Synthese (2021) 199:745–765
https://doi.org/10.1007/s11229-020-02707-y

COMPUTATIONAL MODELLING IN PHILOSOPHY

The strategy of model building in climate science

Lachlan Douglas Walmsley[1]

Received: 1 March 2019 / Accepted: 15 May 2020 / Published online: 25 May 2020

Abstract
In the 1960s, theoretical biologist Richard Levins criticised modellers in his own discipline of population biology for pursuing the "brute force" strategy of building hyper-realistic models. Instead of exclusively chasing complexity, Levins advocated for the use of multiple different kinds of complementary models, including much simpler ones. In this paper, I argue that the epistemic challenges Levins attributed to the brute force strategy still apply to state-of-the-art climate models today: they have big appetites for unattainable data, they are limited by computational tractability, and they are incomprehensible to the human modeller. Along the lines Levins described, this uncertainty generates a trade-off between realistic, precise models with predictive power and simple, highly idealised models that facilitate understanding. In addition to building ensembles of highly complex dynamical models, climate modellers can address model uncertainty by comparing models of different types, such as dynamical and data-driven models, and by systematically comparing models at different levels of what climate modellers call the model hierarchy. Despite its age, Levins' paper remains incredibly insightful and should be considered an important entry into the philosophy of computational modelling.

Keywords Climate models · Robustness analysis · Modelling strategies · Model pluralism · Levins · Model trade-offs

1 Introduction

Computers have had a profound impact on science, revolutionising how humans record and represent the world and our ideas. In model-based science (c.f. Giere 1988, 2010; Godfrey-Smith 2006; Weisberg 2013), where mathematical constructs are investigated to make inferences about theoretical principles or the world, computers have enabled scientists to investigate increasingly complex and realistic model systems (Humphreys

✉ Lachlan Douglas Walmsley
Lachlan.walmsley@anu.edu.au

[1] School of Philosophy, Australian National University, Building 9 Fellows Rd, Acton, ACT 2600, Australia

2004). In the 1960s, theoretical biologist Richard Levins argued that it would be naïve to think that we—living in an imperfect world and without god-like powers—could ever create perfect copies of our world (Levins 1966). And yet it appeared to Levins that this "brute force approach" was precisely the philosophy behind a then new research paradigm in biology. Systems ecologists like Kenneth Watt argued that the simple mathematical models Levins and his colleagues built were too unrealistic to reveal anything of worth about real-world targets (Watt 1956). Watt claimed that only computer simulations could capture all the relevant features of ecological target systems, so we should trust them more than simple mathematical models. For Levins, on the other hand, the brute force approach could not be pursued alone due to three epistemic challenges that highly complex models face: they have a large appetite for data that may be difficult or impossible to acquire; they are computationally intractable; and they are incomprehensible to the human modeller. My aim in this paper is to consider Levins' analysis in the context of contemporary climate modelling. Despite the many decades that have passed and the different subject matter, his arguments are surprisingly relevant.

Levins argued that modellers cannot hope to build one single best model because there are different modelling aims—realism, precision, and generality—that cannot be jointly maximised.[1] John Matthewson (2011) has argued that these trade-offs are particularly problematic within Levins' own field of population biology due to the heterogeneity among ecological target systems: if you pick two ecosystems at random, like the Great Barrier Reef and Yosemite National Park, they are unlikely to share all their important causal features. Making a model more realistic or precise with respect to one target system, then, necessarily reduces how general it is by misrepresenting targets that do not share the properties described in the model and possessed by that first target (Matthewson and Weisberg 2009).

Although causal heterogeneity might be the primary source of the trade-offs in ecology, it does not explain the most important trade-offs in climate modelling. This is because climate modellers frequently aim to describe the one target system over a relatively short period—Earth between 1850 and 2100—thus avoiding the problem of heterogeneity and any demand for generality. Instead, the trade-offs in climate science are better explained with Jay Odenbaugh's (2003, 2006) view that the three epistemic challenges facing brute force models create a demand for simpler models that require less data, require less powerful computers, and, most importantly, are easier to understand. Consequently, the trade-off in climate science is primarily between realistic and precise models with predictive power and much simpler and highly idealised models that facilitate understanding (Charney 1963).

Levins (1966) suggested that model uncertainty could be addressed by building a family of models with different assumptions, a procedure known as robustness analysis (RA). RA has been discussed at length within the philosophy of climate science (e.g. Lehtinen 2016, 2018; Lloyd 2010, 2015; Parker 2011, 2013; Winsberg 2018a, b). A prominent critical view is that ensembles of highly complex climate models do little

[1] Following Michael Weisberg (2006a), I take a more realistic model to be one that explicitly represents more of its target's causal structure, a more precise model to be one with more finely specified parameters, and a more general model to be one that applies to more actual or possible targets. I describe these trade-offs further in Sect. 3.

to reduce model uncertainty because these ensembles represent a very small sample of possible model space, the models are not varied systematically, and the models are not independent (for a more optimistic view see Lehtinen 2016, 2018; see also Weisberg, 2006b). In this paper, I will argue that we should instead consider how RA applies to the practice of comparing complex models with simpler ones, progressing systematically up, down, and through what climate scientists refer to as the model hierarchy.

Here is the plan for the paper. In Sect. 2, I will describe the brute force approach taken in climate modelling today and show how Levins' three epistemic challenges apply. In Sect. 3, I will argue that these problems drive the trade-offs that climate scientists face, rather than the need to account for causal heterogeneity among targets, and that modellers must balance realism and understandability. In Sect. 4, I argue that RA and the systematic movement through model hierarchies can establish a better understanding of key climate processes. In Sect. 5, I conclude.

2 Brute force

The "brute force" approach of systems ecology was to build complex computational models that included as much causal detail about the target system as possible (Watt 1962; Watt and Watt 1968). At first glance, the brute force approach appears to characterise much of climate modelling today.

Increasingly powerful digital computers have enabled modellers to represent the climate system at a higher resolution and to explicitly represent more and more climate processes and components (Washington et al. 2009). Earth system models (ESMs) are a family of climate models that are among recent attempts by climate modellers to represent the climate system in as much detail as possible. They are called "Earth system" models because they include processes required to complete the carbon cycle, distinguishing them from similar models, known as general or global circulation models (GCMs), which are less comprehensive in this respect.

One component ESMs and GCMs share is a dynamical core. This part of the program numerically approximates the so-called governing equations, a set of seven or eight (seven for the atmosphere and eight for the ocean) non-linear differential equations from the laws of fluid dynamics, to represent the circulation of the atmosphere and oceans (Bjerknes 1904; Edwards 2010; Washington and Parkinson 2005, p. 49). Norwegian physicist Vilhelm Bjerknes published the governing equations in 1904, but the equations were analytically intractable, so they could not be solved. In 1922, English mathematician Lewis Fry Richardson developed mathematical techniques for numerically approximating these equations, turning the differential equations into difference equations. Using observations taken on International Balloon Day of 1910, a day for which the observational record was unusually good, Richardson attempted to retroactively produce a weather forecast (Edwards 2010). It took Richardson six weeks to produce his six-hour forecast, which, owing to a calculation error, was terrifically inaccurate.

