the future. Models are simulations, and by the nature of the climate system and the global scope and complexity of climate change, they are not a 100% faithful replication of the system. Even the scientists acknowledge the presence of many large and looming uncertainties: epistemic, ethical, structural, initial conditions, boundary conditions, and parametric. The fact that our state-of-the-art models are gridded means that all data and processes used to simulate are at the grid-scale. On a sub-grid level, many processes (such as cloud formation) are parametrized and that leads to a loss in fidelity of representation. Yet, the models are not discounted from being our best option. In climate science, as in population ecology and other scientific fields, model agreement or robustness of evidence holds weight when it comes to the scientific community's convictions about the truth of those predictions.

Robustness is intuitive – the heuristic idea of diverse and independent sources providing results that agree. But it seems to neglect the intricacies of model building and assessment in favour of the agreement of results. Primarily because the climate models make predictions about the future by simulating the target system based on data, and because it is impossible to say whether a prediction is right or wrong without knowing the answer; we do not know with certainty whether or not any of the present models we have are true representations of the system (as floated by Levins). In this case, we will only know if the predictions for a future year are correct or true when we compare the model predictions to actual conditions in that year. This is an inductive inference which is accompanied by inductive risk, which Parker and Winsberg (2018) argue causes decisions about confidence and inference-making to be eventually guided by non-epistemic values. So, by virtue of not knowing whether a true model exists, we have no access to objective truth against which robust results can be evaluated as being true or false; then what about high-quality observed data that supports the simulations? What happens when the data is contrary to the model results?

Of course, the nature of the inductive step is still consistent when comparing results of an approximate simulation to empirical data. Since the data is that of past observations, one is making the inductive inference that because the models simulate the past accurately, the simulation of the future will be (approximately) consistent with the future truth. And that's

what essentially happens when models are assessed with past observations, reanalyses data, and tuned to provide robust results, it's just that the claim is flanked by considerations of uncertainty, measures of confidence in the form of probability and robustness. Models are packaged with uncertainties that range from errors in initial and boundary conditions to parametric to structural uncertainties. The parametrizations in various models come with their own uncertainties.

In addition, our data sets are not fool proof, in fact, we only have relatively precise data for various climate variables for the last 20 to 30 years because of satellite technology and increased number of weather monitoring stations. Prior to that, climate data is fraught with disproportionalities and lacunae which are bridged statistically. Data for earlier periods (i.e. before the industrial era) where there is little to no climate data is obtained indirectly through paleoclimatic reconstruction, from sources such as ice shelves, tree rings, and mineral analysis in rock layers. What precisely is robustness then? Is it just model agreement, or some heuristic that increases our confidence in our prediction power and theory? Is there any actual notion that one can say that scientists are appealing to when they appeal to robustness?

Robustness Arguments:

An examination of the philosophical literature on the property of robustness and how it is applied in scientific models reveals a rich landscape of ideas and perspectives. The idea of robustness in science, referred to in the literature as a robustness argument or robustness analysis (RA), was formally introduced by Richard Levins (1966) in relation to population biology and ecological modelling. The literature on robustness arguments revolves around two main issues, namely the notion of robustness and the epistemic virtue of robustness. The former concerns the nature of robustness and what it is about robustness that gives us increased confidence in our inferences, while the latter concerns the confidence and validity of the warrants made considering robustness.

Since Levins introduced the idea of robustness as model agreement in 1966, it was uncontested until Orzack and Sober critiqued it in 1993. It was later expanded to incorporate the critique by William Wimsatt (1981), Michael Weisberg (2006), Jay Odenbaugh (2011), and Elisabeth Lloyd (2010) who argued for model agreement to be supplemented by empirical evidence in what is known as the LWWOL approach. James Woodward (1981, 2006) also surveys four other notions of robustness namely, inferential robustness, derivational robustness, measurement robustness, and causal robustness. Jonah Schupbach (2016) argues

for an explanatory notion of robustness analysis, an idea that Eric Winsberg (2018) builds on by proposing that robustness analysis has epistemic virtue beyond model agreement as previous scholars had argued.

The second issue is where the validity of the inferences made based on robustness analysis is called into question. Orzack and Sober's critique to Levins' essentially asserted that a set of models could have a robust result but still be false. This line of argument – that robustness does not justify the inference to truth, has been backed by many including Wendy Parker (2011) and Joel Katzav (2012). In truth, the separation between the two issues is not as stark as one may posit, in fact one could argue that one must address the issue of the notion of robustness in order to address the issue of epistemic virtue. The expansion and development of robustness arguments includes rejecting notions with less epistemic virtue in favour of stronger notions and hence it seems useful to approach the literature in a way that synthesizes the two issues into one.

The LWWOL Approach:

When Levins first published his idea of robustness in the context of models in 1966, he described the truth to be, "an intersection of independent lies" (Levins, 1966), where "lies" refer to the various ways in which experimenters, modelers, and scientists in general could be deceived, misled, or otherwise be caused to make incorrect conclusions. By doing so, Levins asserts that robustness across independent means of testing is inference for truth. For Levins, if multiple different models describing the same natural system produced a robust result, it was evidence that they had some common causal core, and that it was not just an artefact of statistics, perspective, or features of individual models (Wimsatt, 1981). Further, Levins asserts that robust theorems can reveal truths about nature, or that robustness can be evidence for truth.

This idea went unnoticed and uncontested till Orzack and Sober's critical assessment of Levin's work in 1993. They rejected Levins' assertion regarding robustness and truth and argued that this inference to truth is not justified. According to Orzack and Sober, the models could all agree i.e. produce a robust result, but it could still be a false result. In Levins' paper, he argues that if a set of models **M** produces a robust result **R** (i.e. the models are deemed independent and produce agreement in results), then the scientist(s) must know that one of the models in the set is a true representation against which the robust result can be compared to make a truth claim. The dilemma here is that if the scientists do not know that there is some true representation in **M**, then there is no reason to believe that **R** is true, and if there is a true

representation in set **M** that is known to them, then there is no need for a robust result **R** (Orzack and Sober, 1993).

Further, they emphasize that while robustness has a positive confirmatory virtue, it alone cannot be the basis of justification for the inference to truth. They argue that there needs to be more focus on empirical evidence and support for the various models in the set, in addition to the set producing a robust result. For Orzack and Sober, Levins' conception of robustness is confirmatory only in special cases where the true representation is known to us. Levins' rebuttal to Orzack and Sober argued that there are in fact, empirical aspects of robustness and he stresses on the empirical evidence in support of the common causal core and assumptions (Levins, 1993), but as noted by Lloyd, Levins' style of RA does not include empirical support for prediction/retrodiction (Lloyd, 2015). This critique was incorporated by the LWWOL group into what came to be known as the common causal core approach.

Wimsatt's notion of robustness is one of reliability, in that "the probability of failure of the different means of access should be independent" (Wimsatt, 1994) which Schupbach asserts is more accurate interpretation of Levins' notion of "independent lies" (Schupbach, 2016). Essentially, the probability of failure yields a measure of reliability of the means of detection and the independence condition means that if a means of detection fails, then it does so independently of others, i.e. its failure has no effect on the probability of failure of other means. Michael Weisberg's account of robustness involves the outcome or property that is deemed robust and structural elements shared by a set of models. His RA logic is as follows, "Ceteris paribus, if [common core (causal) structure] obtains, then [robust property] will obtain" (Weisberg, 2006). It is with this conception of robustness that we see the development of a form of RA tailored for climate models, more precisely General Circulation Models (GCMs). Weisberg's RA also holds independence of means of detection to be central to the argument, although in the case of climate models, the notion of independence is different.

Firstly, since climate models are representations of complex and dynamic systems, even a slight change in parameters or initial conditions may yield a drastically different result, which is known as Sensitive Dependence on Initial Conditions or the butterfly effect (Frigg, 2014). Hence, one could run single-model ensembles with different background assumptions and parameter values (that is, one model and heterogenous background). Secondly, GCMs are run in multi-model ensembles (MMEs) where models constructed and developed by different groups are run together, either for the same time period or under the same conditions (that is, different models and same background). Either way, this is sufficiently heterogenous for Weisberg. Justus (2016) asserts that Weisberg's analysis does away with the independence

requirement and in fact, recognizes the dependencies between models. He writes, "and in part because the independence assumption has been abandoned, robustness' epistemic payoff is clear: robust theorems link the empirical support of **C** [common causal core] and **R** [robust property]." (Justus, 2016).

For Weisberg, "robustness does not confirm robust theorems; it identifies hypotheses whose confirmation derives from the low-level confirmation of the mathematical framework" (Weisberg, 2006) which demonstrates a shift from Levins' confirmatory conception. Hence for Weisberg, it is not the robustness that warrants confirming robust theorems, but rather the mathematical framework whose confirmation warrants the classification of the relationship between **C** and **R** as one of causal dependence (Weisberg, 2006). While some philosophers like Lloyd (2011) see this as implicitly arguing for evidence supporting model components such as assumptions, idealizations, parametrizations, and so on. Others such as Justus (2016) question Weisberg's claim that the mathematical framework achieves this level of confirmation if it "can adequately represent the phenomenon of interest" (Weisberg, 2006). For Justus, the modality of whether the framework "can" or "does" adequately represent is problematic. The mathematical framework is comprised empirically interpreted mathematical structures whose function is to adequately represent. Justus argues that the confirmation lies in the performance of the structures and not that of the framework they are used in. Additionally, he argues that since confirming the entire framework increases the probability for all structures, this entails an increase in probability for inadequate mathematical structures (Justus, 2016).

Jay Odenbaugh's argument in this matter concerns model idealizations and what effect these idealizations have on our confidence in the model. According to Odenbaugh (2011), idealizations are false propositions that are constructed into a model for the purposes of science qua science, in that they do not serve any explanatory or confirmatory purpose. The problem, for Odenbaugh, is that since we know that our models are imperfect and have idealizations, a confirmed prediction from these models should not increase our confidence in them. We also have no way to know whether the non-idealized parts of the model can produce the same confirmed result without the idealizations (Odenbaugh, 2011). For example, in climate science, cloud formation parametrizations are the most common idealizations in GCMs. We do not know how exactly to simulate cloud formation in a GCM grid cell since it occurs at a sub-grid level and hence, we include a parametrized value. Now, we do not know what the right value is, but we know what range we need to include to allow the model to simulate 20th Century warming accurately. Further, we also know that eliminating the cloud formation parametrizations from the model entirely will result in an inaccurate representation of the

system. Odenbaugh's argument for robustness analysis is that it allows for the varying of parameter values and elimination or "discharge" of idealizations across a set of models, allowing more confidence to be induced in the hypothesis that the non-idealized part is the causal structure that leads to the confirmed prediction (Odenbaugh, 2011). Odenbaugh then considers the problem that idealizations pose to a realist perception of science. In case we do not have said independent support, then robustness cannot placate or discharge our lack of confidence in the model's prediction. Hence Odenbaugh's case is clear: robustness analysis is only useful in increasing confidence when we have independent support for the core structure of a model, an idea that Elisabeth Lloyd refines and expands on, tailoring it specifically for climate models.

Lloyd notes that Weisberg's notion of RA is a realist one (Lloyd, 2015) and that to him, the robust result, in light of the common causal core (which in the case of climate models is the greenhouse gas forcing mechanisms) makes it highly likely that the actual climate system also has a corresponding causal structure, or a Justus (2016) calls it, a two-way flow of empirical support: C => R in the model while R => C in the target system (Justus, 2016). Lloyd writes that while the common causal core approach insists and advocates for the inclusion of empirical evidence and independent support for models, not much has been written about the aforementioned issue of the epistemic virtue of robustness. To properly understand Lloyd's motivations and contributions to the LWWOL group, it is important to examine the background of epistemological background of climate model predictions and the motivations behind Lloyd's complex empiricist view.

As Lloyd notes, there is a debate regarding the objectivity and independence of the various kinds of data used in climate science (Lloyd, 2015). The debate occurs between the direct empiricists who believe that if a model is contrary to the observed data, then the data overrules the results, and the complex empiricists who argue firstly that observed data is itself theory-laden and loaded with assumptions and conventions and hence cannot be used to invalidate model result (Lloyd, 2015). As seen in the case of Spencer and Christy (1990) whose research found that satellite data and model results diverged when it came to predicting temperature trends for the 1979-1988 period, the satellite data was given preference because it was validated by radiosonde data (Spencer and Christy, 1990) and "radiosonde data were taken at face value, as representative measures of the true temperature of the atmosphere" (Lloyd, 2015). The complex empiricist camp argued that radiosonde datasets are not an "unambiguous gold standard" (Santer et al., 2003) because they come parcelled with methodological nuances.

