5-10-2014

When Simulations Conflict: Problems with the External Validation of Computer Simulations

Archie Fields III
Georgia State University

WHEN SIMULATIONS CONFLICT: PROBLEMS WITH THE EXTERNAL VALIDATION OF COMPUTER SIMULATIONS

by

ARCHIE FIELDS III

Under the direction of Daniel Weiskopf and Andrea Scarantino

ABSTRACT

I show that Eric Winsberg's principles of model-building given in *Science in the Age of Computer Simulation* are insufficient to argue for the external validation of simulation data in cases in which simulation results conflict, and that laboratory experiments have an advantage over simulations because conflicting experimental results can be decided between on the basis of reproducibility. I also argue that robustness of predictions serves the same function for simulations as repeatability does for laboratory experiments in either adjudicating between conflicting results or allowing us to say that we do not have sufficient justification to validate the results. Finally, I argue for an interpretation of the argument from robustness that appeals to the convergence of many well-built and diverse models rather than the more common interpretation which appeals to the probability that one of a set of models is likely to be true.

INDEX WORDS: Simulation, Validation, Robustness, Experiment, Model, Data, Philosophy

WHEN SIMULATIONS CONFLICT: PROBLEMS WITH THE EXTERNAL VALIDATION OF COMPUTER SIMULATIONS

by

ARCHIE FIELDS III

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Arts

in the College of Arts and Sciences

Georgia State University

2014

WHEN SIMULATIONS CONFLICT: PROBLEMS WITH THE EXTERNAL VALIDATION OF COMPUTER SIMULATIONS

by

ARCHIE FIELDS III

Committee Chairs: Andrea Scarantino

Daniel Weiskopf

Committee: George Graham

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

May 2014

Acknowledgements

I would like to thank Dan Weiskopf and Andrea Scarantino for their extremely helpful comments on the many drafts of this thesis, as well as George Graham for his help in the later stages of it. I am also grateful to all of my fellow graduate students in the philosophy department, who supported me and provided ample opportunities for discussion of some of the ideas and arguments contained in this thesis. Thanks also to my undergraduate professors, especially Bill Wilkerson and Nick Jones, who helped me see the value of philosophy and encouraged my interest in the subject. Finally, I would like to thank my mother and father, who fostered within me a love of learning at a young age.

Table of Contents

List of Figures

Chapter 1. Introduction

For hundreds of years, progress in science has been limited to using laboratory experiments and field observations as evidence to support various hypotheses. With the advent of computer simulations, scientists have gained a powerful new tool for producing data and studying phenomena. Computer simulations represent a natural system (called the "target system") with a digital model of that system, called the simulation model. The simulation model is constructed to try to find solutions to the differential equations from our best scientific theories which describe the behavior of the target system. Computer simulations are becoming more and more widespread throughout the sciences, from astrophysics to climatology. Indeed, computer simulations are being used by many climatologists to predict climate change on a global scale. Data from these simulations often are used to build other models that can make further predictions. One example I will focus on later in this paper is a model of what areas in North America will be favorable, given the data from climate simulations, for populations of a particular species of insect, the pea leafminer. Such a model, in turn, could be used to predict how populations of that insect will migrate across North American in future decades.

The question arises, though, as to whether computer simulations are (or can be) just as epistemically powerful as more traditional laboratory experiments. By "epistemic power," I mean the capability or potential of some method of investigation to yield accurate and trustworthy information about the subject of the investigation. Many philosophers have recently given arguments for why computer simulations can be considered just as epistemically powerful as material experiments when certain conditions are met. So, simulations meeting these conditions, whatever those conditions turn out to be, can be considered just as trustworthy and informative about the systems they are used to investigate as material experiments. Eric

Winsberg, in *Science in the Age of Computer Simulation*, argues that whether or not we accept as trustworthy a certain set of experimental or simulation data rests on the arguments that can be made for the validation the experiment or simulation. Validation, in the context of computer simulations, "is the process of determining whether or not the chosen model is a good enough representation of the real-world system for the purposes of the simulation" (Winsberg p. 20). Likewise, validating an experiment involves determining whether or not the material system being investigated in a laboratory is a good enough representation of similar natural systems outside of the laboratory such that we can generalize the conclusions we draw about the behavior of the material system in the laboratory to the similar natural systems out in the world. According to Winsberg, validating an experiment or simulation requires an argument for whether we are justified in accepting its results as accurate. Because these arguments for validation can be weaker or stronger for either simulations or material experiments, there is no inherent difference with respect to epistemic power between the two types of investigation. Moreover, Winsberg gives a set of criteria, called principles of model-building, which are supposed to serve as the guidelines for arguing for the validity of a simulation and its results. These criteria are supposed to serve to validate simulation results even if (or perhaps especially when) the relevant data from the system being simulated is impossible to get through normal empirical methods because the system of interest is too unstable, too distant or even too small. Indeed, part of what motivates the use of computer simulations in the sciences is that they provide us with a means to study natural phenomena, or at least very detailed models of those phenomena, which cannot be produced and studied in a traditional laboratory, such as black holes and violent storms.

I will argue that Winsberg's criteria for validating simulations, when certain data from the system being simulated are lacking, will fail to allow us to distinguish between the outputs of

simulations that meet these criteria but give different results. To do this, I present two climate simulation models for which we seem to be able to give strong arguments for their validation. In other words, according to Winsberg's principles of model-building, we should be able to consider these two simulations as trustworthy sources of data. I then present a study which shows that they actually give drastically conflicting predictions in certain cases, so that Winsberg's principles of model-building end up validating contradictory results. Moreover, I will present one criterion for validation which favors laboratory experiments over simulations by helping us distinguish between conflicting results in laboratory experiments but not in computer simulations. This criterion is reproducibility of results, and the fact that it helps validate laboratory experiments but not simulations[1] suggests that we can be more justified in trusting the results of well-conducted laboratory experiments than we can be in trusting the results of well-built simulations. However, following Wendy Parker in "When Climate Models Agree: the Significance of Robust Model Predictions," I will suggest an addition to Winsberg's criteria for accepting simulation results which serves to put computer simulations back on a methodological par with laboratory experiments: robustness of results. Robustness of results, in the context of computer simulations, is obtaining similar results from a variety of methodologically independent[2] computer simulations of the same natural phenomenon. In fact, robustness of predictions may serve a similar validation function for simulations as reproducibility of results serves for laboratory experiments. So ultimately, I wish to save Winsberg's position for equal epistemic power for laboratory experiments and computer simulations by adding another

1 This is not precisely true, since reproducibility of simulation results does enhance the internal validity of a computer simulation, that is, our trust that the simulation is functioning as we intend for it to function. However, the type of validation Winsberg chiefly concerns himself with is external validation of simulations, which, as described above, deals with whether or not we should accept the results of a study as accurate.

2 By "methodologically independent," I mean that each of the simulations differ in the modeling techniques they employ to represent the target system. A more detailed discussion of robustness is found in Chapter 5.

criterion, robustness of results, to Winsberg's list of criteria for the validation of simulation results.