With the assistance of powerful digital supercomputers, Bjerknes' primitive equations can be numerically approximated in a relatively timely fashion with, hope-

fully, fewer errors than Richardson's calculations. ESMs still bear some similarities with Richardson's model. Richardson proceeded by carving Europe into a three-dimensional grid of 90 cells and prepared 23 programs for turning observed data into a forecast. While some climate models remain regional, ESMs use a 3D grid that represents the atmosphere, the oceans, and land of the entire planet. Using this grid, the computer steps through discretised versions of the governing equations. Grid sizes vary, from coarser grids spanning 200 km squared along the Earth's surface, to finer grids spanning 20 km squared (Neelin 2010, p. 150). Generally, the finer the resolution the better, but finer resolution comes at the cost of computational tractability, which places a cap on just how closely the primitive equations can be approximated.

While a significant part of the Earth system—the atmosphere and the ocean—can be represented as a circulating fluid, there are many components that make a difference to climate behaviour that cannot be represented using fluid dynamics alone, such as sea ice and vegetation. So, the other major component of ESMs and GCMs is the model physics, which represents these additional parts of the climate system. The number of relevant processes represented in highly complex climate models has steadily increased in step with easing computational limitations (Washington et al. 2009).

ESMs and GCMs, I suggest, are examples of the brute force approach. Including as much detail as possible is not a bad strategy if the details matter and in ways that may surprise us due to feedback loops and the connectedness of the system, or if the details are relevant for the kinds of questions we wish to answer with our models (Shukla et al. 2010). This is very much the case in climate science. However, ESMs and GCMs face a notorious amount of uncertainty largely because of the three problems that Levins attributed to the brute force approach in his classic paper (Levins 1966, p. 241):

(a) There are too many parameters to measure; some are still only vaguely defined; many would require a lifetime to measure.
(b) The equations are insoluble analytically and exceed the capacity of even good computers.
(c) Even if soluble, the result expressed in the form of quotients of sums of products of parameters would have no meaning for us.

I refer to these kinds of problem as (a) the problem of data hunger, (b) the problem of tractability, and (c) the problem of comprehensibility. In the remainder of this section, I will demonstrate that high-fidelity climate models face these problems and that practices used to manage these problems can themselves be sources of uncertainty.

2.1 The problem of data hunger

The first problem Levins identified with highly detailed computational models is that a large amount of data is required to calibrate these details. Climate science is a very data hungry discipline indeed. Measurements are needed to adjust parameters, set initial conditions,[2] and evaluate model performance. As Edwards argues in his (2010), climate science as a discipline is predicated on a fairly recent global infrastructure of

[2] In some conditions, models do not rely on specific initial conditions. ESMs and GCMs, in fact, are allowed to "spin-up" for some simulation time, falling into their own natural equilibria before forcing scenarios, such as different carbon emission schemes, are imposed and the model system moves away from its equilibrium.

weather stations along with the more advanced technology required to take measurements beyond the Earth's surface, such as weather balloons and satellites. This is to say nothing of the technological requirements necessary for feasibly storing and sharing the large amounts of weather data needed to describe the climate. As Levins argued, data hunger becomes particularly problematic when the relevant data is difficult or impossible to collect, as it is in climate science today.

Here's one way to evaluate how well a model represents Earth's climate: compare the model's output to real-world data (Winsberg 2018a). To do that, we need some data, and while our international system of weather and climate measurement is better today than it was on Balloon Day 1910, our weather stations, ocean buoys, weather balloons, and so on, are not distributed evenly across the Earth's surface, up into the atmosphere, or down into the oceans and soil. Certainly not with a density proportional to the 3D finite difference grid of an ESM. These grids can have points every 100 km across the horizontal of the Earth's surface and at far smaller intervals than that along the vertical up into the atmosphere and down into the ocean. We simply do not have this much equipment. Moreover, as the resolution of climate models increases, so do their appetites for data.

Scientists could, in principle, cover the world in measuring equipment going into some science fiction future—we have smart homes, there is a growing literature on smart cities (see Meijer and Bolívar 2016), so why not a smart Earth? Unfortunately, there would be no way to deploy such information infrastructure on any past Earth. Climate records of the Nineteenth and Twentieth century will never grow, but this is a key period used to evaluate model performance: "data are available for only a few quantities (e.g. temperature, pressure, precipitation), for only relatively recent time periods, and primarily for land locations and near-surface locations, and even these records are incomplete and of variable quality" (Parker 2006, pp. 353–354). Historical data only gets patchier the further back in time we go.

Scientists, ever ingenious, have developed techniques to fill the gaps in the historical records by calculating the likely values for the missing data points. Data sets produced through such a process are known as reanalysis data sets (Parker 2016). Reanalysis, however, is its own source of uncertainty. Reanalysis data sets are constructed using weather models that are not identical to ESMs, but that have a shared history and are based on numerical approximations of the same physical equations representing fluid flow. This similarity introduces epistemic complications when evaluating a model against a reanalysis data set (Lewis 2017). While a fit between a model M and data is typically a good sign, fit between M and a data set partly constructed using algorithms resembling those numerically approximating the same equations in M—as in the case of reanalysis—provides comparatively less confidence about the quality of M. This is because the similar data and model output may be explained by appeal to a common causal factor—approximation of the same dynamical equations—rather than by appeal to the model's skill at reproducing natural patterns.

2.2 The problem of tractability

Levins argued that highly complex models are limited by computational power. Running a simulation that included every causal detail about an ecosystem was simply not feasible given 1960s technology. Causal completeness still isn't feasible in climate science today (Winsberg 2018a). The ideal model, then, cannot be investigated until we make more powerful computers. Until then, we are stuck with non-ideal models that are made with rough approximations, ad hoc adjustments, and parameterisations.

Bjerknes' governing equations, for example, are based on well-established pieces of physical theory, which we might trust to represent atmospheric and oceanic flows. But these equations must be numerically approximated and finer approximations require more computational power, which we do not have. Approximations are made rougher through truncation and rounding errors resulting from memory and computational limitations. The result of some calculation at one grid point may have a long string of numbers after the decimal point, not all of which can be stored in the computer's memory, so the numbers after a certain position, such as the fourth decimal place, are simply dropped and forgotten (truncation), or dropped and forgotten after rounding the last remaining digit up or down (rounding). Although a single instance of truncation or rounding may make little difference, these little differences can accumulate as the simulations run over long time scales and the result of one calculation is used as the input for the next.

Computational limitations also force modellers to use parameterisations to represent important processes that occur at scales too small to be resolved within an ESM's finite difference grid. For example, shallow cloud formation in the lower atmosphere is a major contributor to Earth's Albedo factor—that is, how much incoming solar radiation is reflected out of the system and doesn't contribute to the Earth's energy budget—and consequently to the Earth's climate. Interactions with aerosols can also determine droplet size and alter the cloud's albedo (Pasini 2005, pp. 126–127). Cloud behaviour, however, occurs on scales that even the finer grids on the market cannot resolve (Schneider et al. 2017, p. 4):

> atmosphere models have a horizontal grid spacing around 50–100 km and a vertical grid spacing in the lower atmosphere around 200 m. This is much too coarse to resolve the 10–100 m wide turbulent updrafts that originate in the planetary boundary layer and generate low clouds.