For the complex empiricists, the radiosonde datasets are "subject to a host of complications, including changing instrumentation types, configurations, and observation practices... making long-term climatological studies difficult" (Mears et al., 2003).

Another objection raised against the direct empiricists' "validation" of satellite data was also pointed out by Santer et al. (2003) was that since the radiosonde data had previously been used to calibrate and build the UAH satellite datasets, the validation is ultimately circular (Santer et al. 2003). The complex empiricists on the other hand, when faced with the same situation viz. the disagreement between observed data and model results, propose a more thorough investigation and comparison rather than just goodness-of-fit. Lloyd's sketch of the complex empiricist argument pushes for a stronger emphasis on the strengths of climate models while advocating for more testing and checking for evidence that supports the model. Simply dismissing a model because its results do not fit well with observed data (which has its own discrepancies) is an unfair judgement which downplays the strengths of that model (Lloyd, 2015).

Unlike scientific realism which is the view that the aim of science is truth, antirealism (of which complex empiricism is a subcategory) is the view that the aim of science is empirical adequacy (van Fraassen, 1980). For Lloyd, the aim of climate models is not to predict truth, because of our "human finitude" (Lloyd, 2015) limits us: We do not know whether there is a true simulation or representation of the climate system to compare robust results against, neither do we have a fool-proof, airtight, independent dataset that can be used to validate a robust result. This also means that we do not know what the right or true way is to simplify or idealize the system (Lloyd, 2015 and van Fraassen, 2010). The sources of data argued for by the direct empiricist are neither completely objective, nor are they independent since they are used to tune other sources of data or are themselves tuned.

Lloyd addresses such concerns as a guarantee of truth or model completeness from non-LWWOL philosophers such as Orzack and Sober, Woodward, and Parker as "scientifically unrealistic" (Lloyd, 2015). Hence, the aim of scientists constructing climate models, according to Lloyd, is to make choices and represent the system the best they can. The idea that this construction, the choices made in representing and assessing results are guided by non-epistemic values is discussed in a later chapter. The world currently has many different climate model families with different parametrizations, assumptions, and idealizations as well as core structures. It is this heterogeneity across which Lloyd argues that her form of robustness is confirmatory and confidence inducing.

Although this approach to robustness is explicitly confirmatory, the kind of confirmation that Lloyd argues for in the case of climate models is different from the way that climate scientists and other philosophers who have worked on robustness have conceptualized it (Lloyd, 2015). Lloyd delineates Jim Woodward's measurement robustness (Woodward, 2006) from her own model robustness. According to Lloyd, measurement robustness or "heuristic robustness" is focused on prediction and retrodiction and is used by scientists to gain confidence in the results of their models, that is, to search for a strong basis for their prediction/retrodiction capabilities. In the case of climate science, Woodward's concept of measurement robustness is echoed in Mastrandrea et al. (2010), as well as the IPCC's Guidance Note on Uncertainty in AR5, as stated at the beginning of this chapter. For Lloyd, her version of robustness looks beyond building confidence in prediction and retrodiction. Model robustness is focused on confirming the causal explanation that leads to the prediction/retrodiction result unlike measurement robustness which is focused on confirming the effect (Lloyd, 2015).

One could use a Woodward-style robustness analysis by treating each model as an independent method of measurement, but it would quite constrained and not especially useful (Lloyd, 2015). Concurring with Weisberg (2006), Lloyd states that "many climate models are variants of others, and not usefully independent for these purposes, thus placing tight constraints on the climate-model-version of measurement robustness" (Lloyd, 2015) and that the manner in which physical experiments might be distinct differs from the manner that climate models may be independent because distinct physical experiment are causally independent and their errors and uncertainties may also be different and specific to that apparatus and conditions. Climate models or GCMs on the other hand share many core elements including material and energy flow equations and assumptions about causal mechanisms for radiative forcing (e.g. GHG forcing) while differing in the sub-grid parametrizations such as cloud formation and other idealizations (Lloyd, 2015).

Discarding the notion of independence in favour of the criteria of heterogeneity (Weisberg, 2006), Lloyd further explicates how dependencies between models strengthen the confidence in the causal hypothesis that increases in GHG emissions have caused a majority of 20th Century warming (Lloyd, 2015). "In sum, **T** is a robust result under the combination of the variety of assumptions and parameterizations, **Ai**s, which are themselves usually empirically supported, combined with any individual **Mi**, which includes the **GHG** causal radiative core, (**Mi** & **Ai**s)." writes Lloyd (2015), where **Mi** is any model in the set **M** (i = 1, 2, 3...) and **Ai** is the set of assumptions that **Mi** contains, and it includes cloud parametrizations

and ocean mixing parametrizations. The key insight here comes from the emphasis on combination, implying that if any **Mi** combined with any **Ai** yields **T**, then it is a robust result. In this way, Lloyd clarifies what she calls Weisberg's "implicit appeal to the variety of evidence" (Lloyd, 2009) when he advocates for more heterogenous model sets which still share causal cores.

Recall that Weisberg's argument is that robustness does not confirm robust theorems but identifies where in the mathematical framework does the low-level confirmation occurs (Weisberg, 2006). The robustness he is referring to here is what Lloyd would call the result or effect-focused measurement robustness, which does not serve a confirmatory purpose, but it is the first step in Lloyd's more comprehensive robustness. Lloyd (2015) writes, "we can thus identify the patterns of evidence that support a model-type and its causal core, while tracking the processes of reasoning used in climate modeling and the confirmation of climate models", which, in other words is the identification of the low-level confirmation in the underlying mathematical framework. Then **T** is not just a robust result but in fact, that **T** is robust across all combinations of **Mi** and **Ai**, each of which are supported by a variety of empirical evidence, also increases confidence in the **GHG** hypothesis.

Lloyd writes that model robustness is used in a nested fashion and that the set of assumptions **Ai** is also robust, which further makes the set **M** the strongest contender and most robust model type (Lloyd, 2015). Essentially, what gives model robustness its confirmatory virtue is the variety of support to the various parametrizations and idealizations encompassed by **Ai**. Without this empirical support, **T** is simply just a robust result with no confirmatory virtue. The variety of support for parametrizations in addition to **T** appearing to be robust over **Mi**, "builds confidence in the core's efficacy, accuracy and reality" (Lloyd, 2015).

In the next chapter, we shall discuss some of the refutations offered by other philosophers to the LWWOL formulation which has been the dominant strain of RA in the literature. The first set of refutations is offered by Wendy Parker who in response to Lloyd's claim that confirmatory robustness confirms the climate model itself, poses argument that models are constructed and evaluated in scientific practice with process fidelity in mind, and how that requires a shift in thinking – from the Lloyd's claim to the argument that confirmatory robustness only confirms hypotheses about the model's adequacy for specific purposes (Parker, 2009). Parker also questions the epistemological issues with robustness, specifically questioning the ability of the LWWOL type reasoning to give warrants to truth claims, induce an increase in confidence, and increase the security of these claims (Parker, 2011).

The second set of refutations comes from arguments provided by Joel Katzav (2012) who contends that GCMs used in climate science are hybrid models – models which have causal components that are explanatory and idealizations and parametrizations that are instrumental. As a result of this, hybrid models are not completely explanatory devices and whatever hypothesis is inferred from the model cannot be attributed solely to the explanatory parts. This argument can be extended further to contend that individual models in an ensemble confer these explanatory-instrumental virtues to the ensemble. Hence any degree of confirmation in a hypothesis derived from a Lloyd-style robustness analysis could suspect of being an artefact of the idealizations rather than the causal cores in the ensemble.

Chapter II – Challenges to Robustness

As we saw in Chapter I, the concept of robustness has gone through a process of evolution from Levins (1966) to Lloyd (2015), with there being objections, concerns, and challenges at every turn. Some of the initial and more general arguments against robustness, such as those posed by Orzack and Sober (1993), were explored earlier in Chapter I, while examining the concept of robustness. We will also explore in Chapter III, the arguments posed against the status-quo of robustness by Jonah Schupbach (2016). In this chapter, the goal is to explore three objections – or more correctly – explore two objections by Wendy Parker (2009, 2011) and argue that an epistemological position held by Joel Katzav (2012) undermines robustness in very much the same spirit as Parker's overall position.

The first objection Parker offers directly challenges Lloyd's claim that robustness has a confirmatory virtue by arguing that what robustness confirms is not the model (or ensemble) itself, but rather the hypothesis that the model is adequate for the required purposes of simulation. The second objection is a set of three arguments that downplay the assumed or apparent role of robustness in providing the following inferences: to truth of the hypothesis, to increased confidence in the hypothesis, and to increased security of the evidence. Both objections are based ultimately on the same premise: that we do not have a true representation of the climate system, nor do we know what simulating that would entail. Our models are highly specialized, indubitably sophisticated simulations but at the end of the day they are highly tuned approximations, which makes them flawed.

Katzav's position is not a direct objection to robustness, neither was it posed as one. It is a position that concerns the question of meta-explanation in climate models: Are the inferences made from climate models results abductive inferences, that is to say, are the warrants that we have in light of climate model results provided by Inference to the Best Explanation (IBE)? Katzav claims that it is not a case of IBE where hybrid models (in which he includes GCMs) are concerned because they are not completely explanatory – they have instrumental aspects as well. This means that the warrants provided are not because of the explanatory virtue of a model result, but its explanatory-instrumental virtue which Katzav argues is weaker.

Adequacy-for-purpose:

Wendy Parker's (2009) analysis of the robustness arguments hitherto mentioned begins with her making "adequacy-for-purpose" claims about climate models. Her major objection to Lloyd (2009) who provides the latest version of the LWWOL robustness argument described above, is that robustness does not confirm (a particular set of) climate models as a whole, but rather confirms specific hypotheses related to the adequacy of those models for the purposes of representing specific processes or simulating specific climate variables or phenomena. Additionally, Parker argues that this required shift in thinking, across the board, is difficult to bring about because model evaluation practices are steeped in their way of doing things (Parker, 2009).

Parker differs with Lloyd on what the epistemic goal of climate models is. Lloyd's realist argument is that evidence such as model-data fit and independent support for auxiliary assumptions or parametrization serves as accumulating evidence for the truth of the hypothesis (Lloyd, 2009). Parker, on the other hand, claims that since even our best climate models have "significant simplifications and approximations", the hypothesis that they can accurately embody a complex climate system is undoubtedly false (Parker, 2009). Further, she claims that this also precludes the hypothesis from being empirically adequate (not to be confused with adequacy-for-purpose) in the sense defined by van Fraassen (1980): that whatever the model says about observable entities and phenomena is true (Parker, 2009).

To establish this sense of adequacy-for-purpose, Parker claims that we will need to determine two things: i) what is most likely expected to be observed if the model is adequate for a particular purpose and ii) whether or not what is actually observed fits with the expectation, and the former is more difficult to determine. Additionally, she argues that if the model truly embodied the true or empirically adequate, as Lloyd claims, then determining (i) would be considerably easier than if the model is adequate for some explanatory or predictive purpose (Parker, 2009). This is because if we assume that the model is adequate for some purpose, any information about its adequacy (or lack thereof) for any other purposes or "information about what other properties the model is likely to possess" does not follow from the assumption (Parker, 2009).

The hypothesis that Parker gives as an example is: A climate model is adequate for the purpose of predicting global mean surface temperature (GMST) for each year of the 2050s within a margin of 0.3°C under the scenario that GHG emissions increase rapidly from the present till the 2050s. If this hypothesis is true, that the model is in fact adequate for this

purpose, then we need to determine what it is that we are most likely to observe. Parker claims that there are at least three confirmatory observations we can make by virtue of the model's adequacy-for-purpose: a) observations of model-data fit for GMST for years other than or prior to the 2050s, b) observations of model-data fit for variables other than GMST such as precipitation or wind speeds at various times, and c) properties of the model other than model-data fit.

Regarding the first observation, Parker says that if a model is adequate for accurately predicting GMST for the 2050s, there seems to be no reason to expect that it will fail to do the same for preceding years. Further, she argues that this holds even if the predictive accuracy of the model decreases with the passing of in-simulation time: that is, if the model is overtuned to current conditions or if the relationships between GHG forcing and other feedbacks are misrepresented such that as GHG emissions increase leads to an increased effect of the feedbacks. In this case, we can still reasonably claim that the model will predict GMST in years before 2050 at least as accurately or reliably as it will for the 2050s, a period for which it is adequate (Parker, 2009).

In the case of the second type of observation, Parker notes that it is important to realize that state-of-the-art climate models are constructed with process fidelity in mind, as opposed to system fidelity. This means that a climate model might produce reliable and accurate representations of specific processes but at the cost of proper representation (that is, through the use of parametrizations) of other processes. Hence it is important that we understand how the processes we are interested in, say global mean annual precipitation (GMAP) and GMST, are represented in the model in question and how this is different from our beliefs and understanding about the actual workings of the processes. This includes how the processes are related to each other in the model, whether those relationships are accurate to the actual relationships or if they are exaggerated or too idealized.