Naturally, this work may have broader implications in the philosophy of science, particularly in regard to the nature of scientific evidence and the relationship between theory and evidence. Computer simulations, if a modified version of Winsberg's account is right, allow us to gather valid data about a system without having much in the way of physical access to that system. Moreover, since theories are used to help construct simulations, theory may play an even greater role in data collection than previously thought. However, I will be in a better position to explain the implications of this paper for philosophy of science more broadly after explaining Winsberg's argument for the epistemic equivalence of computer simulations compared to more traditional means of data collection. So, I will return to the implications of this work for both the nature of scientific evidence and the relationship between theory and evidence in the conclusion of the paper.

Chapter 2. Winsberg's Model-Building Principles

Eric Winsberg repudiates the claim that experiments are "epistemically privileged relative to simulations—the claim that they [experiments] 'have greater potential to make strong inferences back to the world'" (Winsberg, p. 70). Winsberg argues for his position by suggesting that "how trustworthy or reliable an experiment or simulation is depends on the *quality* of the background knowledge and the skill with which it is put to use" in designing and running the experiment or simulation (ibid. p. 70). By "trustworthy" and "reliable," Winsberg means that the simulation or experiment in question is validated. Validation here refers to external validity, that is, whether the simulation or experiment is appropriately representative of the real-world system we want to investigate. External validity, Winsberg claims, is determined by "*the character of the argument given* for the legitimacy of the inference from object [under direct investigation] to target [the real-world system we want to know more about] and the *character of the background knowledge* that grounds that argument" (ibid. p. 63, emphasis in original). In the case of material experiments, such arguments require appeals to properly calibrated measuring instruments, providing appropriate control cases to rule out alternative explanations, the amount of knowledge we already have about how the system being investigated works, etc. Essentially, appealing to these sorts of criteria amounts to arguing that the experiment has been conducted using sound methodology by the standards of the scientific community and that the system studied in the experiment is likely to be appropriately representative of the system we want to learn more about.

To better understand what Winsberg has in mind regarding arguments for validation of results, consider a case in which a researcher wants to know what areas of the cerebral cortex a particular thalamic nucleus projects to in macaque monkeys. The researcher might inject an

anterograde tracer[3] into the thalamic nucleus in question in the subject and then, after euthanizing the subject, examine its brain to see which parts of the cortex the tracer reached. Using tracer chemicals is a tried-and-true technique for determining neural connectivity, so the use of that technique helps the researcher argue that her results about are valid. Of course, not all brains are equivalent, so the researcher will likely want to perform the study on several animals, rather than just one. The use of multiple subjects helps establish the conclusion that whatever results are obtained for the test subjects will hold for the general population of subjects. Appeals to the use of particular techniques, such as anterograde tracer studies of neural connectivity, and general experimental principles, such as using an appropriate number of test subjects, are examples of premises used in the arguments that researchers make to establish the validity of their studies.

Likewise, Winsberg suggests that if computer simulations also have been conducted with sound methodology and we can argue that the system being studied (a computer program and computer) is appropriately representative of the target system, their results can likewise be considered externally validated. Of course, since in both the case of a computer simulation and a material experiment an argument is required to validate their results, there is no inherent epistemic advantage enjoyed by the results of material experiments or computer simulations. On Winsberg's account, what determines the epistemic power of a study, that is, the ability of the study to produce results that we can consider to be accurate, is the strength of the argument for that study's external validity. For computer simulations, external validation of results comes from three factors, each a kind of background knowledge, which Winsberg calls "principles of model-building":

3 An anterograde tracer is a chemical compound that, when injected into neural tissue, will travel away from the injection site along the axonal pathways of the affected neural cell bodies towards wherever those axons terminate (Bear et al., p. 41-42).

(1) Soundness of theoretical principles that guide building the simulation.

(2) Soundness of researcher intuitions regarding the system being studied.

(3) Soundness of computational techniques employed by the simulation.

(Winsberg p. 64-65)

It is through appealing to principles such as these that the results of computer simulations are validated to the point that they are deemed trustworthy empirical data (Winsberg p. 65).

The first principle essentially appeals to our confidence in our theoretical understanding of the behavior of the system we are trying to simulate. So, the fact that we have great confidence that Newtonian mechanics describes the motion and behavior of objects larger than the quantum scale moving at speeds much slower than c[4] adds to our confidence in the results of any simulation employing Newtonian mechanics. The second principle appeals to the idea that scientists trained in the theories and practical techniques of modern science will have, in general, good intuitions about how to model physical systems so as to make the model both computationally tractable and a good representation of the system being studied. In other words, the rigorous training of scientists will help them determine what sort of idealizations and approximations are appropriate to make when building a model to ensure that it appropriately represents the system in the ways needed for any given study. This training and the experimental know-how it bestows upon researchers are particularly important because there is no simple step-by-step algorithm for building a model of any given target system (Winsberg p. 30-31).

An example of a specific technique meant for validating the results of simulations that relies upon the training and intuitions of researchers is calibration. Calibration is the process of showing "that the relevant output of the simulation matches what is known about the

4 The speed of light in a vacuum, approximately 3.00×10^8 meters per second (Halliday et al., p. A-3).

phenomena" (Winsberg p. 22). Typically, this is done by checking to see whether simulations can reproduce experimental data. However, the comparison between simulation results and experimental data is not as straightforward as one might think. For example, "simulation data and experimental data are not always obtained from the same spatial location within a system," so proper calibration "requires the skilled judgment of a good observer...there is no [particular] metric of similarity between the different data sets that need to be compared" (Winsberg p. 22). In other words, calibration requires a certain degree of intuition on the part of the researcher to determine whether or not a given simulation sufficiently reproduces past data to warrant increased confidence in its results.

Regarding the third and final principle of model-building, when we have confidence in the computational techniques and tricks used by scientists either to implement a model on a computer or to simplify calculations, this also lends confidence to the results of the simulation. For example, a frequently used computational technique in fluid dynamics is the inclusion of a term called artificial viscosity which helps account for "certain crucial effects that would otherwise be lost [in the simulation]...in particular, the dissipation of kinetic energy into heat" (Winsberg p. 128). As its name suggests, "artificial viscosity" is not real viscosity that exists in the system being simulated—it is a fiction, a fabrication on the part of the researcher. But employing this fictional term is a computational technique with a record of past successes in making accurate predictions. So, when we use a technique like the inclusion of artificial viscosity to help model and study systems we have not simulated before, those past successes contribute to our confidence in the simulation results regarding that system. Indeed, the past successes of our theoretical principles and of the good intuitions and problem-solving skills of scientists contributes to our trust in those principles as well.