Tapio Schneider and colleagues calculate that, given current grid spacing, the improvements in grid spacing required to resolve low clouds, and the pace at which computational power advances, sufficiently fine grids will not be available until the 2060 s. Until then, shallow cloud formation must be parameterised. Unfortunately, we do not yet understand this process or how best to model it (Stocker 2014), and different cloud parameterisation schemes account for the bulk of the disagreement among climate model projections (Schneider et al. 2017, p. 4). As stated in the Intergovernmental Panel on Climate Change's Fifth Assessment Report: "There is *very high confidence* that uncertainties in cloud processes explain much of the spread in modelled climate sensitivity" (Stocker 2014, p. 743).

In response to parameterisation use and computational limitations, a model's equations may be adjusted in ad hoc ways (Edwards 2010; Lenhard and Winsberg 2010; Parker 2006; Winsberg 2010). This can occur during a process known as model tuning: "Tuning a climate model involves making ad hoc changes to its parameter values or to the form of its equations in order to improve the fit between the model's output and observational/reanalysis data" (Parker 2011, p. 587). Rough approximation, parameterisation, and ad hoc adjustment are responses to computational limitations, which obscure the connection between the model's behaviour and the solution to the theoretically principled equations. This is a problem if we rely on the credentials of the governing equations to justify the inferences we make with our model.

2.3 The problem of comprehensibility

Levins argued that highly complex computational models are incomprehensible to human scientists. Complex climate models are millions of lines of code long, typically scripted by teams of scientists, and run on powerful supercomputers, as opposed to a couple of equations, which can be investigated using back-of-the-envelope reasoning that might characterise Levins' simple theory approach. In contrast to simple models, the inner workings of ESMs and GCMs are too complicated, and the output a number array too large, for any human scientist to wrap their head around.

To some extent, the problem of understanding complex simulations has been addressed with computer-assisted analysis techniques which take number arrays and convert them into some form that is comprehensible to a human. Most notably in computational modelling more generally, visualisation techniques can be used to present the values of variables over time in a fashion that appeals to intuitive perception. Some philosophers have even argued in the past that visualisation is an essential aspect of computer simulation (Hughes 1999; Humphreys 2004, pp. 112–114; Rohrlich 1990, p. 515). As Eric Winsberg argues, visualisation appealing to intuitive perception is just one technique among many for analysing simulation behaviour, but such computer-assisted techniques are required to make sense of the large number arrays produced by a complex computational model in just the same way that they are required for making sense of large data sets (Winsberg 2010, p. 33). Of course, there is a danger that these techniques introduce epistemic problems of their own. For one thing, a sense of understanding is often a poor guide to the truth (Trout 2016), and images that present complicated data in a digestible form are seductive but have the capacity to mislead (Klein 2010), so may support a false sense of understanding or confidence.

Complex models like ESMs present a challenge for understanding. Representing the key components of the climate system together in one comprehensive model can obscure causal attribution within the model: "Interpretation of cause and effect linkages may be difficult to trace in a GCM because of the large number of internal degrees of freedom in the model and because of the huge volume of output generated by a high-resolution time-dependent model" (Schneider and Dickinson 1974, p. 486). It can also create the possibility of compensating errors (Winsberg 2010, 2018a, pp. 196–197). That is, when a model gets something wrong (or right, for that matter), we won't know where to place the blame. A result that seemingly vindicates the model—say, a match

between the model's performance and historical data—could be caused by errors in model components that compensate for one another when coupled. This becomes a real problem when we move to modelling future climate scenarios that are radically different from past or present climates. Here, the errors that had once cancelled each other out may no longer do so, leaving us with a model that appears accurate given the data we do have, but is nevertheless wildly inaccurate for the future cases that matter for large-scale decision-making and planning.

3 Modelling trade-offs

3.1 Realism, generality, and precision?

The appropriate response to the problems described in Sect. 2 is not the cessation of all brute force modelling operations. As Levins argues, modellers have different desiderata that they may wish to satisfy with their models:

> It is of course desirable to work with manageable models which maximise generality, realism, and precision toward the overlapping but not identical goals of understanding, predicting, and modifying nature. But this cannot be done. (Levins 1966, p. 422)

On this view, we cannot produce a single model that maximises generality, realism, and precision because these desiderata compete such that modellers can maximise only two of these desiderata at a time (Matthewson and Weisberg 2009).

Levins did not define his concepts beyond an intuitive understanding of them, so I follow Michael Weisberg's (2004, 2006a) analysis of the distinction between realism, generality, and precision. If a modeller aims for realism, they aim to represent as much of the target's causal structure as possible. As stated in Sect. 2, climate modellers have increased the number of processes they include in their GCMs and ESMs over time, suggesting that they aim for this desideratum. Weisberg distinguishes between two kinds of generality. A-generality refers to the number of actual target systems a model describes, and the second kind, p-generality, refers to the number of "logically possible" target systems a model describes. Finally, Weisberg describes precision as "the fineness of specification of the parameters, variables, and other parts of the model descriptions" (Weisberg 2006a, p. 636). Note that precision is not the same thing as accuracy.

Levins argued that a consequence of the trade-offs between these three desiderata was that a single model could exemplify two of them at best. This led Levins to propose a trichotomy of modelling approaches, each sacrificing one desideratum in order to meet the other two. There is a sizable literature regarding Levins' specific proposal (e.g. Levins 1993; Matthewson and Weisberg 2009; Orzack and Sober 1993; Weisberg 2006b). However, I will argue for a different trade-off in the remainder of the paper, so there is no need for further details on Levins' three model types here.

3.2 What generates the trade-offs

Based on close textual analysis of Levins' work (1966, 1968b, 1973), Odenbaugh (2003, 2006) argues that Levins' (1966) was primarily motivated by the epistemic problems facing the brute force approach, described in Sect. 2: the problems of data-hunger, intractability, and incomprehensibility. Although it may be possible in principle to build a highly realistic, general, and precise model, it would be impossible in practice given the empirical, cognitive, and computational limitations that constrain real modellers. The uncertainty facing the brute force approach, then, can be addressed with different types of models that do not aim at high-fidelity representation and so do not suffer from the same epistemic problems. In response, Matthewson argues that the trade-offs in population biology are due to the inescapable heterogeneity among biological systems. In this section, I defend Odenbaugh's view that the modelling trade-offs are generated by pragmatic empirical, computational, and cognitive constraints, at least in the case of climate science.

Let's start with Matthewson's view. Matthewson (2011, p. 328) argues that Odenbaugh's response undersells the importance of Levins' analysis because it focuses on contingent factors rather than inescapable features of the logic of representation (Matthewson and Weisberg 2009) or the subject matter of the discipline:

> We have presumably already overcome many of the practical limitations that existed for Levins and his peers in 1966… the more that Odenbaugh convinces us that these trade-offs are only due to contingent limitations, the less compelling Levins' claims become.

Providing grist to Matthewson's mill, Odenbaugh (2006), argues that the epistemic challenges Levins describes are not as striking in contemporary biology as they once were. Rather, new computational and mathematical techniques, such as agent-based modelling and matrix algebra, have at least partially addressed the problems of tractability and comprehensibility. The measurement problem remains however, and, on his view, has always been "the most serious problem for population biology" (Odenbaugh 2006, p. 620).