Similarly, for the third confirmatory observation, Parker is not optimistic of its viability in practice. The issue thus framed is, "What properties is the model's representation of process $\mathbf{P_1}$ likely to have if the model is adequate for the predictive purpose at hand, *given* that the model's representations of other processes $\mathbf{P_2, P_3, P_4, \ldots}$ have these other properties?". It is important to reaffirm here, that the properties referred to here are not the properties of the actual processes $\mathbf{P_1, P_2, P_3, \ldots,}$ but of the representations of these processes. As stated above, for models constructed in process fidelity in mind, adequacy-for-purpose will involve a "balance of approximations" (Lambert and Boer, 2001), which means that accurate and reliable

representation of other processes may be sacrificed for the adequate representation of the process which is the purpose at hand.

Since we don't know exactly in how many ways or how much our models differ from the target system, we do not know of or possess a perfect representation of our climate system; these questions about representation are difficult to answer (Parker, 2009). Thus, the underlying issue is the same: Since we are in an information deficit and the problem of unknown unknowns looms, there are very few instances in real-world climate model assessment where we can say with confidence that we have enough information about misrepresented or unrepresented processes in a model to argue that it is adequate for a particular purpose. For most cases, Parker (2009) argues, "it is often difficult to collect evidence that can be said to clearly confirm or clearly disconfirm a climate model's adequacy for a predictive purpose, because there is no simple, general principle that can be applied to determine what we are likely to observe if the model is adequate."

By showing that adequacy-for-purpose of climate models is difficult to establish, let alone establish more lofty goals such as truth representation or empirical adequacy, Parker proposes that confirmation may not be the most useful route to follow. Something else noted by Parker (2009), and by Stegenga and Menon (2017) in the broader field of scientific research, is that robustness is sometimes formulated as a realist no-miracles argument. The idea that this agreement between models (or means of evidence in general), if it is not a result of plausible, approximately accurate, reliable representation of the target system, its existence is miraculous. But science and scientific realists do not accept miracles as compelling explanations and hence the community of climate scientists and modellers believe that this agreement is by virtue of plausible, reliable representation (Parker, 2009).

Parker's main point of debate with Lloyd is over what is being confirmed, and she rejects Lloyd's argument that the model is itself confirmed or disconfirmed in the presence of robust evidence. Instead, what is confirmed or disconfirmed, is a hypothesis about the adequacy of the model for a particular predictive or explanatory purpose. Adequacy-for-purpose is difficult to establish in practice as Parker points out. Even alternative approaches to confirmation such as severe testing approach proposed by Mayo (1996), which employs a sort of falsificationist method to try and determine adequacy-for-purpose, face the same challenges as the confirmation approach mentioned above (Parker, 2009). Either way, her argument is that the lack of information about certain aspects of models, along with other uncertainties makes for a very difficult barrier to penetrate or cross and approaches such as confirmatory robustness and severe testing can at best offer hypotheses about plausible adequacy-for-purpose. This

standard of plausibility needs to be much higher in order for policymakers and decision makers to seriously consider model results and scientists' opinions in making important decisions regarding adaptation, mitigation, and assessing vulnerability (Parker, 2009).

Model Agreement: What does it mean?

The next set of challenges that Parker (2011) brings up are regarding the epistemic significance that robust model predictions are supposed to provide. At the outset, Parker's goal is to identify the conditions under which robust predictions have epistemic significance. Once these conditions have been identified, the next step is to assess their validity and if the epistemic significance of robustness holds in the case of climate model ensembles. To that end, Parker investigates whether this significance justifies the inferences made in three cases: a) that an agreed upon hypothesis **H** is likely to be (approximately) true, b) that there is warrant for a significant increase in confidence in **H**, and c) that the security of the claim to have evidence for **H** is increased (Parker, 2011).

Tackling the first inference, Parker asks the question, "under what conditions can an inference from robustness to likely truth be justified?" (Parker, 2011). Drawing on Orzack and Sober's (1993) criticism of Levins' (1966) conceptualization of robustness, and also Woodward (2006), Parker argues that Levins' robustness is largely inapplicable in actual science and modelling. Levins' robustness logic involves comparing robust model results to the results of a true representation of a system. This brings us back to the dilemma we mentioned earlier in Chapter I: In order for a robust result to be inferred as true, we need a model that is a true representation of the system, or at the least, know that it exists and what entails. But if we have that true representation or know what it entails, then we don't need a robust result in the first place.

Parker argues that we do not possess such a true representation of the climate system, that there is an epistemic gap. From the discussion of adequacy-for-purpose above and all that it entails in practice, it is difficult to argue that climate models or even model ensembles can meet this expectation. Further, as shown by Lloyd's complex empiricist argument in Chapter I, observational datasets are not completely independent, and Parker states that climate science analysis, the data used is reanalysis data which comprises "cleaned up depictions" of raw data. This cleaning up extends beyond simply removing noise and correcting for instrumental error in the raw data because the gaps in the raw data are filled using output from weather-forecasting models which are accompanied by idealizations and parametrizations of their own because of

which, "the frequency with which ensembles are found to capture truth, will be artificially inflated" (Parker, 2011).

One of the most common ways that climate models become "data laden" is through tuning, a process that involves making ad-hoc changes to parameter values to reflect a better model-data fit. While this is a common practice across different scientific disciplines and it is not considered a bad practice in and of itself; the assumption that models tuned to present data will perform similarly well with respect to future, as-yet-unseen data is suspect in climate science given the imperfect nature of representation of all current climate models. Hence, Parker concludes this argument by stating that the inference from robustness to truth is not justified because our current crop of climate model ensembles cannot be said to be adequate for the purposes of discerning the truth value of hypotheses regarding future climate change simply because they can approximately accurately represent past climate conditions (Parker, 2011).

The second inference that Parker discusses is the significant increase in confidence in a given hypothesis that robust climate model results are supposedly responsible for. As we saw in the survey of climate scientists carried out by Pirtle et al. (2010), climate model agreement is perceived to be a good thing in that it increases scientists' confidence in the representational abilities of the model ensembles. Parker analyzes these inferences using three frameworks: Bayesian, Condorcet's Jury Theorem, and a sampling-based perspective, each of which she states, runs into problems when concerned with climate model ensemble predictions.

The first perspective deals with Bayes' theorem: **P(H|e) = P(H)*P(e|H)/P(e)** where **P(H|e)** is the updated probability that **H** is true after encountering incremental evidence **e. P(H)** is the probability assignment for hypothesis **H** before obtaining incremental evidence **e**, **P(e|H)** is the probability of obtaining **e** given that **H** is true, and **P(e)** is the expected probability assignment for **e** prior to actually encountering said incremental evidence. In this case, the confidence in **H** is the subjective probability assigned to the hypothesis and it can be updated in the face of new evidence. The important thing to note here according to Parker is that **P(H|e)** should increase if and only if **P(e|H) > P(e|~H)**, that is to say it must be more probable that the incremental evidence is encountered in the case that **H** true than if it is false (or **~H** or not-**H** is true).

Hence, Parker's formulation of the Bayesian argument is as follows:

1. **e** warrants significantly increased confidence in predictive hypothesis **H** if **P(e|H) >> p(e|~H)**
2. **e** = all of the models in this ensemble indicate H to be true.

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|∼H)**

∴ **e** warrants significantly increased confidence in **H**. (Parker, 2011).

Examining each of these premises, Parker argues that it is the third premise which entails a potential insufficiency of the Bayesian framework when ensemble model predictions are concerned. Since the probability assignments are subjective to particular "epistemic agents" that is, specific individual scientists or modelling groups or centres she claims the third premises then requires a more substantial justification (Parker, 2011). Once again, Parker's argument boils down to the idea of the epistemic gap and that neither of the two justificatory arguments that the Bayesian might pose: ensemble construction and ensemble performance, can provide justification for the third premise. Our current understanding of the climate system and how to model it is far from complete and our models are constructed with idealizations and parameters undercuts the ensemble construction argument. As discussed in the previous sections, Parker suspects that the frequency for truth-detection in ensembles may be inflated because of data-ladenness of models through tuning and other processes. She extends this argument to reject the ensemble performance position which claims that multiple trials and runs resulting in high probability of **H** being true should increase confidence (Parker, 2011).

While the Bayesian perspective is the one that is most commonly adopted by the scientific community, Parker analyses two other approaches which also provide disappointing results. The first is Condorcet's Jury Theorem (see Parker, 2011 and Ladha, 1995) runs into problems since model independence is suspect, and exactly in what sense, if at all, models in an ensemble can be treated as independent is a topic of debate in climate science and philosophy of science as well (Weisberg, 2006; Pirtle et al., 2010; Justus, 2012; Lloyd, 2015; Annan and Hargreaves, 2017). Hence that condition that 'votes' of jury members (or results of ensemble members) must be independent and not based on any sort of conference between the voters is one that is most likely to be violated given the way models are constructed and share a lot of core and auxiliary components. Even in multi-model ensembles, models can share many features. The second condition, that each member has more than 50% chance of making a correct prediction, does not have a strong basis because we have high confidence in the reliability of very few models and for very few variables that they are deemed adequate to represent (Parker, 2009). Hence it is difficult to argue that all the models in an ensemble have been analysed to that extent, such that we can assign a probability $p>0.5$ and hence we can reject the CJT justification for increase in confidence (Parker, 2011).

The third and final possible argument that Parker explores in this section is that treating model ensembles as a random sampling of models allows increased confidence in **H** if all the models in the sample indicate that **H** is true. However, the problem is obvious here, since climate model ensembles are ensembles of opportunity, and do not involve systematic or random sampling methodology (Tebaldi and Knutti, 2007). Parker also adds that it is not clear whether a larger model space from which to sample ensembles can even be specified. Additionally, the given how uncertain we are about how to properly represent the climate system, it is more likely that an ensemble of opportunity (where results agree, but construction may differ) is likely to be more heterogenous than a random sample (Parker, 2011). This third argument is logically the weakest attempt, but as Parker demonstrates, even the strongest and usually signature argument of the scientific community: Bayes' Theorem, can only be used to prove increased confidence for very few instances. For a large number of models and large number of purposes (representation of variables, processes, feedbacks and so on), there is not enough analysis done to warrant this increase in confidence (Parker, 2011).

The final inference that Parker considers is that robustness increases the security of the claim that a hypothesis **H** true. Security is defined as "the degree to which an evidence claim is immune to defeat when there is a failure of one or more auxiliary assumptions relied on in reaching it." (Parker, 2011). The general argument provided here is as follows: Using a certain model **M** and some auxiliary assumptions $\mathbf{A_1}$, scientists arrive at evidence claim **E** that provides some confidence in **H**; that **H** can have some minimum strength. If any assumptions in the set $\mathbf{A_1}$ were mistaken or incorrect, they would have to reconsider **E** and the consequent strength of **H**. The scientists run **M** again, but this time with a different set of auxiliary assumptions $\mathbf{A_2}$, $\mathbf{A_3}$, … $\mathbf{A_n}$ and derive the same evidence claim **E**. Now, if each $\mathbf{A_n}$ is at least partially logically independent of all previous sets $\mathbf{A_1}$, $\mathbf{A_2}$, $\mathbf{A_3}$, … $\mathbf{A_{n-1}}$ – in that there is at least one assumption in $\mathbf{A_1}$ such that, even if it is false, all other assumptions in $\mathbf{A_2}$ could still be true, and so on – then, the security of **E** increases.

The focus is on the evidence here, and not on confirming the hypothesis as it has been in the previous cases. If **E** is reliable and confers us with a high level of confidence in **H**, then multiple instances of **E** could be plausibly used to infer **H**. But, if **E** provides only weak confidence in **H**, then multiple instances of **E** do not provide a jointly higher confidence of **H**, all they tell us is that firm evidence of weak confidence **H**. The level of confidence in **H** does not change in either case: what changes is our degree of confidence in **E**. That is to say, the argument from robustness to security works if and only if the evidence claim can provide high level of confidence in the hypothesis being true. If the **E** is weak, then multiple instances of **E**

are still weak, and if we found some other evidence which confers more confidence in **H**, we can discard **E** in favour of that evidence: it is not secure. However, if **E** is strong evidence (say, the best evidence we have) for **H**, then multiple instances of **E** raise the standards for other evidence claims – they would have to be better than our best.