However, it is worth pointing out that validating a simulation according to Winsberg is not the same as "confirming" it in the sense that what grounds the argument for validation does not need to be data from the system being simulated. In other words, when we think of a theory being confirmed, or made more likely to be true, it is confirmed by some piece of observational evidence from systems in the domain of that theory. For example, Kepler's laws of planetary motion were confirmed by observations of the observed paths of planets through the night sky. For simulations, though, we sometimes do not have access to the empirical data that would be needed to confirm or disconfirm the simulation model, such as when we want to create models of systems that are difficult to observe or interact with, like the inside of stars and turbulent storms. Again, part of what is exciting about simulation studies is that they might allow us to learn more about such normally inaccessible systems. But that means that we have to rely upon arguments that our simulations effectively represent the system we are interested in learning more about. Indeed, applying a term like "confirmation" to simulations is somewhat misleading, since confirmation involves *using* evidence to support belief in the truth of some theory or other representational entity, whereas the purpose of validation is figuring out whether an experiment or simulation's results can *count* as evidence. Of course, when building simulations, one wants to use theoretical principles which are well-confirmed by observational evidence and appealing to the use of well-confirmed theoretical principles can be used in the argument for validating the simulation. However, in most cases, simulations also use outright fictions or falsehoods, such as the inclusion of "artificial viscosity" mentioned above, in modeling their target systems. Such fictions are not "confirmed" in the sense that they are thought to represent veridically some part of a real-world system. Nonetheless, using fictions like artificial viscosity help validate simulations in which they are employed by having a record of past success in generating accurate

results or data points regarding the behavior of some real-world systems. So, to summarize, confirming[5] a simulation means arguing that the simulation model of the target system is a veridical representation of the target system while validating a simulation means arguing that the results of the simulation are accurate enough to count as observational evidence or legitimate data from a target system for a given purpose. It is important to realize that Winsberg is concerned more with the validation of simulations and determining whether they produce accurate results and concerned less with determining how veridical the simulation is in representing the actual processes at work within the target system that produce certain outcomes.

5 Again, it is not clear that any simulation is ever truly "confirmed" in this sense since simulations always employ approximations, idealizations and falsehoods in representing their target systems. Of course, some have suggested that simulations provide generally veridical or approximately true representations of their target systems, and as such could be said to be "confirmed" models of certain systems or phenomena. For a more detailed discussion, see Paul Teller's "Fictions, Fictionalization and Truth in Science."

Chapter 3. Conflicting Results from Well-Built Models

3.1 Conflicting Results in Climate Modeling

One problem regarding Winsberg's argument for equal epistemic power for computer simulations is that appealing to model-building principles may not be enough to decide between conflicting simulation results when both simulations are supported by reliable model-building principles. Suppose two different simulations are made of the Earth's atmosphere for the same purpose of predicting global climate change via change in mean temperature. Suppose also that the results of these simulations drastically conflict in some of their predictions, but both are deemed to be built with reliable model-building principles to the extent that we can make equally convincing arguments for the external validity of either simulation. In such a case, it seems that Winsberg's principles of model-building allow us to make an argument that we could trust as being accurate the conflicting results of both simulations, which would be nonsense. Indeed, it seems that just appealing to reliable model-building principles will not be enough to decide in favor of one set of results or the other because both simulations have been built with what the researchers believe to be good model-building principles. The problem is especially troubling in situations in which we are using simulations to make predictions about or measurements of properties of systems that would be difficult or practically impossible to directly measure, such as convection currents inside stars or turbulent wind flows in storms. In such situations, we may not be able to appeal to further empirical evidence to adjudicate between the two because the source of the empirical evidence, that is, the inside of a storm or a star, is not easily observable.

Cases in which simulations are conducted because the systems we want to study cannot be investigated via direct observation or experimentation, are not all that uncommon. Indeed, part of the point of Winsberg's arguments is to provide us with a reason for trusting the output of

simulations built with reliable principles when it is practically impossible to conduct a more traditional laboratory experiment with a system because it is too large, too complex or too far away. Again, though, appeals to model-building principles may not be enough to justify having the same degree of confidence in simulation studies of this type as we have in well-conducted laboratory research in other domains.

Consider studies involving climate models, such as the study conducted in 2010 by Anna M. Mika and Jonathan A. Newman in which they attempt to project what geographical regions in North America will be most environmentally suitable for a certain species of insect called a pea leafminer, or *Lyriomyza huidobrensis* by creating a bioclimatic envelope model for it. A bioclimatic envelope model is a kind of species distribution model "in which the current geographical distribution of species is related to climatic variables so to enable projections of distributions under future climate change scenarios" (Heikkinen et al. p. 751).The motivation for this study is that *L. huidobrensis* is considered to be "an invasive species in North American and a serious economic pest on a wide variety of crops," so being able to predict what regions such pests may be able to expand to in the future would be helpful in agricultural planning (Mika and Newman p. 213). Mika and Newman use two different climate models in trying to make their prediction. Both are the same sort of climate model, called Atmosphere-Ocean General Circulation Models (AOGCM), which, according to Winsberg, are "highly complex computer models that are constructed on the basis of both principled science--including fundamental partial differential equations from mechanics and thermodynamics--and trial-and-error approximations and parameterizations, and everything in between" (p. 107-108). The first climate model used is one developed by the Canadian Centre for Climate Modeling and Analysis, which Mika and Newman shorten to CGCM2, and the Hadley Center climate model,

or HadCM3. Both of these simulations are run in two scenarios called A2 and B2 which differ with respect to the rate of carbon-dioxide emission into the atmosphere. A2 corresponds to a situation with high human population growth and slow development of CO_2 emission-reducing technologies, projecting a relatively high atmospheric CO2 concentration by 2090-99. B2 corresponds to a situation with slower human population growth and more advancement and implementation of environmental protection technologies, projecting a relatively low atmospheric CO2 concentration by 2090-99 (Mika and Newman p. 216). Data from those simulations, including minimum and maximum temperature values, precipitation and relative humidity, was used in another simulation program, called CLIMEX, to create the bioclimatic envelope models for pea leafminers in North and Central America (Mika and Newman p. 216). CLIMEX essentially takes data about the current distribution and abundance of a species along with the current climate conditions then, using information from climate modeling programs like CGCM2 or HadCM3, predicts how that species will be geographically distributed in the future (Mika and Newman p. 215). The resulting bioclimatic envelope models for pea leafminers are found below in Figure 1. As we can see, CGCM2 and HadCM3 result in strongly conflicting predictions. For example, using scenario B2 we can see that CGCM2 projects that the southernmost portion of Texas will be "very favorable" for *L. huidobrensis* by the 2080s while HadCM3 predicts that the same region will be "unfavorable" at the same time (Mika and Newman p. 219). Both models have a similar conflict regarding the suitability of the environment for *L. huidobrensis* in the Yucatan peninsula as well as in other areas of North and Central America. Simply put, it seems that there is far more disagreement than agreement between the predictions resulting from CGCM2 and HadCM3.[6]

6 Unless, of course, you count the agreement between them on the unsuitability of Canada and Alaska as environments for *L. huidobrensis*. However, given the general unsuitability of typically cool climates for most