Instead of seeing the trade-offs as a pragmatic problem generated by the *complexity* of dynamic ecological system, Matthewson argues that the trade-offs, and their unique force within biology, are generated by the *heterogeneity* among different ecological systems. As Matthewson notes, Levins was aware of heterogeneity in his original paper: "the multiplicity of models is imposed by the […] demands of a complex, heterogeneous nature…" (Levins 1966, p. 431; c.f. Matthewson 2011, p. 326). We could build a maximally precise and realistic model of an Airbus A380 that also describes the causal structure of all the A380s, Matthewson argues, because their causal structure, though complex, is homogeneous (Matthewson 2011, p. 331). Having a complete replica of the Great Barrier Reef ecosystem, however, does not give us a complete replica of the Yellowstone National Park ecosystem because the two complex systems are causally distinct in many important ways (see also Weisberg 2004, p. 1078).

Heterogeneity is especially problematic in population biology, Matthewson argues, because population biologists study populations under natural selection and the first two conditions of natural selection demand difference-making variation. Condition 1

states that there must be phenotypic variation within a population, and condition 2 states that this variation must have consequences for survival and reproduction. The third condition—heritability—ensures that this variation endures.

Matthewson also takes heterogeneity to be the main driver of the trade-offs because, as he sees it, other disciplines which do not deal with heterogeneous systems do not discuss trade-offs (2011, p. 330):

> The fact that trade-offs hold more in population biology than in other natural sciences is evidenced by the fact that although Levins' work influenced many population biologists, his ideas did not noticeably filter through to the other natural sciences… if trade-offs are important and ubiquitous in modelling, then even if physicists or chemists had never heard of Levins or his work, we would expect them to have their own version of "The Strategy" in the relevant literature.

If Matthewson is right and heterogeneity among target systems really is the primary driver of the trade-offs, then this aspect of Levins' work is unlikely to apply well in the context of climate science. Afterall, climate scientists are typically interested in one target system: The Earth's climate. Moreover, climate modellers often aim to describe that system for a brief window of time: between the years 1850 and 2100.[3]

One possible response here is that, although high-fidelity climate models typically investigate one system—the terrestrial climate—some need for generality remains. Alkistis Elliott-Graves (2018, p. 1109) distinguishes between two kinds of heterogeneity. Intersystem heterogeneity involves variation across a number of different systems, and intrasystem heterogeneity involves variation within a single system across time. Although there is only one actual terrestrial climate, there are multiple possible climate scenarios, which may be enough to create problematic intersystem heterogeneity. For example, modellers want to investigate the Earth where everyone slowly stops emitting by 2050 and the Earth where everyone continues to emit as usual. Representing different emissions scenarios does not create the kind of intersystem heterogeneity Matthewson describes, however, because a single brute force model—that is, a single simulation program—can represent these different scenarios by manipulating the one model rather than by building different models.

A convincing case for problematic intrasystem heterogeneity in climate science is likewise lacking. Although the causal structure of the climate system could change in principle—humans could, in principle, introduce a new artificial component into the climate system—the changing climate is, for the most part, a matter of the climate system and its components occupying different states. It is not a matter of structural change as you might see in ecology when an invasive species is introduced into an ecosystem. While some of these state changes have big consequences for the rest of the state of the system, as in the case of climate tipping points, this should still be conceived

[3] This claim is based on what can be found in the Intergovernmental Panel on Climate Change's Summary for Policymakers of *The Physical Science Basis* (Stocker, 2014), a report which focuses specifically on the phenomena of Twentieth and Twentifirst Century climate change. Of course, many climate modellers are interested in different time periods. Paleoclimatologists, most obviously, investigate phenomena in the distant past, far outside of the roughly 350-year window that is most relevant to anthropogenic climate change. For the purposes of this paper, however, I am focusing only on one (very prominent!) branch of climate science.

of as a change to the system's state rather than the system's structure. Just as with different emissions scenarios, a single high-fidelity model, if it correctly represents the underlying physical processes, can represent both a pre- and post-warming Earth.

To the extent that climate modellers grapple with intrasystem causal heterogeneity, the problem is caused by empirical and computational limitations, rather than, as it in ecology, the very nature of the subject matter. As discussed in Sect. 2, parameterisations are often used to describe poorly understood processes top-down precisely because we do not have the technology to resolve those processes bottom-up from better understood underlying physics. ESMs and GCMs are often tuned to available data sets, including reanalysis data sets, a process through which they are adjusted to achieve a better fit with the data and, hopefully, become a more reliable guide to future climate behaviour in the face of different emissions scenarios. However, these adjustments may make ESMs less likely to generalise successfully to future cases if they are very much unlike the past in unforeseen ways. This is a real possibility. A model biased towards historical conditions, for example, may not accurately represent and predict the behaviour of an altered climate system.

Although Matthewson might be right about the primary source of trade-offs in population biology, Odenbaugh's view of the modelling trade-offs as generated by pragmatic empirical, computational, and cognitive constraints, is a natural fit for climate science. Even if Odenbaugh is right and these pragmatic problems have been largely addressed in ecology, they persist in climate science and it is unlikely that modellers will overcome these limitations any time soon (Schneider et al. 2017, p. 4). Indeed, the incomprehensibility and causal opacity of ESMs may only get worse with technological improvements if climate modellers include yet more components in their state-of-the-art models when the technology permits. In the remainder of this section, I will make one final argument against Matthewson, focusing on his claim that, if a scientific discipline faced a trade-off, it would have its own literature on an equivalent of "The Strategy." As we will see, climate modelling does, in fact, have such a literature.

3.3 Balancing comprehension and comprehensiveness

Jule Charney, an atmospheric modeller from the early post World War II days and beyond, argued that climate modellers, or anyone representing similar complex systems, must "choose either a precise model in order to predict or an extreme simplification in order to understand" (Charney 1963, p. 289; c.f. Dalmedico 2001, p. 415). The central trade-off in climate science, then, is between realistic and precise models and comprehensible ones.[4] Highly realistic models have their strengths: representing as much causal structure as possible will hopefully help avoid the problem of omitting a potential difference-maker or missing some unforeseen feedback loop, and more detailed models can be used to investigate more detailed counterfactuals (how are precipitation patterns in Australasia likely to change?) (Shukla et al. 2010). As Charney suggested, these strengths make them well-suited to the task of prediction.

[4] Levins (1993, p. 554) acknowledged that understandability was another important modelling desiderata beyond the three discussed in his (1966).

Simpler models, on the other hand, are far more comprehensible than complex ones and, as such, are far better suited to fostering understanding of fundamental climate processes. This is not just because they isolate key difference-making processes in contrast to realistic or comprehensive models that try to include as many processes as possible, although they can do this very well. Simpler models are also better suited to increasing understanding within an epistemic community because simpler models can have greater longevity. While realistic models may become obsolete as more powerful computers become available and the state-of-the-art changes, simpler models retain their value precisely because their value is not, for the most part, determined by technological feasibility. The longevity of simple models, Nadir Jeevanjee and colleagues (2017) argue, could foster a greater understanding of fundamental climate processes because a smaller set of simpler models lends itself to thorough investigation by researchers over time. A large set of models hidden away in many different publications, they argue, is less apt to foster such collective comprehension and progress.