For Parker, this argument is not applicable to climate model ensembles on because of this underlying assumption that individual climate models provide relevant, reliable evidence for some credible strength of belief in **H**. In practice, individual model results are weak lines of evidence, which is why the ensemble approach is utilized in the first place (Parker, 2011). Individual climate models are very approximate and highly tuned representations of the climate system; there are various kinds of errors involved that make individual model results suspect. One of the chief reasons for using climate model ensembles is because they can be constructed in a way that many errors cancel out. As climate models cannot individually provide the high level of confidence in specific hypotheses, the implication here is that at most what a robust climate model ensemble result can show is that it can provide only as much confidence as the "weakest" member of the ensemble (Parker, 2011).

Climate Models and Inference to the Best Explanation:

Parker mentions the idea of the epistemic gap multiple times in both her papers. The argument that we possess an incomplete understanding of the Earth's climate and its true dynamics, that there is no Laplace's Demon to make perfectly accurate representation of the evolution of the climate system (Frigg, 2014), our models are at best idealized approximations that sacrifice process fidelity for predictive accuracy. Katzav (2012) characterizes the epistemic gap in three dimensions of ignorance: "our ability mathematically to model well-understood processes within it is limited; our understanding of many known processes within it is limited; and our understanding of which processes and types of processes do occur within it is limited" (Katzav, 2012). He uses the idea that models (and by extension model ensembles) are not completely explanatory devices to argue that the warrants and justifications their results provide to make inferences to truth, such as those noted by Parker (2011) above, cannot be considered as instances of inference to the best explanation. Further he argues that inference to the best explanation (IBE) can only provide these warrants "only to model implications about which there is real uncertainty as to their accuracy" or known unknowns. In the case of climate models, the case of unknown unknowns weakens the case of IBE.

In the scope of this dissertation, it is Katzav's first argument which is more important as it concerns the construction and composition of climate models. As we have seen so far in this dissertation, climate models are commonly accepted to be made up of a causal core with physical equations and parametrizations which comprise sub-grid processes and idealizations (Weisberg, 2006; Lloyd, 2015; Parker, 2009, 2011). He states that the former possesses explanatory virtues, that is to say, the causal core is designed and constructed with the purpose of simulating causal explanations and hence provides inferential warrants by virtue of this explanatory power. The latter possesses virtues he calls more-than-explanatory virtues or explanatory-instrumental virtues (Katzav, 2012), although one could argue that they could be called less-than-explanatory since he claims they reduce the explanatory power of climate models and hence the power of the subsequent inferential warrants.

He writes that "are successfully designed to compensate, by means other than just increasing explanatory quality, for the limitations on model inferential reliability that result from a limited ability to model realistically relevant target complex systems" (Katzav, 2012). Parametrizations could be non-simulated, empirical equations for processes which are plugged into models, or they could be empirically derived correlations between feedbacks, or used to cover gaps in the simulations. This agrees with and is a detailed, technical expansion of earlier arguments about parametrizations and model tuning being used to achieve predictive accuracy and success. The conclusion that follows is that since these non-explanatory components are essential to the success of models in that they act in a compensatory manner, the explanatory-instrumental virtues they possess, are also essential to the success of models (Katzav, 2012).

In Chapter I, we saw that Lloyd's confirmatory robustness called for independent support for model components such as parametrizations. This is often in the form of theoretical justification, which seems to be sufficient to warrant an explanatory inference, but Katzav argues that it is not so (Katzav, 2012). Similar to Parker's (2006) adequacy-for-purpose argument, the conclusion here is that theoretical justification oftentimes does not bestow full explanatory strength of the theory guiding their construction. This follows from the fact that we do not know if a perfect representation of the system exists or what constructing that would entail. We would not need parametrizations if we knew either of those two things. Hence, by design, parametrizations are approximations and "possess the explanatory-instrumental virtue of being adequate substitutes for at least part of this underlying theory" (Katzav, 2012).

Although Katzav's argument is not posed directly in opposition to robustness analysis, it clearly points to a caveat in robustness arguments so far. The goal of robustness arguments has been to provide the warrants that justify the inference to truth. The justification for this has

spanned from simple model agreement (Levins, 1966) to the common causal core approach (Weisberg) to confirmatory robustness (Lloyd, 2015) who argued that the confirmatory aspect comes from the added confidence that independent support for model parametrizations brings. Lloyd argued that the robustness analysis should not be satisfied with mere agreement in models but in fact be used to confirm a given hypothesis as the causal explanation for that agreement (Lloyd, 2015).

Katzav's position can be used to undercut the causal explanation part of Lloyd's by arguing that parametrizations are not explanatory components and hence possess explanatory-instrumental virtues which are conferred to the model, and by extension the explanatory-instrumental virtues of various models are conferred to the ensemble. Hence whatever inferences we make across the ensemble are not by virtue of the best explanation, but by virtue of something weaker than a true causal explanation: a small set of model implications (Katzav, 2012). As we noted in the previous section, Parker (2011) also considers individual model results to be weak lines of evidence owing to our modelling limitations, data limitations, and uncertainties, and that robust weak evidence was not really useful. It could be similarly argued using Katzav's position that with model implications (explanatory-instrumental) being weaker than true explanations (explanatory), inferences provided by robust model implications (in that they are true across the ensemble) are also weaker than inferences provided by robust true explanations.

The objections raised against robustness revolve around one central theme: the epistemic gap. Earlier criticism by Orzack and Sober (1993) pointed out a version of this gap in Levin's (1966) initial robustness argument, which took the form of the dilemma of the true representation of a system. In order to infer truth about set of models from a robust result, we would need to know that one of the models or representation in a set is the true representation, without which the agreement could be meaningless as the models could all agree and be incorrect. But if we know that one of the models is a true representation, we would not need a robust result in the first place. With Parker (2006, 2011) and Katzav (2012), this challenge takes a slightly different form since in the case of climate science we know that there no perfect representation exists and hence we do not know what designing and constructing one might entail because of our multidimensional ignorance (Katzav, 2012). One could interpret a more theoretical reading from Katzav when he writes:

“If IBE did capture the warrants we have in light of hybrid model successes, the warrants in question would allow us to infer the correctness or approximate correctness of successful hybrid models.” (Katzav, 2012).

The implication here being that if IBE was at work in the case of a climate model, then we would have our (near) perfect causal explanation which would then mean that we know that the perfect representation exists in the set or ensemble. This would take us back to the start of the dilemma pointed out by Orzack and Sober (1993): if we have a perfect representation, then we do not require robustness.

Chapter III – Robustness post-LWWOL

So far, we have looked at the various robustness arguments that have been made and how they have been developed into a form of analysis that can be applied to modelling in various fields of scientific inquiry. Robustness analysis is also utilized in climate science and modelling, where a result that is robust across our best multi-model ensembles, is said to attribute confidence to our predictive or explanatory hypotheses about the climate system. We have also seen arguments that downplay the usefulness of robustness analysis in the particular field of climate modelling: specifically, to do with the epistemic gap and our inability to completely model the system.

In this chapter, we shall explore the state of robustness analysis post-Lloyd (2015). This exploration begins first, with an exploration of the concept of independence that is touted by both scientists and philosophers as being an important notion in robustness analysis. We examine the arguments offered by Annan and Hargreaves (2017) which attempt to justify the notion of independence used by climate scientists. We also look at arguments regarding ontic independence and conditional probabilistic independence by Stegenga and Menon (2017). Although they show that both kinds of independence do not provide the critical warrant for making a robustness argument, they are less pessimistic about the usefulness of independence than Schupbach.

A further step in the evolution of the concept of robustness comes from Schupbach (2016) who analytically and rigorously rejects the former concepts put in place by the LLWOL approach, while establishing his own notion of robustness. For Stegenga and Menon (2017), notions of independence, while they may not justify an inference by robustness, still have positive confirmatory virtue; Schupbach conversely argues that probabilistic independence is not useful for even those purposes. His explanatory notion of robustness is applied to climate science by Eric Winsberg (2018) who argues that robustness means moving beyond simply model result agreement and incorporating multi-modal evidence to reach our epistemic goals. Finally, we look at the problems that emerge when using multi-modal evidence to make robustness arguments as formulated by Stegenga (2009).

Independence in Climate Science:

Annan and Hargreaves (2017) discuss several notions of independence found in climate science literature, noting that most of them are largely qualitative in nature, with no quantitative explication of independence. The first such notion of independence called "truth plus error" or "truth centred" interpretation, which is described by as being the fundamental assumption behind the practice of model ensemble construction: that if models are independent, then their errors are independent, and hence averaging or combining results over multiple independent models should cause the errors to cancel out, hence decreasing uncertainty in the results of the ensemble (Tebaldi and Knutti, 2007).

However, as studies with Coupled Model Intercomparison Project Phase 5 (CMIP5) ensembles have shown, there are strong correlations in the patterns of climatological biases that are exhibited by the results, which implies that there is a strong correlation between errors of different models (Annan and Hargreaves, 2017). This has been a point of criticism from philosophers such as Weisberg (2006), Lloyd (2009, 2010, 2014), Parker (2009, 2014), and Justus (2012) among others, that since models many common elements, even in "sufficiently heterogenous" (Weisberg, 2006) multi-model ensembles, there is a very high chance that the errors of the models will also be correlated. This problem crops up again towards the end of this chapter when we see Winsberg's conception of RA.

Another interpretation of independence is offered by Abramowitz and Gupta (2008) who suggest that scientists should base their notion of independence on inter-model differences in the ensemble. For models that produce very similar outputs, they suggest negative weighting. There is an obvious flaw here as Annan and Hargreaves (2017) argue: in the event that models produce similar outputs because they actually happen to be correct (and not just as a result of sharing similarities in construction), their results will still be negatively weighted. Hence there is no objective or absolute criterion that can be used to determine independence. They argue, similar to the philosophers mentioned above and in previous chapters, that since all models are designed to be as close to ideal simulations of the real climate system as possible, all tuned to current climate observations. Hence it is unsurprising that their results have much in common (Annan and Hargreaves, 2017).

The approach of Sanderson et al. (2017) seems to be the most coherent and complete qualitative approach to independence. They utilize the idea of inter-model differences in output

and down-weighting the most similar models, as proposed by Abramowitz and Gupta (2008). But to counter the problem faced by Abramowitz and Gupta, Sanderson et al. compare inter-model difference with data-model differences, that is to say they look for a data-model fit. This provides a candidate for the independence criteria that was missing from the earlier interpretation. Although, as Annan and Hargreaves (2017) point out, this method only reduces dependency by some degree instead of eliminating it, something that is also pointed out by Lloyd (2010) in the complex empiricism debate – that comparing models to data does not provide an independent yardstick because of the data-ladenness of models.

Probabilistic Independence:

In order to formulate a quantitative explication of independence in relation to climate models, Annan and Hargreaves (2017) propose the use of conditional probabilistic independence (or conditional independence). Two or more events or outcomes are probabilistically independent if the occurrence of one does not affect the probability of the occurrence of the other. This has been a staple focus of the philosophical literature regarding robustness and scientific literature, especially in the area of climate science and modelling. Conditional independence entails the probabilities of two events relative to a third event: "two events, A and B, are defined to be conditionally independent given a third event, S, if their joint probability conditional on **S**, $\mathbf{P(A \cap B|S)}$, is equal to the product of their individual probabilities both conditional on **S**, $\mathbf{P(A|S).P(B|S)}$" (Annan and Hargreaves, 2017).

In climate models, the events (**A, B, C**…) usually denote the confirmation a hypothesis **H** by models ($\mathbf{M_1, M_2, M_3 \ldots M_n}$). Hence the notation for conditional probabilistic independence becomes: $\mathbf{P(M_1 \cap M_2 \cap M_3 \ldots M_n|H) = P(M_1|H).P(M_2|H).P(M_3|H) \ldots P(M_n|H)}$. The authors argue, "If a researcher does not know how to improve their prediction of a particular model, in light of being given a particular set of outputs from another named model, then this pair of models is in fact absolutely independent to them in statistical terms." The advantage that the authors purport their formulation of independence is that instead of a qualitative description of independence, it is an absolute criterion for independence and is not a measure of relative differences between models. Further it is relative to the hypothesis at hand. (Annan and Hargreaves, 2017). However, another thing they state is that their formulation is not at all related to model and ensemble performance, and that developing a useful formulation that can inform model and ensemble performance is an important task.