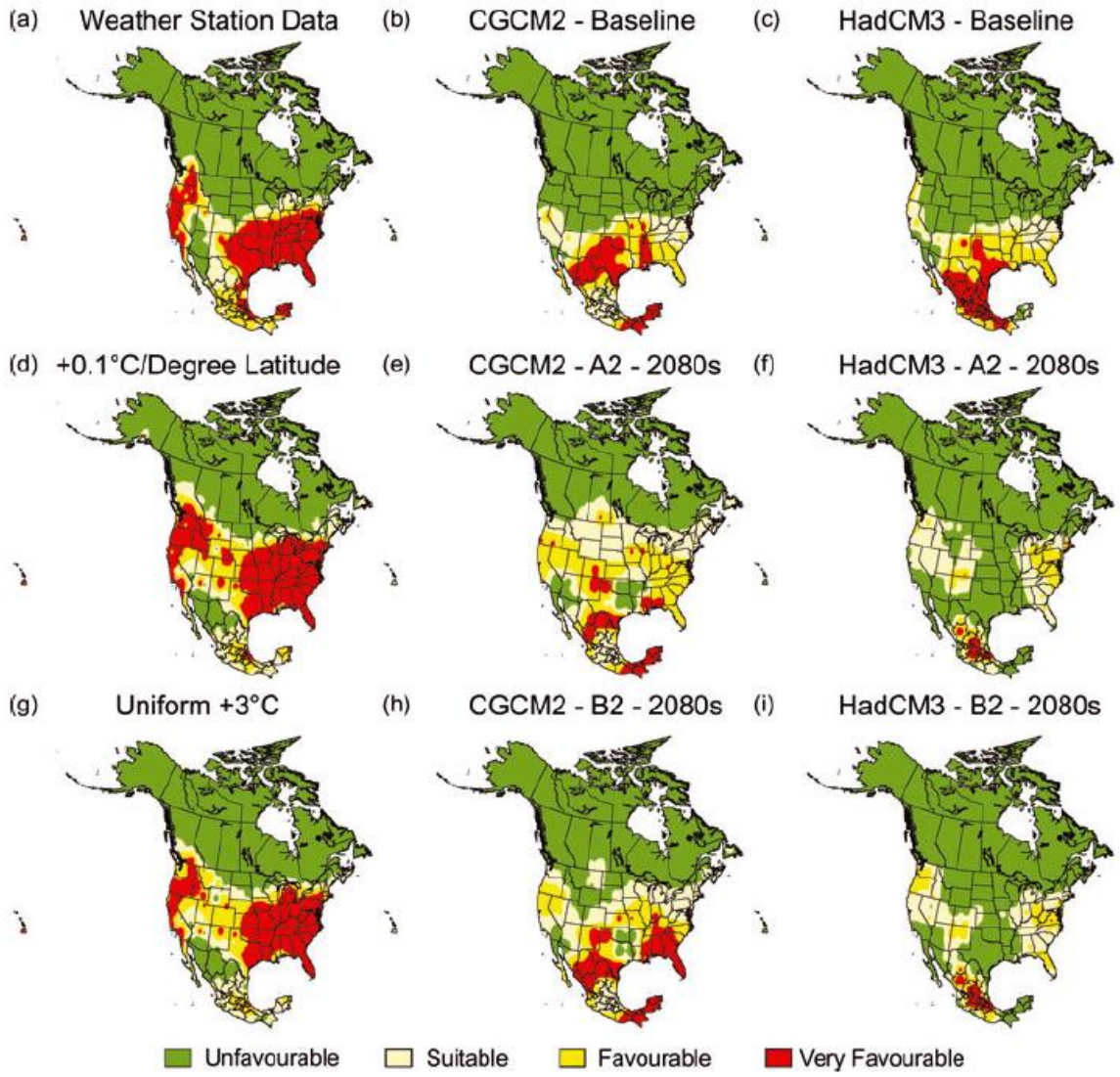

Figure 2 Ecoclimatic index (EI) values for *Liriomyza huidobrensis* in North America using the (a) weather station data and the (b) Canadian Centre for Climate Modelling and Analysis GCM (CGCM2) and (c) Hadley Centre climate model (HadCM3) baseline period (1961–1990). The results from the bioclimatic envelope model (BEM) when using (d) an increase of 0.1°C per degree latitude or (g) a uniform increase of 3°C everywhere are shown. The BEM results when using GCM climate change projections in combination with the A2 scenario for (e) CGCM2 and (f) HadCM3 are shown for the 2080s. The B2 scenario projections for (h) CGCM2 and (i) HadCM3 are also shown for the 2080s. 'Unfavourable' climate represents EI < 10, 'suitable' is 10 < EI < 20, 'favourable' is 20 < EI < 30, and 'very favourable' climate is EI >30.

Figure 1 Conflicting Bioclimatic Envelope Model Results (Mika and Newman p. 219)

insects, such agreement may not be very surprising or interesting. Nonetheless, in accordance with the robustness of prediction criterion suggested later, we could certainly be justified in trusting that aspect of CGCM2's and HadCM3's predictions.

3.2 Problems for Winsberg's Model-Building Principles

Can we decide between the conflicting results of CGCM2 and HadCM3 on the basis of Winsberg's criteria? Winsberg's criteria are meant to allow us to validate simulation results, or deem them acceptable data, but it appears that we can make a fairly strong case for validation of either set of results. Both use what can be considered good model-building principles. For example, CGCM2 employs a widely-used computational technique known as "flux adjustment" to manually adjust or account for certain parameter values which helps prevent the simulation from drifting into an unrealistic depiction of the Earth's climate. Moreover, deciding how specifically to make these flux adjustments required a process of "trial and error" which can be described as relying to some extent upon the intuitions of the researchers (Flato et al. p. 454). HadCM3, on the other hand, appeals to its increased grid resolution[7] for the oceanic component over its predecessor, HadCM2, which should allow it to make more fine-grained predictions (Collins et al. p. 62). HadCM3 also refers to a host of improvements over its predecessor's atmospheric model component, including the introduction of a new radiation scheme, the representation of convective momentum transport and a reformulation of treatment of clouds in the simulation, just to name a few (Collins et al. p. 62). We can try to sort out whether the individual techniques or principles used in one simulation are more validated than the techniques and principles used to build the other, but in many cases that will be very difficult, perhaps even practically impossible. As Winsberg points out, "[one] cannot attribute the predictive success or failure of a given model to a particular localized component of that model because the components of models are strongly coupled to one another and hence interact with one another in

7 Grid resolution in computer simulations is how finely the computer model divides the target system for calculating changes in the target system's behavior. In general, the greater the grid resolution is, the more detailed the simulation will be.

significant ways" (Winsberg p. 110). To clarify, Winsberg means that the modular[8] nature of our most complex simulations prohibits us from decisively analyzing how those simulations' results are produced because we cannot track how the hundreds of thousands of numbers are exchanged and calculations are carried out between different components of those simulations. By the same token, no individual technique or model-building principle of the simulation can be validated completely independently because it is only through the conjunction of several such techniques and principles that we are able to represent the behavior of a target system with any accuracy whatsoever. Also, we cannot wait around for 80 years to see which simulation results end up being correct, since the whole point of building the simulations is to know what the climate will be like without having to physically measure the state of the climate in 80 years.