Climate science has an old an on-going literature on its own version of the strategy, which I will continue to discuss in Sect. 4.2. Moreover, I will argue that, within this literature, climate scientists have proposed a method that we may use to refine some of Levins' work. Below, I will describe Levins' procedure of robustness analysis and introduce the notion of a model hierarchy, which I will argue can be used to refine and systematise robustness analysis.

4 Robustness analysis and model hierarchies

In his (1966), Levins introduced the massively influential notion of robustness analysis (RA). Levins' RA involves building a family of models with a fixed causal core and a set of varying auxiliary assumptions. If, despite the variety of different assumptions, the models produce the same result, then the relationship between the causal core and the result is robust and modellers can formulate a "robust theorem" describing their relationship (Levins 1993, p. 553); something like *a general biocide* (the causal core) *favours the relative abundance of the prey* (the result) (Weisberg and Reisman 2008). In this section, I examine how RA has been discussed in the philosophy of climate science and argue that these discussions have focused too heavily on RA in the context of highly complex, predictive models, and would benefit from greater consideration of RA among much more idealised models suited to fostering understanding. My further aim in this section is also to show how the notion of a model hierarchy from climate science is also highly beneficial to the general literature on RA as a template of how to conduct RA systematically.

4.1 Robustness analysis in climate science

Robustness analysis appears to be commonplace in climate science. One way in which modellers deal with the uncertainty of ESMs and GCMs is to perform ensemble studies (Lloyd 2010; Odenbaugh 2018; Parker 2006, 2013). There are two kinds of model ensembles. Perturbed physics ensembles are specifically meant to address para-

metric uncertainty—that is, the uncertainty regarding what values parameters should take—and involve running the same model with a range of different parameter settings. Multi-model ensembles, on the other hand, involve using multiple different models from different modelling centres and comparing their behaviour in the same scenarios. Multi-model ensembles are intended to address structural uncertainty—that is, the uncertainty about which processes should be represented and how—by, for example, using different parameterisation schemes for poorly understood but important processes like cloud formation or by using different atmosphere or ocean circulation components (Stocker 2014).

Many philosophers argue that ensemble modelling does not meet the conditions allowing for successful RA despite the similarities between the two practices (Lloyd 2010, 2015; Parker 2006, 2010, 2011, 2013; Winsberg 2018a, b), although some are more optimistic (Lehtinen 2016, 2018). For a start, these ensembles cover a very small region of possible model space, especially considering the large number of adjustable assumptions in highly complex models like ESMs. Varying every structural assumption in an ESM would require ensembles far larger than the collection of fifty or so models seen in the ensembles of the last Intergovernmental Panel on Climate Change assessment report (Stocker 2014). Even within these small ensembles, questionable assumptions are not varied one at a time and one research group's model and another's may have many structural differences.

To make matters worse, the models within these multi-model ensembles are not independent. Naturally, they are built upon the same background knowledge of climate theory and are benchmarked against the same data. But, more problematically, they also have a common history, with some model components, such as the algorithms that approximate the primitive equations, being shared between research groups (Edwards 2010). Modellers also move between research groups and may take their methods with them. To summarise the argument: multi-model ensembles are not prepared systematically or independently but are ensembles of opportunity. That is, they are constructed from whatever models existing research groups contribute, so agreement between them is not good evidence of a robust result.

In response to these criticisms, Fulvio Mazzocchi and Antonella Pasini (Mazzocchi and Pasini 2017; Pasini and Mazzocchi 2015) have argued that RA can be better applied within climate science if modellers also consider alternative kinds of models (see also Katzav and Parker 2015). Mazzocchi and Pasini describe two kinds of data-driven modelling frameworks that could provide independent lines of model-based evidence in climate attribution studies: neural network models (Pasini et al. 2006; Schönwiese et al. 2010; Verdes 2007), and Granger causality analyses (Attanasio et al. 2012, 2013; Pasini et al. 2012). Neural network modelling falls within the space of machine learning. Networks are trained on a training set and validated with an independent data set. Once trained, they can identify nonlinear relationships between causes and effects, such as greenhouse gas concentration and increasing global temperatures that would otherwise go undetected (Mazzocchi and Pasini 2017, p. 5). Unlike neural network analysis, Granger causality analysis is a linear method and, very roughly, involves predicting the value of some variable at t_2 based on the value of some variable at t_1 (Granger 1969). Taking a variable x to represent an external forcing like GHG concentration, and another variable y to represent global average temperatures, we

can ask whether the x and y at t_1 make for a better predictor of y at t_2 than y at t_1 alone. If so, then x is deemed to be a Granger-cause of y. These studies need not be bivariate but can be performed with multi-variate data sets for a more sophisticated picture of the causal relationships between variables.

My own take on RA in the context of climate science is influenced by Levins' general message that modellers should not become overly fixated on the brute force approach of using highly complex predictive simulation models. Instead, they must remember the value of much simpler models. Although Levins did not explicitly advocate for the use of simple models in his (1966) paper, instead advocating for the use of general models, he nevertheless built and investigated simple models himself. Levins was one of four eminent biologists—the others being E. O. Wilson, Richard Lewontin, and Robert MacArthur—who met regularly to discuss a unified theory in population biology (Odenbaugh 2003, 2006). While some biologists favoured the brute force approach of building highly complex simulation models, a practice Levins derided as "FORTRAN ecology" (Levins 1968a), a reference to the programming language, Levins and colleagues believed the complexity of population biology could be represented with multiple simple mathematical models which were mostly qualitative. Wilson called this the *simple theory* approach (Chisholm 1972, p. 177; c.f. Odenbaugh 2006). Simpler models can be useful because they apply to more target systems, as Levins desired (though see Evans et al. 2013), or because they foster understanding, as Charney (1963) suggested.

To present my view, let me begin with two senses of RA that Tarja Knuuttila and Andrea Loettgers (2011) argue can be extracted from Levins' paper. The first sense Knuuttila and Loettgers call *independent determination* RA. This involves increasing researchers' confidence regarding a result by using multiple lines of evidence, such as multiple different models, but which could be generalised to include empirical sources of evidence such as the use of multiple experimental paradigms or different observational and measurement equipment (Wimsatt 1981). As Mazzocchi and Pisini argue, this kind of RA would indeed be strengthened through the use of different model types, such as data-driven models, to add greater independence into model ensembles.

The second notion of RA Knuuttila and Loettgers identify is called causal isolation RA. Instead of focusing on increasing a researcher's confidence in a result, causal isolation RA targets the causal mechanism driving the robust result (Knuuttila and Loettgers 2011, pp. 777–778). Multiple models are used as a means of investigating a possible mechanism, varying parameters or components to assess which combination of factors are sufficient to produce the focal result. This notion of causal isolation RA is similar to other views in the RA literature which similarly do not consider the procedure to be a means of increasing our confidence in the model results (Forber 2010; Odenbaugh and Alexandrova 2011). Patrick Forber (2010), for example, characterises RA in the context of evolutionary theorising as an exercise of how-possibly modelling (p. 37): "The formal inquiry exemplified by robustness analysis and simulation provides global how-possibly explanations that constrain what counts as a biological possibility" (see Gelfert 2016, pp. 87–93 for more on exploratory modelling and how-possibly explanation).