Annan and Hargreaves (2017) clearly state that in the climate modeling literature, when people referred to independence, they were implicitly referring to conditional probabilistic independence. This explication of independence serves as an important first step to our exploration of the relationship between robustness and independence. Since their interpretation is unrelated to model performance, the next step involves connecting robustness and independence – we shall see Stegenga and Menon (2017) do this in the next section using Bayesian networks

Robustness and Ontic Independence:

So far, we have seen the issue of independence being raised be of one type: probabilistic independence. Schupbach (2016), we shall see below, rejects the various kinds of probabilistic independence being the underlying notion at work when philosophers and scientists appeal to robustness. In his view, as long as means of evidence can eliminate competing hypotheses, they are RA-diverse, regardless of whether they are independent or not. Probabilistic independence is not a sufficient condition for RA-diversity for Schupbach, but he does not comment on whether it is a necessary condition. Most probably, having means of detection that are probabilistically independent in addition to already being RA-diverse in an explanatory sense would make analysis easier than having interdependent means.

A second type of independence that has been missing from our conversation on model robustness ontic independence, defined as, "when the multiple lines of evidence depend on different materials, assumptions, or theories" (Stegenga and Menon, 2017). It is often thought that if a hypothesis is supported by ontically independent evidence, that is, the theoretical or material bases for different modes of detection are independent, then one can build a robustness argument. Stegenga and Menon point out a caveat in this assumption – if we have concordant, multi-modal evidence, that is to say that multiple various modes of evidence agree and support our hypothesis in a way that seems to warrant a robustness argument, but if this evidence is not actually ontically independent, then the argument that follows is pseudorobust in nature (Stegenga and Menon, 2017).

This points out to the idea that multi-modal-evidence is not necessarily ontically independent by virtue of it being multi-modal. For example, as we have seen so far in climate science and modelling, the multi-modal evidence (from models, observations, paleoclimatic data, geological phenomenon etc.) is truly independent. Models share a genealogy (Masson and Knutti, 2011), models are also tuned to current climate data, and as Lloyd examines at

length, radiosonde and satellite data are linked through the act of calibration. Hence it appears that Stegenga and Menon's claim about pseudorobustness would apply to climate science, although this is not the only type of pseudorobustness that can crop up.

A second type of pseudorobustness comes from what Stegenga and Menon term as dissynergistic evidence – where individual lines of evidence confirm a hypothesis but when in conjunction, the combined evidence seems to support a competing hypothesis. This type of pseudorobustness cannot be avoided by ensuring these individual lines of evidence are ontically independent, but rather one must ensure they are conditionally probabilistically independent. Although, they also note by avoiding this second type of pseudorobustness, conditional probabilistic independence provides a "minimum justification" for robustness arguments even though the requirements are very demanding (Stegenga and Menon, 2017). By this they mean that in most cases it is unclear whether scientists have access to conditionally probabilistically independent evidence. This level of justification is does not really warrant robustness arguments though, as they point out, it does not warrant the "special epistemic oomph" that robustness is held to provide.

Their argument also challenges Parker's (2009, 2011) analysis of model agreement, specifically, conditional probabilistic independence seems to make the confirmatory value of model agreement redundant. In Stegenga and Menon's view, conditional probabilistic independence avoids the problem of dyssynergystic evidence, hence ensuring that the evidence will jointly provide more confirmation than individual means. Further ascertaining the probability of the hypothesis given agreement of results amounts to double counting as the avoidance of dyssynergy already provides concordance or agreement.

In this way they agree with Schupbach (2016) who, we will see ahead, argues that probabilistic independence does not provide a strong enough justification for robustness arguments. Although they only agree up to a certain extent. While they agree with Schupbach on the limits of conditional probabilistic independence, Stegenga and Menon (2017) disagree with him regarding the overall usefulness of probabilistic arguments. While they argue that probabilistic independence arguments have some value that helps in confirmation, Schupbach rejects the usefulness of probabilistic arguments in general as seen in the next section.

Often, appeals to robustness are actually cases of pseudorobustness either due to means of evidence being ontically or conditionally probabilistically dependent. In the case of climate science, we have already seen that even though models may be structurally independent,

models in an ensemble are tuned to current climate data which makes them ontically dependent not only within the ensemble but to observational and reanalysis data to which the ensemble fit is compared. Thus, what appears to be robustness in climate modelling, actually might be pseudorobustness due to failure of ontic independence (Stegenga and Menon, 2017). Further, the problem of dyssynergystic evidence crops up if ontic independence holds. In that case, conditional probabilistic independence is a necessary condition for avoidance of dyssynergy. But in the case of climate science, the failure of ontic independence ensures that modes of evidence are synergistic, and hence the requirement of conditional probabilistic independence seems to be an even stronger condition.

Robustness Analysis as Explanatory Reasoning:

Jonah Schupbach provides an updated account of robustness in which he develops a "single sense of evidential diversity that drives our reasoning in RAs" (Schupbach, 2016), a notion which he argues must specify certain conditions and must have a normative epistemic appeal. That is to say that the criterion over which models or theories are judged should be one that is in alignment with what is accepted by scientific practice. Before developing his own notion of RA diversity, Schupbach first explores the notions that have thus far been proposed as the idea underlying robustness, namely the various modes of independence that have varying levels of support from Levins till Lloyd. Although the notion of independence has been refined over the years by philosophers in the LWWOL group itself, Schupbach is the first to provide an analytical argument for why independence does not fit the criteria for RA diversity. Schupbach critiques recent robustness arguments in the philosophical literature by stating that Weisberg's idea of "sufficient heterogeneity" (Weisberg, 2006) is too vague and that philosophers are too dependent on Bayesian probabilistic independence which does not correspond to the notion which is actually at work in RA.

First, Schupbach analyzes the two types of probabilistic independence: unconditional probabilistic independence and reliability probabilistic independence. According to the unconditional independence account, he writes, "if two means of detection are RA-diverse, then the fact that R [a proposition describing the result that has been robustly detected by various means] is detected via means *i* should have no bearing whatever on the probability that R will be detected using means *j*: $\mathbf{P(R_i \& R_j) = P(R_i) \times P(R_j)}$" (Schupbach, 2016). This further necessitates that $\mathbf{P(R_i) = P(R_i|R_j)}$ and $\mathbf{P(R_j) = P(R_j|R_i)}$. But this account fails to provide Schupbach's notion of RA diversity because as Orzack and Sober (1993) argued, Levins' idea

of independent modes of detection was flawed since the common biological assumption was an important part of Levins' robustness and having a common core structure precluded models from being truly independent. As such, detection of a result **R** by a previous model would inform the probability of it being detected by another model, that is to say that $\mathbf{P(R_i) < P(R_i|R_j)}$ (Schupbach, 2016).

The notion of reliability independence is one that was proposed by Wimsatt (1981), discussed earlier as the independence of the probability of failure of means of detection. In essence, this means that each means of detection is or is not reliable, independent of each other (Schupbach, 2016). Apart from being a more accurate interpretation of Levins' original idea of truth lying at the intersection of independent lies (Levins, 1966), Schupbach argues that reliability independence demonstrates the epistemic appeal of (evidential) diversity since it creates a web of independent lines of justification which is no weaker than its strongest member while the former unconditional independence manifests as a linear chain of justification where the chain is only as strong as its weakest link (Schupbach, 2016).

Despite this epistemic appeal, Wimsatt's reliability independence does not meet the criteria for Schupbach's notion of RA diversity because once again, the LWWOL program requires a common causal core, which Schupbach argues can potentially hamper the independence of reliability. He writes "To the extent that we are aware of such overlapping sources of potential unreliability, learning that one of these experiments [or model] is leading us astray provides relevant information when deciding whether to trust the other. In particular, such information will often greatly reduce our estimate of how reliable the other is" (Schupbach, 2016), and this extends to means that are considered fully RA-diverse under this notion.

The third type of independence notion that Schupbach examines is "confirmational independence" as proposed by Lloyd (2009, 2010) who argues for model robustness as a confirmational virtue (Lloyd, 2015). As mentioned earlier in Chapter I, by the time philosophers such as Weisberg and Lloyd entered the conversation surrounding robustness, the traditional notion of independence had been done away with in favour of embracing the interdependencies in models (Justus, 2016). Explicating the idea of conformational (or conditional) independence, Schupbach writes that relative (or conditional) to a hypothesis H, "Two means of detection are RA-diverse, according to this account, only if their results incrementally confirm / disconfirm H (raise / lower H's probability) to the same extent regardless of whether we have detected and learned the results using the other means" (Schupbach, 2016).

Formulating this logic in a Bayesian notation, he writes that if the *i*th and *j*th means of detection are RA-diverse with respect to H, then $\boldsymbol{c(H, R_i|R_j) = c(H, R_i)}$ and $\boldsymbol{c(H, R_j|R_i) = c(H, R_j)}$ where *c* is some adopted Bayesian measure of incremental confirmation (Schupbach, 2016). Like Wimsatt's notion of reliability independence, Lloyd's confirmational independence demonstrates the appeal of evidential diversity since confirmationally independent means jointly confirm H to a greater extent than any individual means of detection. Yet, despite this, confirmational independence fails to meet Schupbach's criteria for RA-diversity since the means could be confirmationally independent, but still be incorrect about the result.

In summary, Schupbach's argues that the problem is not that RA-diverse means fall short of full independence in any of its three forms (unconditional, reliability, and confirmational), and it is not clear that being closer to independence implies more RA-diversity. Rather, his criticism is that those means which are "clearly and recognizably RA-diverse may not even come close to being independent" (Schupbach, 2016). With this analysis in place, he sets out to propose an explanatory notion of RA-diversity by building on the work of Horwich (1982) whose account of evidential diversity focused on probabilistic elimination of competing hypotheses instead of probabilistic independence, which is a notion that Schupbach claims lures philosophers away but is not a notion that is at work in actual RA (Schupbach, 2016).

According to this method of RA by explanation and elimination, each newly cited means of detection "RA-differs" from previous means if and only if it can rule out potential explanations offered by the previous means (Schupbach, 2016). This leads to the elimination of competing hypotheses that could explain the result, and the most robust hypothesis would be the one remaining after all other competing hypotheses are eliminated. Schupbach writes that such means are "explanatorily discriminating" between two competing hypotheses H and H', that is to say that H would explain the detection of a result R by a new mean, something that H' would fail to do and hence this new mean would eliminate H'.

Under this notion, means of detecting R are RA-diverse with respect to potential explanation (or target hypothesis) H and its competitors if their detections (R_1, R_2, … R_n) can be sequentially arranged in a way that any member is explanatorily discriminating between H and some competing hypothesis which is not yet ruled out by the $|R_{i-1}{}^{th}$ means of detection (Schupbach, 2016). This means that RA-diverse means of detection are a way of successively eliminating competing hypotheses by providing incremental evidence which is explanatorily discriminating. In addition, a more salient feature of this notion of RA-diversity is that it does not require means of detection to be "strongly diverse" or "sufficiently heterogenous" as stated

by Weisberg (2006) in some absolute sense, instead what matters is that they are explanatorily discriminating (Schupbach, 2016) which is more aligned with traditional scientific practice and methodology than notions of probabilistic independence.

Beyond model agreement: Cumulative Epistemic Power

All the forms of robustness we have seen so far were focused on the idea of model agreement. The agreement of the results of models in an ensemble is seen as a good thing in that it adds confidence to inferences that we make from it. Earlier robustness arguments such as those proposed by Levins (1966), Wimsatt (1981), and Weisberg (2006) focused on the agreement of results and as stated in Chapter I, Lloyd (2015) argues that this isn't enough, that RA should not simply stop at the agreement of results. The agreement of results might be confidence building, but it does nothing in pointing us towards a causal explanation of why those results came about. For Lloyd, the goal of RA should be to add confirmatory value to our hypothesis, which in this case is that anthropogenic greenhouse gas emissions are the driving force behind increased climate warming since the 20th Century.

Even so, while Lloyd's confirmatory conception of RA is very comprehensive in that it includes the common core structure, model-data fit, and independent support for the parametrizations, as Schupbach (2016) points out, that by itself is not enough. The confirmatory power is not useful if it confirms the wrong hypothesis. Let the GHG forcing hypothesis be denoted by **H**, and say there is some alternative hypothesis **H'**, the related forcing or feedback for which is also part of the common core structure of all the climate models in an ensemble. Further, if we assume that the forcing for **H'** is biased or in some way improperly simulated such that the simulation results are very similar to those one would expect if **H** was true. In this case, Lloyd's formulation of RA will not be able to nail down the causal explanation to either of the competing hypotheses. Schupbach's explanatory RA reasoning proposes that in addition to being confirmatory and confidence increasing, RA-diversity implies the "chopping away" of alternative, competing hypotheses (Schupbach, 2016).