So, the next best thing is comparing the models to past data and seeing which models best fit the data. Of course, building a simulation to reproduce past data, or calibrating the simulation, is one aspect of the model-building principles that Winsberg argues helps validate simulation results, since the ability to reproduce such data give us confidence in both our models and in our computational techniques. Problematically, both CGCM2 and HadCM3 have been built in accordance with and checked against various sets of climate data, so both can be said to be appropriately calibrated to the point that we can be justified in trusting their results (Flato et al. p. 466; Collins et al. p. 78-79). Certainly, each model might fit slightly differently to the available empirical data. For example, one simulation might reproduce very closely the temperature changes that have been measured in one region of the globe while another simulation reproduces very closely the data from another region. Nonetheless, both simulations do an acceptable

8 A "modular" simulation is one that has different modules, or components, to represent various aspects of the phenomenon that the simulation as a whole seeks to represent. For example, a modular climate model might have a separate component that accounts for precipitation and another that accounts for wind currents. These modules then communicate and share data with each other to produce the final results of the simulation.

enough job overall of reproducing past climate data that scientists consider them appropriate sources of data to use in studies for which climate projections are needed. So again, it seems like appealing to principles of model-building and past successes will not give us an appropriate means of distinguishing between the results of different simulations for the same system to tell which results are actually externally valid.

However, laboratory experiments sometimes give conflicting results as well. If so, then it seems that laboratory experiments which conflict in that way are no better off in terms of external validation than the simulations just discussed, *unless* we have a criterion for distinguishing between the results of conflicting experiments that is unavailable for conflicting simulations (or which does not externally validate the simulation results). My suggestion is that reproducibility of results is just such a criterion.

Chapter 4. Reproducibility

One major difference between our ability to validate the results of simulation and material experiment in terms of distinguishing between conflicting results lies in the significance of the reproducibility of the results. Here I argue that reproducing the results of a material experiment speaks, to a certain extent, for the external validity of the experiment, while repeating a simulation and getting similar results does not help externally validate the simulation. The reason for this is that when a simulation is run multiple times and shows consistent results, all that is demonstrated is that the simulation is functioning correctly, that is, the program is running on the computer as it was intended to run. Indeed, a given run of a simulation usually repeats its calculations many times, so each data set produced by a simulation is essentially a set of reproduced results. Nonetheless, as Winsberg suggests, other arguments are required to establish the external validity in such a case, like appealing to the successful history of the techniques involved in building the simulation or arguing that the model accurately represents the target system in relevant respects. For example, we can run a simulation that predicts global mean temperature change multiple times and get similar predictions or data sets each time, but that in itself does not give us any reason to believe that the predictions of that simulation are accurate other than showing that the prediction at least stays consistent and does not vary randomly.[9] After all, the simulation could be using a model that simply misrepresents the climate, but which runs just fine when implemented on a computer so that it consistently produces inaccurate results.

In the case of a material experiment, on the other hand, repeating the experiment and getting the same or similar results helps establish to a certain extent that a particular entity or

9 Of course, if multiple runs of the same simulation program making use of the same input data produced very dissimilar data sets, that would be a reason to not trust the results as well because the simulation is functioning inconsistently.

phenomenon has been isolated and exists in the real world as theorized or modeled. The reason is that repeatedly measuring or producing some quantity or effect in a laboratory environment means that we are, at the very least, manipulating a real-world system and gathering data from it in some reliable way. If nothing else, repeating a laboratory experiment and getting similar results shows that, for the kind of material system being studied in the laboratory, a certain effect is more likely to follow from a certain cause. For example, the repeatability of the famous double-slit experiment, in which a light source is shone through a pair of narrow slits which causes an interference pattern on a screen beyond the slits, is taken to demonstrate something about how light behaves in the world outside of the laboratory. In the case of simulations, running a simulation over and over again and amassing a pile of results which all largely fall within a certain close range of values only demonstrates the precision of the simulation's results but not necessarily their accuracy with respect to any real-world system or entity.

Moreover, reproducibility becomes even more important in establishing the external validity of material experiments when we consider the importance of reproducing the results not just by the original researchers using their own equipment, but also when other researchers in different labs reproduce the results. When other researchers reproduce results generated by a different lab, we can take that as evidence that the original result did not come from some experimenter error in the original study. Rather, the results are more likely to be describing accurately some natural phenomenon, whether they are measurements of some quantity such as the elementary charge or the rate of cell division in human embryos. Even when other researchers use different equipment or techniques to study the same phenomenon, if those different methods yield results in agreement with previously established results, it builds confidence that the results obtained are accurate. There are, for example, many ways to measure

the elementary charge, *e*. One traditional way involves Millikan's famous oil-drop experiment, while more modern methods make use of single electron tunneling devices that can manipulate individual electrons (Millikan 1913, Feltin et al. 2009). When these different methods nonetheless produce the same or very similar measurements of *e*, our confidence in those results as accurate measurements of the elementary charge grows.

One example of how reproducibility helps establish the external validity of material experiments, or rather, how being unable to repeat results is taken as a sign of external invalidity, is the controversy surrounding Stanley Pons and Martin Fleischmann's claim in the late 1980s that they had discovered "cold fusion" reactions. Pons and Fleischmann proposed that they had observed theoretically unaccounted-for excess heat and nuclear byproducts such as neutrons and tritium from electrochemically compressed deuterium in a palladium cathode that suggested that a nuclear fusion reaction had occurred at room temperature (Pons and Fleischmann p. 306). Given what was known about nuclear fusion at the time, namely that it typically occurs at very high temperatures such as those found inside stars, Pons and Fleischmann were presenting what could be considered conflicting experimental results. However, their claims came under fire and were ultimately rejected by most of the scientific community because most scientists were unable to reproduce Pons and Fleischmann's results. Had scientists been able to consistently reproduce the sort of effect described by Pons and Fleischmann, it would have been taken as evidence of the existence of a real phenomenon. Of course, we might try to blame the failure of replication on the scientists. However, assuming that the scientists attempting to recreate Pons and Fleischmann's results were trained in many of the same experimental techniques and had much of the same theoretical background as Pons and Fleischmann and had at least a general idea of how the original study was conducted, it seems unlikely that so many should fail to

reproduce the phenomenon if such a phenomenon actually existed. Again, all of this is to say that laboratory experiments enjoy increased confidence in their external validity when they are repeated with the same or similar results, while computer simulations are not externally validated by repeating them and getting the same results. Since the question of whether or not simulations are just as epistemically powerful as laboratory experiments lies within the extent to which we can validate their results, it seems that laboratory experiments have an advantage over simulations in arguing for their external validity because we can decide between conflicting experimental results based on the reproducibility of those results.

This is not to say that the reproducibility of results will always settle the question of the external validity of conflicting results between laboratory experiments. My point, though, is that reproducibility of results *can* help settle questions of external validity for conflicting laboratory experiments but not for computer simulations. Two different computer simulations may give conflicting results initially, but if we discover that one of those simulation's results cannot be reproduced, that does not support inferring that the other simulation is therefore producing externally valid results. It only supports the inference that the other simulation, because its results are reproducible, is at least a candidate for producing externally valid results. Other arguments are required to establish the simulation's external validity by appealing to Winsberg's model-building principles. For laboratory experiments, if two similar studies of the same phenomenon yield conflicting results, but only one set of results ends up being reproducible (say, a consistent measurement of some constant), then we seem justified in thinking that the reproducible result is the one that correctly reports facts about the phenomenon in question. Of course, there may be cases in which laboratory experiments produce conflicting results repeatedly. In such cases, researchers will have to begin making arguments for why their results

are externally valid while explaining away the conflicting results, and will be no worse off than simulationists in similar positions. So, sometimes material experiments have the problem of conflicting results which cannot be decided by appeals to reproducibility. However, these cases seem relatively rare. More frequently, as seen in the case of Pons's and Fleischmann's results, questions of the external validity of conflicting results in laboratory experiments get decided by a difference in the reproducibility of those results. For computer simulations, differences in the reproducibility of results may help us decide between conflicting results, but such differences will not help us establish the reproducible results of a simulation as externally valid. Remember, it is the external validity of simulations, or their ability to be informative about the systems they purport to represent, that Winsberg is concerned with. So, the advantage enjoyed by laboratory experiments over simulations in our ability to externally validate them threatens Winsberg's case for epistemic equivalence between the two methods of investigation.