Both the independent determination and causal isolation notions of RA described above can be more generally characterised in terms of Jonah Schupbach's (2016)

notion of *explanatory robustness*, which Winsberg has recently advocated for as the right view of robustness in climate science. The relationship between Schupbach's explanatory robustness and independent determination RA is not too difficult to illustrate. On Schupbach's view, robustness is fundamentally a matter of ruling out competing hypotheses. Independent determination RA is likewise concerned with ruling out competing hypotheses, where these competing hypotheses are those concerning possible artefacts that may introduced by the choice of observation equipment, experimental designs, experimental subjects, models, or whatever else (Wimsatt 1981). To co-opt an example from Ian Hacking's (1983), if I observe that subjects under a microscope appear to be covered in unusual globules, I may use another kind of microscope to check whether the flint-glass achromatic lenses are introducing this visual artefact.

In the last part of this section, I will demonstrate how the so-called model hierarchy of climate science supplies a framework for the systematic variation of assumptions in model families that can be used for both independent determination and causal isolation RA.

4.2 Exploring model hierarchies

Charney described a process of climbing a "hierarchy of models" in which climate models would slowly become increasingly comprehensive until researchers reached the most realistic model at the top of the hierarchy (Edwards 2010). Modellers were to get a better understanding of climate processes as they built and climbed this hierarchy. Unfortunately, Charney's hierarchies were somewhat forgotten for several decades while modellers pursued the alternative brute force approach of building ensembles of highly complex models—a trend facilitated by the rapid development of computational hardware (Parker 2014). As modellers have encountered the limitations of this approach, however, there has been some renewed interest in model hierarchies:

> The models used to simulate the climate are themselves complex, chaotic dynamical systems. To work with them effectively requires not only the careful examination of alternative formulations of these comprehensive models but also the construction of a hierarchy of models in which elements of complexity are added sequentially. (Held 2014, p. 1206)

There have been a few different ways of conceptualising the dimensions of the model hierarchy (e.g. Held 2005; Jeevanjee et al. 2017, p. 1762; McGuffie and Henderson-Sellers 2013, p. 52). In recent papers, Sandrine Bony et al. (2013), Jeevanjee et al. (2017), and Penelope Maher et al. (2019) have all discussed model hierarchies in climate science, with Maher et al. focusing specifically on atmospheric models. Bony et al.'s representation is the simplest, as they picture a two-dimensional space, with target system complexity along the x axis, running from basic particle and fluid systems at one end of this spectrum to entire Earth and Earth-Human systems at the other, and model simplicity *relative to the target system* along the y axis. An energy balance model, which can be described with only a few parameters and can be manipulated with back of the envelope reasoning, for example, is far simpler than an ESM relative to their (nearly) shared target: Earth. Much like Charney, Bony et al. see a trade-off

Table 1 Jeevanjee et al. propose six dimensions along which climate models may be more or less realistic (Adapted from 2017, p. 1761)

Fluid	Rotation	Ocean	Surface	Convention	Radiation
Compressible	Coriolis	Dynamical	Land + ice	Explicit moist	Spectral
Hydropstatic	β-plane	Column	Real land	Super-param.	Gray
QG	f-plane	Slab	Ideal land	Parameterised	Newtonian
static	none	Non-uniform T_S	Aqua	Large-scale	Fixed
		Uniform T_S		Dry	

between understanding and realism, where simpler models with simpler targets are more understandable, and where more complex models with complex targets are more realistic.[5]

Jeevanjee et al. claim that the notion of a hierarchy can be misleading since there is not a strict ordering of less to more complex models (Jeevanjee et al. 2017, p. 1761): "how, for instance, can one compare a moist, non-rotating cloud-resolving simulation in a planar geometry to a dry, rotating, coarse-resolution global simulation? One is not clearly more realistic than the other, at least in any general sense." They, and Maher et al., who follow Jeevanjee et al.'s lead, conceptualise of the hierarchy as a multi-dimensional model space, with different dimensions along which climate models or, in Maher et al.'s case, atmospheric models, can vary in complexity. The six dimensions Jeevanjee et al. describe are shown in Table 1. With respect to representations of the Earth's surface, for example, a model that represents real land masses and sea ice is more realistic than a model that represents the Earth as an aquaplanet, completely covered with oceans. While it is obvious that a model maximally complex along all dimensions (contained in the accompanying table) is more complex than a model maximally simple along these dimensions, it is not obvious, for Jeevanjee et al., that a model maximally complex along one dimension and maximally simple across all the rest is more complex than a model maximally complex along some other dimension and maximally simple along the others. Likewise, if one model is maximally complex along three of Jeevanjee et al.'s six dimensions and maximally simple along the other three and another model is moderately complex along all six, it is not obvious which is more complex than the other. In the absence of a strict, global ordering, a framework like that depicted in Table 1 allows researchers to compare models of varying complexity by decomposing model complexity into different dimensions, which researchers can cite when comparing models.

A hierarchy like the one shown in Table 1 can be used for both kinds of RA. Jeevanjee et al. (2017, p. 1764) demonstrate that hypothesis testing is a function

[5] In Bony et al.'s two-dimensional model space, there are three regions that remain unoccupied. First, there are few if any complex models of simple systems because they are of limited scientific value. Second, in a region Bony et al. label a "conceptual abyss," there are few simple models of highly complex systems due to a lack of understanding. Finally, in a region they label a "computational abyss," there are few very complex models (relative to the target system) of very complex target systems due to computational limitations. Bony et al. place a "descriptive horizon" between the second and third unoccupied regions and the rest of the model space.

of the model hierarchy and give two examples. First, polar amplification is a well-known climate phenomenon. The effects of increased atmospheric CO_2 levels, such as increased average surface temperatures, are more exaggerated toward the poles than they are at lower latitudes. One possible explanation of polar amplification is an albedo feedback process initiated by highly reflective polar ice, which decreases as rising surface temperatures cause the ice to melt. Alexeev (2003) shows how such a hypothesis can be tested by descending down the model hierarchy, building an aquaplanet model without any land or ice, in which the surface has a uniform albedo rather than concentrations at the poles. By reproducing polar amplification without the typically assumed cause, Alexeev created space for hypotheses about alternative causal mechanisms, which can be articulated and tested through further modelling (see Alexeev et al. 2005).

The second example involves the phenomenon Atlantic multidecadal oscillation (AMO), a cycle of variation in sea surface temperatures of the Northern Atlantic Ocean with a period of about 70 years. Here too there is uncertainty regarding the causal mechanism responsible for AMO, but one popular contender is Atlantic meridional overturning circulation, which is the chunk of ocean circulation that takes place in the North Atlantic. Amy Clement and colleagues (2015) tested the explanatory robustness of Atlantic meridional overturning circulation as the driver of AMO. In their study, GCMs with coupled atmospheric and oceanic circulation components were uncoupled and the circulating atmospheric component was coupled to a non-circulating ocean component. This represents a movement through model space along Jeevanjee et al.'s ocean dimension, from a dynamical to a slab ocean. Clement et al. found the AMO was stable across the two conditions and so could not be explained primarily by ocean circulation, even if heat transfer between atmosphere and ocean is still part of the explanation. These are just two examples, but they illustrate how movement through the model space can facilitate a better understanding of key climate processes through the systematic hypothesis testing suggested by Schupbach's notion of explanatory robustness.