Eric Winsberg (2018) uses Schupbach's analysis to lay the foundation of an explanatory reasoning-based RA specifically tailored for climate models and climate science in general. He argues that an explanatory reasoning-based RA will mean having to move beyond model agreement, further than Lloyd has already advocated. He agrees with Lloyd's idea that beyond model agreement, climate scientists should make use of independent support and evidence outside models that supports the target hypothesis. Winsberg's argument is that one must look

beyond models and to the entire spectrum of evidence available including model simulations, paleoclimatic data, proxy data, theoretical calculations, and so on, which support the target hypothesis. He writes, "Rather than trying to show how RA is part of a complex epistemic landscape in which other sources of evidence also play a role, our aim will be to show how RA can offer a comprehensive picture of that complex landscape—of how all these sources of evidence work together." (Winsberg, 2018)

On adopting Schupbach's explanatory notion of RA-diversity, RA then becomes not about models but about the target hypothesis **H** in question, as Winsberg writes, "It is a property of a set of detection methods with respect to a particular hypothesis" (Winsberg, 2018). He proposes that the evidence for **H** from various epistemic activities be used in conjunction with model results, model agreement, and whatever level of confirmation that brings to eliminate alternative, competing hypotheses. In the case of climate science, explanatory RA-diversity does not give a context-free criterion for believing in a hypothesis and that we can need context-specific epistemic standards to determine how much confidence is enough to warrant belief. RA-diversity implies that as we add more epistemic activities that are RA-diverse in that they eliminate competing hypotheses, we approach *complete* justification for our belief in hypothesis **H**. How RA-diverse this set of activities needs to be in order to provide *sufficient* justification for belief in **H**, is a further question, one that "will always be a matter of judgment, context, considerations of inductive risk, etc." (Winsberg, 2018).

Examining the IPCC value and confidence interval for Equilibrium Climate Sensitivity (ECS), Winsberg performs his own RA to show that the methods of evidence used to make hypotheses that ECS lies between 1°C and 6°C (very unlikely), with the most likely value lying in the range 1.5°C to 4.5°C. He uses three means of detection: models, instruments, and paleoclimatic data, each of which is subject to some out of (but not all) six forms of uncertainty: a) data uncertainty, b) lack of equilibrium, c) lack of independence from internal variability, d) simulation model error, e) different responses from different kinds of forcings, f) different base state. Each of these forms of uncertainty can be treated as alternative hypotheses which could lead us away from the true hypotheses. Winsberg proceeds to show how eliminating these alternative hypotheses across the three means of detection allows us to narrow down the explanations that fit with our target hypotheses:

"Suppose, for example, that using instrument data associated with a particular volcanic eruption, we find that the data support the hypothesis that ECS is between 1.5 and 4C. We can count this as a method of detection for this hypothesis. Thus, to do RA, we would want to ask: in addition to the truth of the hypothesis, what other explanations are there of the fact that this method detects that hypothesis? *We can then go through a–f and determine which ones provide such alternative explanations. Here, we can rule out a,d and f.* We can rule out 'a' because the hypothesis is 'wide' enough (spanning the range from 1.5 to 4C) to cover likely measurement error. *We can rule out model error, because it is not a simulation-model-based detection. And we can rule out the possibility that the detection is due to the data coming from a different base state, because we are using data from the recent past.* Thus, any second detection method that didn't suffer from these sources of error could help to rule out either b or e, and hence would count as RA diverse compared to our first method. Paleoclimate data is especially helpful here. That's because *paleoclimate data are not especially susceptible to b,c,d and e. But they are especially susceptible to a, and f.* They are also somewhat susceptible to b if ECS turns out to be especially high. Thus, in conjunction with each other, volcano and solar cycle data, combined with paleo-data, are pretty good at ruling out alternative explanations of the detection of the hypothesis that ECS> 1.5, and hence already by themselves provide robust evidence of that hypothesis." (Winsberg, 2018). (Winsberg's italics)

Robustness and Multi-Modal Evidence: Discordance and Relevance

Winsberg's argument assumes that the various modes of error are independent of each other and are not correlated. Since his analysis requires the sources of error to be diverse in order to understand the diversity of modes of evidence, if the errors happened to be correlated then it lowers the degree of confirmation that is assigned to the hypothesis at hand. What seems to be required is a criterion for establishing independence, as pointed out in Annan and Hargreaves (2017) earlier in this chapter. Is it plausible that they are uncorrelated, and if so, on what basis? Jacob Stegenga calls this the individuation problem: "having independent modes of evidence and knowing that they are properly independent are difficult; since robustness requires multiple modes of evidence, an incomplete or vague individuation of evidential modes will leave robustness as an incomplete or vague notion, and hence robustness-style arguments will be vague or inconclusive" (Stegenga, 2009).

The arguments that Stegenga offers as problems facing robustness involving multi-modal evidence comprise three "easy problems" and the "hard problem" (Stegenga, 2009). The first easy problem is that scientists do not necessarily possess multiple techniques of investigation or modes of evidence, the second is the aforementioned individuation problem, and the third is that concordant, multi-modal evidence will not necessarily yield a correct conclusion. In terms of climate science, scientists do have multiple modes of evidence such as observational data, models, paleoclimatic data and hence the first easy problem is resolved. The individuation problem as pointed out above, is still persistent in Winsberg's analysis since it is assumed that errors are independent, hence it is still unresolved. The third easy problem seems to follow the same logic as Parker's (2011) arguments regarding robustness guiding inferences to truth, confidence, and security in Chapter II. hence remaining unsolved.

The hard problem of discordance questions the degree of support that the target hypothesis receives in the event that multi-modal evidence disagrees. If concordant or agreeing evidence provides a high degree of support, it is unclear how much support, if any, discordant evidence provides (Stegenga, 2009). Winsberg claims that evidential diversity of a certain kind – RA-diversity, provides greater cumulative epistemic power, and Stegenga agrees, but points out something interesting. He writes that while robustness prescribes that concordant, multi-modal evidence is good, it also requires the acquiring or generation of more data from diverse, independent sources. And while there are many formulations of utilization of robust data, there seems to be no formulation or systematic way of amalgamating and assessing discordant data if and when it shows up (Stegenga, 2009).

If the data is discordant, then it is a matter of choosing the evidence that supports your hypothesis, opposing evidence, and incongruent evidence (Stegenga, 2009). Data in climate science can be argued to be discordant. As we noted in Chapter I, the entire debate between the direct empiricists and the complex empiricists (Lloyd, 2012) can be characterized as being about discordance between two modes of evidence. Further, those modes are not ontically independent: models are tuned to reanalyses data sets and current climate data (Parker, 2011), gaps in observational data sets are filled by synthesizing it with short term weather forecasting, further climate models themselves are not independent and share many similarities in core structure and auxiliary assumptions (Masson and Knutti, 2011).

Thus, the problem of discordance does not reject robustness, but rather points out if multi-modal data is discordant, then there is no systematic method of dealing with it. Stegenga

argues that in many cases, the appeal to robustness is a "philosophical cheap trick" (Stegenga, 2009). As seen from Stegenga and Menon's (2017) above, in light of failure of ontic independence, what is claimed to be robustness is actually a kind of pseudorobustness. The problem of discordance lowers the epistemic value of robustness – the more discordant the multi-modal data is, the less useful robustness is. Further, he argues that there is no universal standard to determine what evidence is relevant.

Now, Winsberg could possibly argue that unlike a sufficient diversity condition which requires a universal standard or "context-free criteria" to determine how much confidence in a hypothesis is enough confidence, explanatory-reasoning based RA-diversity does not require a context-free criterion – how much confidence is enough confidence can be determined by the context of the particular epistemic setting (Winsberg, 2018). This argument could be extended to both discordance and relevance: let the particular epistemic context of climate science decide how to deal with discordant data and determine what data is relevant. Although Winsberg does not provide any explicit reason for this type of claim, it's plausible that by deferring this question to the particular epistemic context, he is deferring to the expertise of the climate scientists.

In summary, in this chapter, we saw arguments from. Along with Annan and Hargreaves' (2017) technical account of independence in climate science, we also examine arguments analysing the concepts of ontic independence and conditional probabilistic independence from Stegenga and Menon (2017). While their arguments show that notions of independence do provide some degree of additional confirmation of a hypothesis, they do not warrant the "special epistemic oomph" that realist positions supporting robustness argue for. Schupbach (2016) who rejects the use of notions of probabilistic independence in favour of an explanatory reasoning-based notion that he argues is actually at work in robustness. We also saw the Winsberg (2018) apply Schupbach's notion of explanatory-reasoning based RA-diversity to climate science, explaining how emergent constraints (EC) analysis, a form of RA that is commonly used in climate science. He also expands the idea of RA beyond just model agreement, to include evidence in support of the hypothesis from different modes such as experiment, observational evidence, and paleoclimatic data. The last section of this chapter deals with Stegenga's (2009) the kind of problems that such multi-modal evidence might bring: the three "easy" problems and the "hard problem of discordance" and how those problems are translated into and persist in climate science.

The next chapter involves examining actual scientific publications that seem to use robustness arguments to justify their claims. An interesting claim that Schupbach makes is that his explanatory notion of robustness is well aligned with actual scientific practice. Specifically, the focus is on what notion it is that they appeal to when they invoke robustness: some sort of independence or explanatory reasoning as claimed by Schupbach.

Chapter IV – Robustness Analysis in Scientific Practice

After a detailed discussion regarding both issues mentioned at the outset: the underlying notion of robustness and its epistemic virtue, we shall examine results of scientific analysis from the field of climate science. The aim here is to see what scientists are referring to when they appeal to robustness, that is, are they making inferences from robustness based on some type probabilistic independence or explanatory power as discussed in the previous chapter. In light of the views offered by Stegenga and Menon (2017), it will be interesting to see if these appeals to robustness also turn out to be a variation of pseudorobustness. The first analysis we shall examine is that of Annan and Hargreaves (2006) whose explication of conditional probabilistic independence was provided in Chapter III (Annan and Hargreaves, 2017). In the second part of that paper, they extend the justification for conditional independence to this earlier estimation of climate sensitivity using observational (non-model) constraints.

The next two analyses involve emergent constraints analysis, which Winsberg states is a type of robustness analysis that utilizes Schupbach-style explanatory reasoning. The first analysis is that of Borodina et al. (2017) who use relationships found between model predictions of short-term sea ice metrics and future temperature variations to constrain the range of the ensemble predictions to show that it depends on the skill in selecting model and is robust across emission scenarios and seasons. They assume that this constraining is more effective in models tuned to current observational data (Borodina et al., 2017). The second analysis by Qu et al. (2018) is a survey and critical analysis of four claimed estimations of climate sensitivity using emergent constraints method. The goal here is to firstly investigate whether the constraints are emergent constraints on climate sensitivity as claimed, and secondly to investigate whether they are emergent constraints at all.

Early Robustness Analysis in Climate Science:

In 2006, Annan and Hargreaves published a paper called "Using multiple observationally-based constraints to estimate climate sensitivity" where they used Bayes' theorem to combine independent lines of evidence to reduce the uncertainty in the spread of ECS. The observational methods mentioned are: 20th Century warming, volcanic cooling, and the last glacial maximum (LGM). Each of these methods has a constraint described by the authors as (x, y, z) where y represents the temperature value (in Celsius) with highest probability while x and z represent the 95% confidence intervals on either side of y. In this

way, considering the intricacies and uncertainties of each method, 20th Century warming has a constraint of (1, 3, 10), volcanic cooling has a constraint of (1.5, 3, 6), and LGM has a constraint of (0.6, 2.7, 6.1) (Annan and Hargreaves, 2006).

This analysis was carried out before modelling was a sophisticated endeavour with multiple international modelling centres, multi model ensembles, and before GCMs became our go-to epistemic device for predictive or explanatory purposes. At this time, modelling was in its nascent stage in the absence of ensemble studies and results, the only modelling data they had was from perturbed parameter ensembles which involve single models run with different parameter values. As we saw in Chapter II, both Parker (2011) and Katzav (2012) are pessimistic of the value of individual models because of high, multi-dimensional uncertainties. Hence, they exclude perturbed parameter ensembles from their analysis because the results were not well confirmed across multi-model ensembles at the time of their analysis.

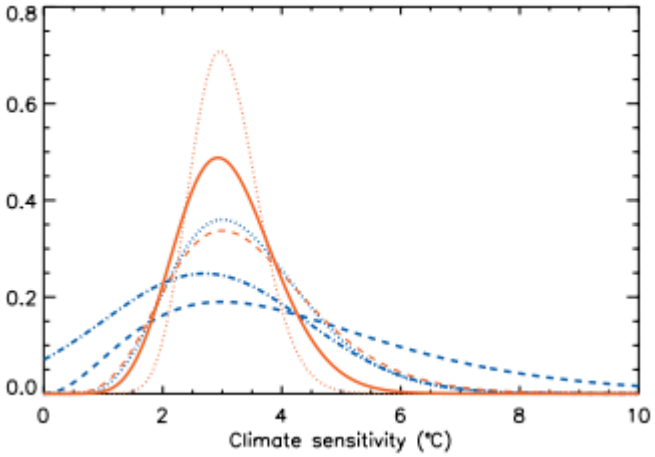

Figure 2: Pdfs and likelihood functions for climate sensitivity based on various observational constraints. Blue dashed line: 20th century warming (1, 3, 10). Blue dotted line: volcanic cooling (1.5, 3, 6). Blue dot-dashed line: LGM cooling (0.6, 2.7, 6.1). Red solid line: combination of the three constraints (1.7, 2.9, 4.9). Thin red dashed line: combination of three copies of widest constraint (1.5, 3.0, 6.3). Thin red dotted line: five constraints (2.0, 3.0, 4.3).