Chapter 5. The Addition of Robustness to Winsberg's Criteria

5.1 Defining and Defending Robustness

While the problem presented with conflicting climate predictions suggests that Winsberg's principles of model-building are insufficient to warrant the same degree of trust in simulation results as we might have in the results of well-conducted laboratory experiments, the problem also suggests an additional criterion popular with other philosophers of modeling and simulation: robustness of results. Robustness of results, as Richad Levins originally describes it in "The Strategy of Model Building in Population Biology," occurs when "we attempt to treat the same problem with several alternative models each with different simplifications but with a common biological assumption. Then, if these models, despite their different assumptions, lead to similar results we have what we can call a robust theorem which is relatively free of the details of the model" (p. 423). While Levins discussed the concept of robustness in the context of mathematical models used in population biology, we can easily apply the idea to computer simulations in any area of scientific inquiry. When looking for robust results, we would examine the results of a wide range of simulations built with certain common assumptions, say, the central theory or equations said to describe the phenomenon we want to study. But these simulations should differ in how they are conducted, such as by using similar models but different computational algorithms to find solutions to the differential equations or by using different simplifying assumptions in creating the model of the phenomenon. For example, perhaps the simulations might differ in which aspect of the overall phenomenon we want to study, such as emphasizing precipitation over air currents, or vice versa, in studying climate change. Such simulations might have more complicated, realistic modules for simulating precipitation or air currents at the expense of having more simplified modules of other climatological phenomena.

Nonetheless, if these simulations converge on certain results, despite their differing simplifying assumptions and methodologies, then according to those who take robustness analysis seriously we should have greater confidence in those results.

Of course, some have objected to the idea of robustness of results as providing any additional confirmation of those results. In "A Critical Assessment of Levins's *The Strategy of Model Building in Population Biology (1966)*," Steven Orzack and Elliot Sober attack the notion of robustness as an indicator of truth by criticizing the form of the argument from robustness. Orzack and Sober present the argument from robustness in the following form (call it version 1 or V1):

"(1) M1 [a model] or M2 or … or M*n* is true.

(2) For each *i*, M*i* implies that some result, R, is true.

Therefore,

(3) R is true" (Orzack and Sober p. 538).

Orzack and Sober criticize this argument on the basis that if we change the first premise of the argument to either "we know each of the models is false" or "we do not know that one of the models is true," then the conclusion no longer follows from the premises (Orzack and Sober p. 538). Essentially, they claim that nothing follows from many agreeing model predictions because all of the models could be wrong.

However, V1 cannot be the best or most charitable interpretation of the argument from robustness. As Levins claims regarding drawing conclusions from models, "our truth is the intersection of independent lies" (Levins p. 423). All models involve idealizations and idealizations are falsehoods, therefore no model can be considered "true" in any straightforward sense. So, it is clear on standardization V1 that premise 1 must be false, not that it simply could

be false. While the argument as Orzack and Sober present it is valid, they have presented it in such a way that guarantees the first premise will come out false, which seems to assume the worst of appeals to robustness.

Wendy Parker also recognizes that Orzack and Sober's formulation of the argument from robustness is flawed in the way just mentioned in "When Climate Models Agree: The Significance of Robust Model Predictions," (p. 583-4). So, she constructs a different version of the argument that may be more appealing (call it version 2, or V2):

"(1') It is likely that at least one simulation (or model) in this collection is indicating correctly regarding hypothesis *H.*

(2') Each of the simulations (or models) in this collection indicates the truth of *H*.

Therefore,

(3') It is likely that *H"* (Parker p. 584).[10]

Notice that (1') accounts for the fact that simulations typically are not, in a straightforward sense, true. Recall that in Chapter 2 I pointed out that simulations often incorporate deliberate falsehoods, or fictions, that nonetheless have a proven track record of getting the data right and producing correct results and predictions. By understanding robustness of results to be about the likelihood of a model producing accurate results (rather than being a perfectly accurate representation of a phenomenon), we avoid the problem of having the first premise always come out false. In avoiding that problem, V2 is superior to V1.

10 Parker uses the somewhat awkward phrasing "indicates correctly" to talk about the results of a simulation being able to serve as evidence indicating the truth of some hypothesis without being committed to the simulation itself being a "true" representation of a phenomenon (p. 584). Although Parker does not explicitly say so, this move may be motivated in part by the problems facing the possibility of taking a realist attitude towards the representations of phenomena in simulations. After all, these simulations often employ, as mentioned earlier in chapter 2, fictional components that correspond to no real processes or entities in the target system being modeled.

Nonetheless, I think V2 is still problematic because it does not capture what I take to be the argumentative force of robustness. Consider instead the following interpretation[11] of the robustness argument (call it version 3 or V3):

(1*) M1 and M2 and … and M*n* are all considered well-built models by the scientific community which employ a diverse range of methods for representing the target phenomenon.

(2*) For each *n*, Mn implies that R is true.

Therefore,

(3*) R is more likely to be true (or, increased confidence in R is warranted).

V3 has several advantages over V1 and V2. V1's first premise involves a disjunction with the assumption that one of the disjuncts (a model) is true. However, if we have that knowledge alone, it seems as though the joint implications of the other models (which may or may not be true) that R is true would not add to our confidence in R because they could be conceptually ill-conceived models. In other words, robustness is not really doing any work in V1: the fact that *one* of the models is true means that of course something implied by that specific model will be true, so the fact that there are other models (which may not be true) making the same prediction should not really increase our confidence in R. V2 makes a similar appeal to "at least one" of a set of models or simulations producing a correct result. So, it is unclear what contribution the other models, which may or may not be "indicating correctly regarding hypothesis *H*," are making to our confidence in the shared results of those simulations. The first premise of V3, on the other hand, appeals to all of the models under consideration being approved at large by the scientific community on the basis of being "well-built" (which could include features like general empirical adequacy, appealing to the past history of success of modeling techniques used

11 I credit Wendy Parker's article "When Climate Models Agree" for inspiring this line of thought, and it meshes well with Winsberg's criteria of model-building also.

in constructing the model, etc.). If we have reasonably justified confidence in a set of models, and all of those models imply some fact R then it seems like we would be more justified in believing R than if we were only to justify belief in R by looking at any individual model. Notice that now it is the agreement between models that is really doing the work in justifying belief in R, not the fact that we think it is probable that one of a set of models is right and makes a certain prediction. I think this interpretation of the argument from robustness better captures why scientists might trust robustness as a justification for belief in some prediction or theorem R. It is not that belief in R is warranted based on a belief that one of a number of models must be true (perhaps in some probabilistic sense that "out of all of these, surely *one* must be true"), so if all of those models agree on a certain R then R must be true. Rather, if a set of models that are considered generally reliable or are well-founded all imply that R is true, that gives us a better reason to think that R is true than if only one such model implies R is true.