5 Conclusion

Levins' "The Strategy" is an important work for the philosophy of computational modelling that remains pertinent today. Written at a time when simulation science was just breaking into population biology, Levins' articulation of the epistemic challenges faced by the brute force approach remains accurate in climate science where the strategy flourishes. As Levins described, highly complex computational models have big appetites for data that can sometimes be unobtainable, they are intractable in their ideal forms and must be adjusted in various ways to run on available computers, and their complexity makes them difficult to understand. These problems create trade-offs. In climate science, the most important trade-offs are between the realism or comprehensiveness of complex models at the top of the model hierarchy and the comprehensibility of simpler models forming the hierarchy's lower rungs or foundations. I wish to draw two conclusions from my discussion of the causal isolation reading of Levins' RA and model hierarchies. First, the philosophy of climate science literature

on RA may benefit from looking at simpler climate models with this notion of RA in mind, which has, until now, only been discussed in the philosophies of biology and economics. Second, the literature on RA in the philosophy of science more generally would benefit from considering the use of model hierarchies in their approaches to exploratory and how-possibly modelling through causal isolation RA.

Acknowledgements I would like to John Matthewson and Kim Sterelny for their many helpful comments on this material as well as two anonymous reviewers whose advice led to a much-improved paper. I would also like to thank audiences at the AAP/NZAP2018 in Wellington and PSA2018 in Seattle, and particularly Joel Katzav and Michael Weisberg, for their feedback on earlier versions of this paper presented at those events. Funding was provided by Australian Research Council (Grant No. ARC FL13).

References

Alexeev, V. A. (2003). Sensitivity to CO2 doubling of an atmospheric GCM coupled to an oceanic mixed layer: A linear analysis. *Climate Dynamics, 20*(7–8), 775–787.

Alexeev, V. A., Langen, P. L., & Bates, J. R. (2005). Polar amplification of surface warming on an aquaplanet in "ghost forcing" experiments without sea ice feedbacks. *Climate Dynamics, 24*(7–8), 655–666.

Attanasio, A., Pasini, A., & Triacca, U. (2012). A contribution to attribution of recent global warming by out-of-sample Granger causality analysis. *Atmospheric Science Letters, 13*(1), 67–72.

Attanasio, A., Pasini, A., & Triacca, U. (2013). Granger causality analyses for climatic attribution. *Atmospheric and Climate Sciences, 3*(04), 515.

Bjerknes, V. (1904). Das Problem der Wettervorhers-age, betrachtet vom Standpunkte der Mechanik und der Physik. *Meteorologische Zeitschrift, 21,* 1–7.

Bony, S., Stevens, B., Held, I. H., Mitchell, J. F., Dufresne, J.-L., Emanuel, K. A., et al. (2013). Carbon dioxide and climate: Perspectives on a scientific assessment. *Climate Science for Serving Society* (pp. 391–413). New York: Springer.

Charney, J. G. (1963). Numerical experiments in atmospheric hydrodynamics. In *Experimental Arithmetic, High Speed Computing and Mathematics.* Proceedings of Symposia in Applied Mathematics (Vol. 15, pp. 289–310).

Chisholm, A. (1972). Philosophers of earth: Conservations with ecologists. *Philosophers of earth: Conservations with ecologists*. New York: E. P. Dutton.

Clement, A., Bellomo, K., Murphy, L. N., Cane, M. A., Mauritsen, T., Rädel, G., et al. (2015). The Atlantic Multidecadal Oscillation without a role for ocean circulation. *Science, 350*(6258), 320–324.

Dalmedico, A. D. (2001). History and epistemology of models: Meteorology (1946–1963) as a case study. *Archive for History of Exact Sciences, 55*(5), 395–422.

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

Elliott-Graves, A. (2018). Generality and causal interdependence in ecology. *Philosophy of Science, 85*(5), 1102–1114.

Evans, M. R., Grimm, V., Johst, K., Knuuttila, T., De Langhe, R., Lessells, C. M., et al. (2013). Do simple models lead to generality in ecology? *Trends in Ecology & Evolution, 28*(10), 578–583.

Forber, P. (2010). Confirmation and explaining how possible. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 41*(1), 32–40.

Gelfert, A. (2016). *How to do science with models: A philosophical primer*. New York: Springer.

Giere, R. N. (1988). *Explaining science: A cognitive approach*. Chicago: University of Chicago Press.

Giere, R. N. (2010). *Scientific perspectivism*. Chicago: University of Chicago Press.

Godfrey-Smith, P. (2006). The strategy of model-based science. *Biology and Philosophy, 21*(5), 725–740.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society, 37*(3), 424–438.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge: Cambridge University Press.

Held, I. (2005). The gap between simulation and understanding in climate modeling. *Bulletin of the American Meteorological Society, 86*(11), 1609–1614.

Held, I. (2014). Simplicity amid complexity. *Science, 343*(6176), 1206–1207.
Hughes, R. I. G. (1999). The Ising model, computer simulation, and universal physics. *Ideas In Context, 52,* 97–145.
Humphreys, P. (2004). *Extending ourselves: Computational science, empiricism, and scientific method*. Oxford: Oxford University Press.
Jeevanjee, N., Hassanzadeh, P., Hill, S., & Sheshadri, A. (2017). A perspective on climate model hierarchies. *Journal of Advances in Modeling Earth Systems, 9*(4), 1760–1771.
Katzav, J., & Parker, W. S. (2015). The future of climate modeling. *Climatic Change, 132*(4), 475–487.
Klein, C. (2010). Philosophical issues in neuroimaging. *Philosophy Compass, 5*(2), 186–198.
Knuuttila, T., & Loettgers, A. (2011). Causal isolation robustness analysis: The combinatorial strategy of circadian clock research. *Biology and Philosophy, 26*(5), 773–791.
Lehtinen, A. (2016). Allocating confirmation with derivational robustness. *Philosophical Studies, 173*(9), 2487–2509.
Lehtinen, A. (2018). Derivational robustness and indirect confirmation. *Erkenntnis, 83*(3), 539–576.
Lenhard, J., & Winsberg, E. (2010). Holism, entrenchment, and the future of climate model pluralism. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics, 41*(3), 253–262.
Levins, R. (1966). The strategy of model building in population biology. *American Naturalist, 54*(4), 421–431. https://doi.org/10.2307/27836590.
Levins, R. (1968a). *Ecological engineering: Theory and technology*. Stony Brook: Stony Brook Foundation, Inc.
Levins, R. (1968b). *Evolution in changing environments: Some theoretical explorations*. Princeton: Princeton University Press.
Levins, R. (1973). The limits of complexity. In H. H. Pattee (Ed.), *Hierarchy theory—The challenge of complex systems* (pp. 109–127). New York: George Braziller.
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.
Lewis, S. C. (2017). *A changing climate for science*. Springer.
Lloyd, E. A. (2010). Confirmation and robustness of climate models. *Philosophy of Science, 77*(5), 971–984.
Lloyd, E. A. (2015). Model robustness as a confirmatory virtue: The case of climate science. *Studies in History and Philosophy of Science Part A, 49,* 58–68.
Maher, P., Gerber, E. P., Medeiros, B., Merlis, T. M., Sherwood, S., Sheshadri, A., et al. (2019). Model hierarchies for understanding atmospheric circulation. *Reviews of Geophysics, 57,* 250–280.
Matthewson, J. (2011). Trade-offs in model-building: A more target-oriented approach. *Studies in History and Philosophy of Science Part A, 42*(2), 324–333.
Matthewson, J., & Weisberg, M. (2009). The structure of tradeoffs in model building. *Synthese, 170*(1), 169–190.
Mazzocchi, F., & Pasini, A. (2017). Climate model pluralism beyond dynamical ensembles. *Wiley Interdisciplinary Reviews Climate Change, 8*(6), e477.
McGuffie, K., & Henderson-Sellers, A. (2013). *a climate modelling primer*. New York: Wiley.
Meijer, A., & Bolívar, M. P. R. (2016). Governing the smart city: A review of the literature on smart urban governance. *International Review of Administrative Sciences, 82*(2), 392–408.
Neelin, J. D. (2010). *Climate change and climate modeling*. Cambridge: Cambridge University Press.
Odenbaugh, J. (2003). Complex systems, trade-offs, and theoretical population biology: Richard Levin's "strategy of model building in population biology" revisited. *Philosophy of Science, 70*(5), 1496–1507.
Odenbaugh, J. (2006). The strategy of "The strategy of model building in population biology". *Biology and Philosophy, 21*(5), 607–621.
Odenbaugh, J. (2018). Building trust, removing doubt? Robustness analysis and climate modeling. *Climate Modelling* (pp. 297–321). New York: Springer.
Odenbaugh, J., & Alexandrova, A. (2011). Buyer beware: Robustness analyses in economics and biology. *Biology and Philosophy, 26*(5), 757–771. https://doi.org/10.1007/s10539-011-9278-y.
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. https://doi.org/10.2307/3037250.
Parker, W. S. (2006). Understanding pluralism in climate modeling. *Foundations of Science, 11*(4), 349–368. https://doi.org/10.1007/s10699-005-3196-x.