Their analysis involves multiplying the likelihood functions together and renormalizing to arrive at a composite likelihood function of (1.7, 2.9, 4.9). This is a substantial decrease in uncertainty over the previous estimation performed by the National Academy of Sciences in 1979, and in fact this would be one of the first uses of RA to be used in climate science to constrain the value of ECS (Annan and Hargreaves, 2006).

Prominently, we can see that the emphasis is on the independence of lines of evidence since these lines of evidence license the direct multiplication of likelihood functions. Considering the independence of the various observational constraints under analysis in their 2006 paper, they write later that they assumed the various observational constraints to be independent, although it was not clearly explained or demonstrated. This was because it is reasonable to assume that analyses multi-modal evidence will yield precise results than single lines of evidence. Further they argue that one can assume for simplicity's sake that the observational uncertainties themselves are independent, although it is not clear whether the resulting probability distribution functions are also independent (Annan and Hargreaves, 2017).

These assumptions seem reasonable until we bring in Stegenga's (2009) proposed problems, specifically the second easy problem or individuation problem, and the hard problem of discordance. The individuation problem is that it is difficult to know that multiple modes of evidence are actually independent, as is evident above. The independence is assumed to be a reasonable assumption which allows for multiplication of probability functions. As stated by Annan and Hargreaves (2017), "The significance of this conditional independence is that if we already have likelihoods P(A|S) and P(B|S), then conditional independence allows us to directly create the joint likelihood P(A∩B|S) by multiplication, rather than requiring the construction of p(A∩B\S) as an additional step".

The hard problem of discordance crops up when the authors exclude the perturbed-parameter model output and the Maunder Minimum estimations of ECS. While they state that the perturbed-parameter models were single model ensembles and hence they did not have high degrees of confirmation, or that both the Maunder Minimum estimations and model outputs were not independent enough and shared biases; they ultimately conclude that these estimates were excluded from their analysis because they did not support their results (Annan and Hargreaves, 2006). In the face of discordant evidence, scientists choose relevant evidence, that which supports their hypothesis; although as Stegenga (2009) argues that there is no standard criterion for relevance. They assumed the independence of the means of evidence that supported their hypothesis without any real explanation (Annan and Hargreaves, 2017), but for the evidence that does not support the hypothesis, independence and shared bias seem to be a concern.

Further, in the 2006 paper, they add that even if they had included the two results, it would not have changed their likelihood function by a significant amount. But that is not likely to be the case every time as Stegenga writes that sometimes the evidence that disagrees with our hypothesis may be high quality and its inclusion or exclusion may have a significant impact (Stegenga, 2009). The assumptions of independence also run into problems with failure of ontic independence as explicated by Stegenga and Menon (2017). Failure of means of independent to be ontically independent leads to a case of pseudorobustness, and as Annan and Hargreaves themselves wrote, there is no explanation or demonstration of independence. Hence in absence of established independence, an inference from robustness is actually a case of pseudorobustness (Stegenga and Menon, 2017).

Emergent Constraints:

Emergent Constraints (EC) analysis is a novel form of explanatory RA that is used to constrain likelihood functions and that is currently prevalent in climate science is. As Winsberg explains: In order to perform an EC analysis, we need a climate variable *v* whose long-term likelihood function that we wish to know, and which has a short term (seasonal) analogue *w* (also called an observable metric). We can simulate *w* over some recent period and determine it to be in an interval (a, b), and additionally we need to have reliable, historical observational data (since it is seasonal) that puts *w* in an interval (c, d) where ideally a< c< b< d. There is an emergent functional relationship (correlation) *f* that arises between the inter-model likelihood function of *w* and that of *v.* Ideally, and Winsberg does stress on the ideal aspect here, this would lead to the likelihood of *v* being constrained to an interval *f* (c, d). In order to make that call, we need to fulfil the final condition: there must be a single plausible physical explanation that describes the emergent functional relationship between *v* and *w* (Winsberg, 2018).

Many studies have been carried out that use EC analysis to constrain the spread or likelihood functions of climate variables. Borodina et al. (2017) use the CMIP5 model ensemble to produce constrained ensembles using a functional relationship between future temperature variability and (short term, seasonal) sea ice variability. Their goal is to "quantify the uncertainty in changes in projected temperature variability by combining observations and emergent relationships across models" and they also aim to show that the improved inter-model agreement depends on the choice of observable metrics (*w*) and the skill in selecting models according to biases amongst other things (Borodina et al., 2017). They successfully

demonstrate the power of EC analysis by showing strong relations across models between their selected metrics and projected changes in variability of seasonal temperature and retreat of sea ice. They also show that these relations are consistent with our physical understanding of Arctic climate and they use this physical understanding to constrain the range of projections by evaluating models in conjunction with observational data sets and reanalyses data which seems to be an example of EC analysis and a Schupbach-type explanatory RA.

Any functional relationship that emerges from the simulation result is then suspect of being spurious because there is a high probability that it could emerge by chance. Even in the case of sea ice and temperature variability where the physical explanation is well established, Borodina et al. (2017) stress on trying to reduce the probability of spurious relationships to a minimum, multiple times. They calibrate the ensemble, selecting models which had a low present-day bias relative to observable climate since a low bias is more likely to result in a better local representation. They determine reliability by ranking this calibrated subset according to bias. In addition to this, they write that ignoring the presence of duplicate models (models from the same centre or even across centres share common cores and parametrizations) could led to overestimation of model agreement and an exaggerated reduction in spread (Borodina et al., 2017).

The process described above itself dilutes the independency condition since the authors show that the models that are calibrated and tuned to current climate observations produce better constraints than those that are not. Those supporting an explanatory notion of robustness would argue that independence is not required since all that matters is elimination of competing hypotheses. Schupbach or Winsberg could potentially argue that by calibrating and ranking the ensemble members according to reliability, accounting for double counting of models, and trying to eliminate spurious correlations, the authors' means of detection are RA-diverse for this specific context or relative to this specific hypothesis, which more than makes up for a loss of independence in terms of quality of output.

The EC analysis is based on nailing down a statistical relationship that is found between a present-day observable metric and projections to a single physical explanation. One variable that Winsberg is hesitant to apply EC analysis to is the aforementioned Equilibrium Climate Sensitivity (ECS) because it is a very complex process that is a result of the simulation of many sub-processes, many of which have their own sub-processes. As a result, the final and most important condition for EC: the single plausible physical explanation for ECS cannot be

ascertained (Winsberg, 2018). Borodina et al. (2017) demonstrated how meticulously EC has to be treated even for variables with single physical explanations. Hence, for variables with multiple independent physical influences, it is even more difficult to ignore the possibility that an emergent functional relationship is spurious and because of chance or as Winsberg (2018) writes, "any purportedly uncovered relationship between intermodel variation in a predictor [*w*] and a predictand [*v*]could be the result of compensating errors".

Still, climate science literature has several studies which use EC analysis to constrain the likelihood function of ECS. A more recent analysis of such studies by Qu et al. (2018) has revealed a rather interesting finding. In this analysis, the authors look at four observable metrics for EC analysis conclude that firstly, all four metrics are significantly related to Short Wave (SW) low cloud feedback, which itself is established to be a major source of spread in the likelihood functions of ECS, and it is this relation that in turn drives the relationship between the respective metrics and ECS. Secondly, they concluded that anti-correlation between SW low cloud feedback and other factors may diminish and suppress their correlations with ECS.

The authors note that in recent years, there has been work done by scientists to try and constrain SW low cloud feedback as well as ECS, even though ECS is shaped by many different feedback processes. The four metrics they choose (Sherwood et al., 2014; Su et al., 2014; Tian, 2015; Zhai et al., 2015) are reasonably correlated with ECS. The authors want to check if there is a clear physical basis, because that is required in order to perform an EC analysis. They also want to check how different or similar these correlations are to the emergent constraints' estimates obtained for SW low cloud feedback in order to investigate whether it is the relationship between this feedback and the metrics that drives the relationship between the metrics and ECS.

Although this is not an easy task owing to the complexity of the ECS process in climate models, Qu et al. (2018) develop a statistical framework to investigate the driving factor behind the correlation between ECS and any given metric M across 26 models in the CMIP5 ensemble. Further they employ multivariate regression and backward selection methods to identify feedbacks of ECS that show a statistically significant relation with M, and additionally decompose the relation between ECS and M into the contributions of each feedback. While they used the four metrics mentioned above, Qu et al. also created artificial metrics that were designed to be well correlated to ECS as a control set-up.

The authors found significant correlation with SW low cloud feedback in all of the four selected metrics as well as in all the artificial metrics. Further they conclude that it is this relationship that drives the correlation between the metrics and ECS. They also found that correlations between other forcings in the models and ECS were diminished by anti-correlations between those forcings and SW cloud feedback. For instance, the anti-correlation between Long Wave (LW) cloud feedback and SW cloud feedback cancels out the contribution of LW cloud feedback to ECS (Qu et al., 2018). Their analysis is restricted to a statistical aspect and does not explore the physical connections between the chosen metrics and SW cloud feedback.

In their analysis, Qu et al. (2018) emphasize multiple times on the fact that EC analysis requires one single physical explanation that drives the relationship and that ECS is too complex of a phenomenon which is shaped by multiple processes, forcings, and feedbacks. Winsberg (2018) also argues to some length about why it is not fruitful to carry out this analysis on ECS writing that "we would have to break it down into all of its components… and find ECs for each and every one of them.". Qu et al. suggest that these "so called" emergent constraints may be reframed to be emergent constraints on SW low cloud feedback and not on ECS itself. In fact, they argue that these are only potential constraints and that to become "true" emergent constraints, the physical relationship between each metric and SW low cloud feedback requires more detailed analysis (Qu et al., 2018).

This is where we see the caveat in EC analysis. A single physical, plausible, causal explanation must be established in order to infer that the statistically significant relationship is actually a causal one. What makes EC analysis a type of RA-reasoning is that it involves using emergent functional relationships and observational data to constrain the spread of the variable's likelihood function, and in doing so, scientists are eliminating competing explanations. But there must be a plausible physical relationship between the variable and the metric, as we see with Qu et al.'s results. Because if that is not the case, then the correlation can be explained by something else, and the inference from correlation to causation is not justified. In Schupbach's terms, means of detection are RA-diverse if and only if they eliminate a competing alternative hypothesis. As seen here, the strong correlation between the metrics and ECS can explained by either some causal explanation or by the fact that all four metrics are related to SW cloud feedback which itself is a major forcing contributing to ECS.

The question of independence also arises here and is quickly answered since the metrics turn out to all be related to SW cloud feedback. Not only are the four metrics related to SW cloud feedback, the authors state that any metric that is correlated with climate sensitivity in the CMIP5 ensemble is most likely to also be correlated with SW low cloud feedback, which makes it likely that there is a failure of conditional probabilistic independence, and hence this could be characterized as pseudorobustness. Hence, Qu et al.'s (2018) analysis clearly demonstrates that purported-EC analyses on climate sensitivity were not cases of true emergent constraints. From here it follows that it's likely that a process as complex as ECS cannot be constrained using true emergent constraints. From the discussion above, and from the fact that EC is a type of RA (Winsberg, 2018), it also follows that it's likely that explanatory-reasoning based RA will not work for means of estimating ECS; not until the explanatory intricacies of climate sensitivity have been rigorously worked out.

To summarize, in the experiments we saw in this chapter, there are different kinds of robustness analyses at work. These are accompanied by claims of independence of the means of detection, which are not usually true, although the means are assumed to be independent anyways. As seen in Annan and Hargreaves (2006), they assumed the various observational constraints to be independent without any real explanation or demonstration of said independence. Borodina et al. (2017) calibrate models to current climate and rank according to reliability to in fact demonstrate that skill in picking the right subset is key in establishing emergent constraints. Qu et al. (2018) show that almost all metrics in the CMIP5 ensemble are correlated with SW low cloud feedback and hence, will also likely show a correlation with climate sensitivity.