The question remains, of course, whether the fact that robustness of results itself contributes anything to our confidence in those results over and above the individual contributions of the models and simulations that generate the results. In other words, we might ask whether the fact that a given result is robust has any extra power to boost our confidence in those results other than from simple aggregation of our confidence in the simulations which produced the results. My answer to the question is that robustness of results contributes an extra degree of confidence over the combined confidence we have in the contributing simulations proportional to the number of contributing simulations and the degree of difference between those simulations. Thus, we can express the extra confidence from robustness in the following

way[12]:

$$R = N * D$$

Where *R* is the additional confidence from robustness, *N* is the number of simulations contributing to the robustness of the result and *D* is the subjective assessment of diversity between the *N* simulations.

This extra degree of confidence aligns nicely with the form of the argument from robustness given in V3 because the number of contributing simulations matters, as does the degree of difference in methodology between each of the simulations. As diversity of models increases, so too does the likelihood that we obtain a robust result decrease, since it becomes more and more likely that different methods and models used in the simulations result in greater discrepancies in their results. So, a greater number of simulations with greater methodological differences nonetheless converging on the same result gives us greater confidence in that result because agreement due to mere coincidence becomes less and less likely. Notice also that if the diversity of models is zero (there's only one contributing simulation and thus no methodological difference to be spoken of), then we get the intuitive result that no additional confidence from robustness is warranted. Likewise, running the same simulation over and over again and getting the same result will not count as a robust result (and thus cannot generate any increased confidence from robustness) either, since there will be no model diversity.

12 This equation is not intended to be a precise mathematical formulation to be filled in with units and other specifics. Rather, it is a simple heuristic device for explaining where the additional confidence in a result from robustness comes from. Regarding the form the equation takes, I give equal weight to number of involved simulations and subjective assessment of diversity in the simulations because the argument from robustness seems to rely equally on the number and diversity of simulations producing robust results. For without diversity in simplifying assumptions, the appeal to robustness loses its force. And if there is only one simulation, the argument cannot be made at all. Moreover, the form cannot be additive, since each simulation cannot have its own unique "diversity coefficient" because the sense of diversity important to robustness is a result of the differences between the set of simulations involved taken as a whole.

Here one might take issue with the robustness criterion by questioning what exactly is meant by "getting the same result," or getting "similar" results. The data sets generated by simulations are often large and cumbersome, requiring a fair amount of work to make presentable so that meaningful conclusions can be drawn from them.[13] How then are we to determine what counts as two simulations getting "the same" or "similar" results? My suggestion is that determining whether there is sufficient agreement between simulation results to qualify them as robust results is a matter of judgment by members of the scientific community. Recall from Chapter 2 that one of Winsberg's principles of model-building involves the soundness of researcher intuitions. I would like to make a similar appeal here and claim that it would be unreasonable to require multiple simulations to produce exactly "the same" results in order to declare a robust result and that any uncomfortable vagueness resulting from the term "similar" results from simulations can be dealt with by appealing to researcher experience and training.

Clearly, if what is required in order to declare a given result robust is that all of the simulation generating the result produce the exact same result to the last significant figure, then almost no simulation results will ever be robust. Such an exacting standard would render robustness of results a practically useless criterion in determining the external validity of simulation studies. Simply appealing to the similarity of results seems to be an undesirable solution because it seems too vague since we have no metric of similarity. Moreover, it seems unclear how there could even be a metric of similarity that consistently applies to all of the diverse disciplines of science. For example, in astrophysics, being in the same order of magnitude might be sufficient for robustness because of the great scale involved in studying the

13 For a more detailed discussion of how and why "raw data" from simulations are processed into a form "usable by scientists," see Patrick Suppes's "Models of Data" and Todd Harris's "Data Models and the Acquisition and Manipulation of Data" (Harris p. 1510).

cosmos, whereas a population biologist might require a greater degree of precision when deciding whether the results of several studies are all close enough to warrant declaring a robust result.

The different standards of similarity for different disciplines suggest a natural answer to the question of how to determine similarity between simulation results: it is a matter of judgment by scientists. If we trust in the training and competence of the scientific community, we can appeal to that training and competence as being able to determine when a set of results from different simulations are similar enough to declare the results robust. Because of their expertise in comparing data sets and interpreting results, researchers are capable of judging the degree of similarity required for robustness in the same way that, say, skilled billiard players can judge angles and difficulties of shots. Of course, this approach to understanding the degree of similarity required for robustness leaves open the possibility of developing a more detailed metric for determining similarity of results in the future. Developing such a metric, though, is up to the members of the scientific community who are engaged in the different disciplines of the sciences.

5.2 Adding Robustness to Winsberg's Criteria

Robustness of results, or widespread agreement on a given result amongst the majority of our best models and simulation results, needs to be added to Winsberg's list of criteria for validating simulation results. Following Parker, I do not think that robustness alone is sufficient as a standard upon which to base our confidence in simulation results. Parker suggests that for the current state of climate modeling, we could (and do) have many computer simulations which largely agree regarding some results, but which we are not justified in placing our confidence in that prediction based upon their agreement (Parker p. 581). One reason robustness by itself is not

sufficient to warrant confidence in a set of simulation results is that one could artificially create a sense of robustness by creating a number of conceptually very poor simulations which one fine-tunes or manipulates to agree with a more conceptually well-designed model built in accordance with Winsberg's principles. Certainly, we could have confidence in the results of the conceptually well-designed model insofar as it is built in accordance with Winsberg's principles of model-building, but in the situation just described robustness would contribute nothing to our confidence in those results because the other simulations are admittedly poorly designed. However, if all of the models are in general agreement on a certain result (making it a robust result) *and* all of the models which make that prediction can be said to be built more or less in accordance with Winsberg's principles of model-building, then perhaps we could have just as much confidence in those results as we do in the results of well-conducted laboratory experiments. Certainly, robustness of predictions is not quite the same thing as repeatability, but it might nonetheless serve the same function for externally validating simulation results as repeatability does for laboratory experiments. After all, robust results give us confidence in those results by showing that the same results can be obtained from a variety of well-built simulations. Or, if robustness is lacking because multiple studies are producing conflicting results, then we know we do not have sufficient justification for trusting the results of the simulation. Likewise, in laboratory experiments, repeating and experiment and reproducing its results indicates that some phenomenon that actually exists in nature has been isolated or, if the experiment does not yield consistently reproducible results, we know we should not trust the results of the initial experiment. Additionally, in the same way that the convergence on a certain result by a diverse array of models increases our confidence in that result by indicating that the result is unlikely to be an artifact of a particular simulation used to obtain the result, reproducibility of results in

laboratory experiments helps rule out that a given result was produced by an error on the part of a particular experimenter. Robustness of results, in other words, is how we make reproducibility of results meaningful in computer simulation studies. Running the same simulation program again and again and getting the same results boosts our confidence very little in the external validity of those results, but as I have argued above, reproducing results with a variety of well-built models and simulations does support their external validity.