Parker, W. S. (2010). Predicting weather and climate: Uncertainty, ensembles and probability. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics, 41*(3), 263–272.
Parker, W. S. (2011). When climate models agree: The significance of robust model predictions. *Philosophy of Science, 78*(4), 579–600.
Parker, W. S. (2013). Ensemble modeling, uncertainty and robust predictions. *Wiley Interdisciplinary Reviews: Climate Change, 4*(3), 213–223. https://doi.org/10.1002/wcc.220.
Parker, W. S. (2014). Simulation and understanding in the study of weather and climate. *Perspectives on Science, 22*(3), 336–356.
Parker, W. S. (2016). Reanalyses and observations: What's the difference? *Bulletin of the American Meteorological Society, 97*(9), 1565–1572.
Pasini, A. (2005). *From observations to simulations: A conceptual introduction to weather and climate modelling*. Singapore: World Scientific.
Pasini, A., Lorè, M., & Ameli, F. (2006). Neural network modelling for the analysis of forcings/temperatures relationships at different scales in the climate system. *Ecological Modelling, 191*(1), 58–67.
Pasini, A., & Mazzocchi, F. (2015). A multi-approach strategy in climate attribution studies: Is it possible to apply a robustness framework? *Environmental Science & Policy, 50,* 191–199.
Pasini, A., Triacca, U., & Attanasio, A. (2012). Evidence of recent causal decoupling between solar radiation and global temperature. *Environmental Research Letters, 7*(3), 34020.
Rohrlich, F. (1990). Computer simulation in the physical sciences. In *PSA: Proceedings of the biennial meeting of the philosophy of science association* (Vol. 1990, pp. 507–518). Philosophy of Science Association.
Schneider, S. H., & Dickinson, R. E. (1974). Climate modeling. *Reviews of Geophysics, 12*(3), 447–493.
Schneider, T., Teixeira, J., Bretherton, C. S., Brient, F., Pressel, K. G., Schär, C., et al. (2017). Climate goals and computing the future of clouds. *Nature Climate Change, 7*(1), 3.
Schönwiese, C.-D., Walter, A., & Brinckmann, S. (2010). Statistical assessments of anthropogenic and natural global climate forcing. An update. *Meteorologische Zeitschrift, 19*(1), 3–10.
Schupbach, J. N. (2016). Robustness analysis as explanatory reasoning. *The British Journal for the Philosophy of Science, 69*(1), 275–300.
Shukla, J., Palmer, T. N., Hagedorn, R., Hoskins, B., Kinter, J., Marotzke, J., et al. (2010). Toward a new generation of world climate research and computing facilities. *Bulletin of the American Meteorological Society, 91*(10), 1407–1412.
Stocker, T. (2014). *Climate change 2013: The physical science basis: Working Group I contribution to the Fifth assessment report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.
Trout, J. D. (2016). *Wondrous truths: The improbable triumph of modern science*. Oxford: Oxford University Press.
Verdes, P. F. (2007). Global warming is driven by anthropogenic emissions: A time series analysis approach. *Physical Review Letters, 99*(4), 48501.
Washington, W. M., Buja, L., & Craig, A. (2009). The computational future for climate and Earth system models: On the path to petaflop and beyond. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 367*(1890), 833–846.
Washington, W. M., & Parkinson, C. (2005). *Introduction to three-dimensional climate modeling*. Sausalito: University Science Books.
Watt, K. E. F. (1956). The choice and solution of mathematical models for predicting and maximizing the yield of a fishery. *Journal of the Fisheries Board of Canada, 13*(5), 613–645.
Watt, K. E. F. (1962). Use of mathematics in population ecology. *Annual Review of Entomology, 7*(1), 243–260.
Watt, K. E. F., & Watt, K. E. F. (1968). *Ecology and resource management: A quantitative approach*. New York: McGraw-Hill.
Weisberg, M. (2004). Qualitative theory and chemical explanation. *Philosophy of Science, 71*(5), 1071–1081.
Weisberg, M. (2006a). Forty Years of "The Strategy": Levins on Model Building and Idealization. *Biology and Philosophy, 21*(5), 623–645.
Weisberg, M. (2006b). Robustness analysis. *Philosophy of Science, 73*(5), 730–742. https://doi.org/10.1007/s001900100162.

Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press. Retrieved from http://books.google.com/books?id=rDu5e532mIoC&pgis=1.
Weisberg, M., & Reisman, K. (2008). The robust volterra principle. *Philosophy of Science, 75*(1), 106–131. https://doi.org/10.1086/588395.
Wimsatt, W. C. (1981). Robustness, reliability, and overdetermination. In M. Brewer & B. Collins (Eds.), *Scientific inquiry and the social sciences* (pp. 124–163). San Francisco: Jossey-Bass.
Winsberg, E. (2010). *Science in the age of computer simulation*. Chicago: University of Chicago Press.
Winsberg, E. (2018a). *Philosophy and climate science*. Cambridge: Cambridge University Press.
Winsberg, E. (2018b). What does robustness teach us in climate science: A re-appraisal. *Synthese*, 1–24.