There are also appeals to explanation in the literature, specifically in the EC literature. Borodina et al. (2017) and Qu et al. (2018) stress multiple times that there needs to be a single physical, plausible causal explanation between the long-term variable to be constrained and the short-term observable metric. If this explanation is missing or we are unsure about it, then the relationship that emerges and the constraints we place on the likelihood function of the variable could be because of some other factor or forcing, or worse, they could be spurious correlations. Even in the case where our physical explanations are quite well established, Borodina et al. take extreme care to eliminate spurious correlations. On the other hand, it looks like the enthusiastic and ambitious work that is going on in the form of EC analysis on climate sensitivity may be not be a good idea since we definitely lack explanatory clarity in this case (Qu et al., 2018; Winsberg, 2018).

Conclusion:

An exhaustive study of the philosophical literature as well as that of climate science reveals some very important things. We have learned what supposedly goes into robustness and what supposedly comes out of it. Many in the philosophical literature claim that robustness requires the modes of evidence be at least ontically independent, but Stegenga and Menon (2017) prove that ontically independent and conditionally probabilistically independent evidence are necessary but not sufficient conditions to warrant a robustness argument. While Lloyd (2015) argues that such conditions are sufficient to warrant confirmation from robustness, Stegenga and Menon disagree. Their position is that both types of independence allow us to avoid two different types of pseudorobustness and hence, they still have some epistemic value.

Schupbach (2016) on the other hand argues that no notion of conditional independence has any importance when robustness is concerned. The only notion that matters is explanatory power. As long as every additional means of detection eliminates a competing hypothesis, it is RA-diverse, and we can make an inference from robustness. This extends across all means of detection – models, observations, experiments, paleoclimatic data (Winsberg, 2018). Nevertheless, the literature of climate science reflects that conditional independence is considered an important and desired virtue of the evidence.

Failure of ontic independence seems become an issue for explanatory reasoning-style RA-diversity. If means of detection are ontically independent, then there's no question of the result being explained by any shared dependencies. But if we fail to establish ontic independence, then it is possible that the result is an artefact of some shared dependencies or similarities, a hypothesis that cannot be eliminated. Looking at the different modes of evidence in climate science, it is evident that they are not ontically independent – models share structural and design similarities, idealizations, they are tuned to current climate data, gaps in observed datasets are filled with weather datasets, and so on.

The reason why Winsberg (2018) discourages the use of emergent constraints (EC) analysis on a process as complex as Equilibrium Climate Sensitivity (ECS), can be argued to be the same reason given above – at the level of complexity of ECS, ontic independence fails. As we saw in Qu et al. (2018), what were thought to be emergent constraints between chosen metrics and ECS were actually correlations driven by the relation between those metrics and

SW cloud feedback, a major contributing feedback in ECS. This means that the relationship between each of the metrics and ECS was explained by an artefact of a dependency. Ontic independence cannot be ensured by changing metrics because any metric correlated with ECS in CMIP5 will be likely correlated with SW cloud feedback.

The only way to ensure ontic independence is to step down the levels from ECS until we arrive at feedbacks or forcings that can be explained with a single, plausible, physical explanation. At this level we can be much more reasonably sure that the emergent constraints have only one explanation and that there is no alternative hypothesis, as was demonstrated by Borodina et al. (2012). We move upward, working out the physical, causal basis and emergent constraints for each forcing and sub-process until we have a complete, coherent causal explanation for ECS (Winserg, 2018; Qu et al., 2018). That is a mammoth task as even in the case of Borodina et al., for just one forcing and one variable, they had to be very careful about screening out spurious correlations and ruling out alternate hypotheses, that too for a case where the causal explanation we require was well established.

Although Winsberg and Qu et al., have already explicated this, the goal here is to argue that the risk of failure of ontic independence increases with increase in the complexity of the process or feedback that is the target of the EC analysis. At higher levels of complexity, there are more processes that might be correlated, and this correlation may inflate (as in the case of SW cloud feedback) or diminish (in case of LW cloud feedback) the correlation of other forcings. While at a level of a singular process or forcing, nothing else can reasonably affect the correlation except the actual purported physical explanation.

The epistemic gap or the problem of completeness continues to be a problem, albeit, it figures amongst what Lloyd calls "scientifically unrealistic" requirements along with truth. We don't have access to an ideal representation of the climate system, in fact, we probably would not know if we even had one. We don't know the truth about future conditions and hence whether or not our predictions are correct, close, or incorrect until the readings can be corroborated at that future time. As we have seen from Parker (2009, 2011), claims of adequacy-for-purpose, inferences to truth, increased confidence in the hypothesis, and increased security of the evidence cannot be successfully made using robustness because of these reasons: truth and completeness. Whatever notion that is appealed to in light of model agreement, could be argued to be another type of pseudorobustness.

Similarly, if we consider the extension of Katzav's argument from Chapter II regarding explanatory-instrumental virtues that are conferred to an ensemble from its members, which in turn get them from their respective explanatory parts (causal core) and instrumental parts (parametrizations); the implication is that warrants we have from the ensemble cannot be attributed solely to the explanatory virtue, but to the explanatory-instrumental virtue, that is to say that it could be an artefact of some parametrization instead of the causal core. Hence, as Katzav (2012) argues, these warrants are not the same as those that we have in light of inference to the best explanation (IBE). This seems to be directly knocking the realist no-miracles argument for robustness – it is the best explanation for our models providing concordant evidence, the only other way it could be true is a miracle. Once again, what occurs in this case could be argued to be another kind of pseudorobustness.

Hence the requirements of true robustness are very high – explanatory reasoning-based RA-diversity, ontic independence, truth, and completeness. To achieve RA-diversity, we need ontic independence. To achieve ontic independence, we need to perform detailed physical and EC analysis at the most elementary, single process level of forcings. Still, these requirements are implausible, not impossible. It would take a long time, many people, and a lot of computing power; but we could plausibly carry out the large-scale exercise of establishing ECs for climate sensitivity. It is impossible to know the truth of the future climate system right now as it also impossible to create a complete representation of the system. In light of these strict requirements, all other appeals to robustness are simply cases of pseudorobustness. If we do manage to carry out the full EC exercise for climate sensitivity and hence acquire proper ontic independence required for RA-diversity, then we seem to be using a sort of stronger form of pseudorobustness that is closer to the true form, but still not complete.

References:

1. Abramowitz, G., & Gupta, H. (2008). Toward a model space and model independence metric. *Geophysical Research Letters*, *35*(5).
2. Annan, J. D., & Hargreaves, J. C. (2006). Using multiple observationally-based constraints to estimate climate sensitivity. *Geophysical Research Letters*, *33*(6).
3. Annan, J. D., & Hargreaves, J. C. (2017). On the meaning of independence in climate science. *Earth System Dynamics*, *8*(1), 211-224.
4. Borodina, A., Fischer, E. M., & Knutti, R. (2017). Emergent constraints in climate projections: a case study of changes in high-latitude temperature variability. *Journal of Climate*, *30*(10), 3655-3670.
5. Frigg, R., Bradley, S., Du, H., & Smith, L. A. (2014). Laplace's demon and the adventures of his apprentices. *Philosophy of Science*, *81*(1), 31-59.
6. Horwich, P. (1982). How to choose between empirically indistinguishable theories. *The Journal of Philosophy*, *79*(2), 61-77.
7. Justus, J. (2012). The elusive basis of inferential robustness. *Philosophy of Science*, *79*(5), 795-807.
8. Katzav, J. (2012). Hybrid models, climate models, and inference to the best explanation. *The British Journal for the Philosophy of Science*, *64*(1), 107-129.
9. Ladha, K. K. (1995). Information pooling through majority-rule voting: Condorcet's jury theorem with correlated votes. *Journal of Economic Behavior & Organization*, *26*(3), 353-372.
10. Lambert, S. J., & Boer, G. J. (2001). CMIP1 evaluation and intercomparison of coupled climate models. *Climate Dynamics*, *17*(2-3), 83-106.
11. Levins, R. (1966). The strategy of model building in population biology. *American scientist*, *54*(4), 421-431.
12. 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.
13. Lloyd, E. A. (2009, June). I—Elisabeth A. Lloyd: Varieties of Support and Confirmation of Climate Models. In *Aristotelian Society Supplementary Volume* (Vol. 83, No. 1, pp. 213-232). Oxford, UK: Oxford University Press.

14. Lloyd, E. A. (Ed.). (2010). Confirmation and robustness of climate models. *Philosophy of Science*, *77*(5), 971-984.
15. Lloyd, E. A. (2012). The role of 'complex'empiricism in the debates about satellite data and climate models. *Studies in History and Philosophy of Science Part A*, *43*(2), 390-401.
16. Lloyd, E. A. (2015). Model robustness as a confirmatory virtue: The case of climate science. *Studies in History and Philosophy of Science Part A*, *49*, 58-68.
17. Masson, D., & Knutti, R. (2011). Climate model genealogy. *Geophysical Research Letters*, *38*(8).
18. Mastrandrea, M. D., Field, C. B., Stocker, T. F., Edenhofer, O., Ebi, K. L., Frame, D. J., ... & Plattner, G. K. (2010). Guidance note for lead authors of the IPCC fifth assessment report on consistent treatment of uncertainties.
19. Mears, C. A., Schabel, M. C., & Wentz, F. J. (2003). A reanalysis of the MSU channel 2 tropospheric temperature record. *Journal of Climate*, *16*(22), 3650-3664.
20. Odenbaugh, J. (2011). True lies: Realism, robustness, and models. *Philosophy of Science*, *78*(5), 1177-1188.
21. 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(4), 533-546.
22. Parker, W. S. (2009, June). II—Confirmation and adequacy-for-purpose in climate modelling. In Aristotelian Society Supplementary Volume (Vol. 83, No. 1, pp. 233-249). Oxford, UK: Blackwell Publishing Ltd.
23. Parker, W. S. (2011). When climate models agree: The significance of robust model predictions. Philosophy of Science, 78(4), 579-600.
24. Parker, W. S., & Winsberg, E. (2018). Values and evidence: how models make a difference. *European Journal for Philosophy of Science*, *8*(1), 125-142.
25. 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.
26. Qu, X., Hall, A., DeAngelis, A. M., Zelinka, M. D., Klein, S. A., Su, H., ... & Zhai, C. (2018). On the emergent constraints of climate sensitivity. *Journal of Climate*, *31*(2), 863-875.

27. Sanderson, B. M., Wehner, M., & Knutti, R. (2017). Skill and independence weighting for multi-model assessments.
28. Santer, B. D., Wehner, M. F., Wigley, T. M. L., Sausen, R., Meehl, G. A., Taylor, K. E., ... & Brüggemann, W. (2003). Contributions of anthropogenic and natural forcing to recent tropopause height changes. *science*, *301*(5632), 479-483.
29. Stegenga, J. (2009). Robustness, discordance, and relevance. *Philosophy of Science*, *76*(5), 650-661.
30. Stegenga, J., & Menon, T. (2017). Robustness and independent evidence. *Philosophy of Science*, *84*(3), 414-435.
31. Schupbach, J. N. (2016). Robustness analysis as explanatory reasoning. The British Journal for the Philosophy of Science, 69(1), 275-300.
32. 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.
33. Van Fraassen, B. C. (2010). Scientific representation: Paradoxes of perspective.
34. Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.
35. Weisberg, M. (2006). Robustness analysis. *Philosophy of Science*, *73*(5), 730-742.
36. Wimsatt, W. C. (1981). Robustness, reliability, and overdetermination. *Scientific inquiry and the social sciences*, 124-163.
37. Wimsatt, W. C. (1994). The ontology of complex systems: levels of organization, perspectives, and causal thickets. *Canadian Journal of Philosophy*, *24*(sup1), 207-274.
38. Winsberg, Eric. "What does robustness teach us in climate science: a re-appraisal." *Synthese* (2018): 1-24.
39. Woodward, J. (2006). Some varieties of robustness. Journal of Economic Methodology, 13(2), 219-240.

Urkund Plagiarism Check Report

URKUND

Urkund Analysis Result

Analysed Document:	Parjanya Joshi MA Dissertation Final - M2017CCSS007.docx (D49090981)
Submitted:	3/14/2019 10:37:00 AM
Submitted By:	urkund.mumbai@tiss.edu
Significance:	4 %

Sources included in the report:

https://pdfs.semanticscholar.org/4280/5b4cae8b3172ab8fc50f92cff758f414de11.pdf
https://link.springer.com/article/10.1007%252Fs11229-018-01997-7
https://link.springer.com/article/10.1007/s13194-015-0117-x
59fc1518-b240-4961-a22a-bd716ce0869f
a4ef228b-7b66-4e27-bcb2-d32456b9bfdd
9bc49116-e34d-45aa-95bf-b45df92461d2

Instances where selected sources appear:

23