The problem raised by Mika and Newman's study for Winsberg's criteria was that it seems we could justify confidence in the results of either bioclimatic envelope model for *L. huidobrensis* based on CGCM2's or HadCM3's predictions if we rely only upon Winsberg's criteria. However, by adding the robustness criterion, we can say that neither result is sufficiently validated because the predictions do not agree, that is, the results are not robust. Moreover, assuming we continue to create bioclimatic envelope models for *L. huidobrensis* using other climate simulations in which we can also justify confidence based on appeals to Winsberg's principles of model-building, if we discover that many of those climate models turn out to create agreeing bioclimatic envelope models then we can be justifiably confident in those bioclimatic envelope models. So, by combining Winsberg's principles of model-building and the robustness criterion, we are able to produce a stronger set of criteria for externally validating simulation results that has the potential to give us just as much justification for trusting simulations as we might have for trusting well-conducted and widely reproduced laboratory experiments.

Chapter 6. Conclusion

Winsberg's point in arguing for equal epistemic power for laboratory experiments and computer simulations is that both require arguments for their external validity. I began by explaining his criteria for the validation of simulations, the principles of model-building. I then showed that his principles of model-building were insufficient to argue for the external validation of simulation studies in certain kinds of cases in which simulation results conflict, and that laboratory experiments have an advantage over simulations in that conflicting experimental results can be decided between on the basis of repeatability. I then suggested that robustness of predictions could serve the same function for simulations as repeatability does for laboratory experiments in either adjudicating between conflicting results or allowing us to say that we do not have sufficient justification to externally validate the results. I also argued for a certain interpretation of the argument from robustness that involved appealing to the convergence of many well-built and diverse models rather than the more common version which involves appealing to the probability that one of a set of models is likely to be true. My interpretation strengthens the case for taking robustness as an additional requirement for the validation of simulation results and ultimately supports the idea that computer simulations, under certain conditions, can provide us with information about the world that is just as trustworthy as data from more traditional laboratory studies.

Arguing for the epistemic equivalence of computer simulations and laboratory experiments may have interesting consequences for how we think about the nature of scientific evidence. For much of the history of science, for data to count as evidence in favor of some particular hypothetical description of a phenomenon, the data are supposed to come from some physical system that falls under the descriptive domain of the hypothesis being tested. But, if the

results of computer simulations can count as data in which we can place as much confidence as we might place in the results of laboratory experiments or field observations, then it is clear that, at least when certain conditions are met, we can get novel information about particular instances of natural phenomena without having a great deal of physical access to those instances of phenomena.

Moreover, the ability of computer simulations to generate novel information about real-world systems complicates the relationship between theory and evidence. According to Winsberg's model-building principles, part of what justifies taking the results of computer simulations as data about a real-world system is that the simulation is based upon sound scientific theory. The typical relationship between theory and evidence is often thought to be that one proposes a theory, derives predictions from it, and then designs an experiment to see if those predictions are correct. But in the case of computer simulations, one uses theory to create a computerized model which then produces results that are *themselves taken to be data*. No further empirical work is necessary to "confirm" the prediction of the theory in these cases; the results of the computer simulation are taken to be data regarding the natural system being modeled.

There are other questions raised by this paper that cannot be properly addressed within its scope. For example, the fact that outright fictional components, such as the artificial viscosity term mentioned in Chapter 2, can contribute to very successful predictions may be a threat to the success-to-truth rule of inference that serves as the cornerstone of the "no miracles" argument for scientific realism. Indeed, Winsberg raises this question in the penultimate chapter of *Science in the Age of Computer Simulation*. Another question that remains is what role computer models play, or can play, in scientific explanation. That many computer models can employ differing simplifying assumptions, yet nonetheless generate robust results, only complicates the matter.

Suppose a diverse array of computer models of a phenomenon did generate robust results about the behavior of that phenomenon: which model, if any, can be said to explain the behavior of the phenomenon? None of them? All of them, but in different ways? Whatever the answers to these particular questions turn out to be, the growing popularity of computer simulations across various scientific disciplines makes it very likely that continued exploration of computer simulations will yield fruitful new directions for answering some of the most central questions in the philosophy of science.

References

Bear, M.; Connors, B.; and Paradiso, M. *Neuroscience: Exploring the Brain*. Third edition. Philadelphia: Lippincott Williams & Wilkins, 2007.

Collins, M.; Tett, S.F.B.; and Cooper, C. "The internal climate variability of HadCM3, a version of the Hadley Centre coupled model without flux adjustments." *Climate Dynamics*, 2001. Vol. 17: pp. 61-81.

Feltin, N.and Piquemal, F. "Determination of the elementary charge and the quantum metrological triangle experiment." *The European Physical Journal Special Topics*, 2009. Vol. 172, pp. 267-96.

Flato, G.; Boer, G.; Lee, W.; McFarlane, N.; Ramsden, D.; Reader, M.; and Weaver, A. "The Canadian Centre for Climate Modelling and Analysis global coupled model and its climate." *Climate Dynamics*, 2000. Vol. 16: pp. 451-467.

Fleischmann, Martin and Stanley Pons. "Electrochemically induced nuclear fusion of deuterium." *Journal of Electroanalytical Chemistry*, 1989. Vol. 261: pp. 301-307.

Halliday, D.; Resnick, R.; and Walker, J. *Fundamentals of Physics*. Seventh edition. John Wiley & Sons, Inc., 2005.

Harris, Todd. "Data Models and the Acquisition and Manipulation of Data." *Philosophy of Science*, Vol. 70, No. 5: pp. 1508-17.

Heikkenen, R.; Luoto, M.; Araujo, M.; Virkkala, R.; Thuiller, W.; and Sykes, M. "Methods and uncertainties in bioclimatic envelope modelling under climate change." *Progress in Physical Geography*, 2006. Vol. 30, No. 6: pp. 751-777.

Levins, Richard. "The Strategy of Model Building in Population Biology." *American Scientist*,1966. Vol. 54, No. 4: pp. 421-31.

Mika, Anna and Jonathan Newman. "Climate change scenarios and models yield conflicting predictions about the future risk of an invasive species in North America." *Agricultural and Forest Entomology*, 2010. Vol. 12: pp. 213-221.

Millikan, Robert. "On the Elementary Charge and the Avogadro Constant." *The Physical Review*, 1913. Vol. 2, No. 2: pp. 109-43.

Parker, Wendy. "When Climate Models Agree: the Significance of Robust Model Predictions." *Philosophy of Science*, 2011. Vol. 78, No. 4: pp. 579-600.

Suppes, Patrick. "Models of Data." In *Logic, Methodology, and Philosophy of Science: Proceedings of the 1960 International Congress*. ed. E. Nagel, P. Suppes, and A. Tarski. Stanford: Stanford University Press, 1962. pp. 252-61.

Teller, Paul. “Fictions, Fictionalization, and Truth in Science.” In *Fictions in Science: Philosophical Essays on Modeling and Idealization*, ed. M. Suarez. New York: Routledge, 2009. pp. 235-247.

Winsberg, Eric. *Science in the Age of Computer Simulation*. Chicago: University of Chicago Press, 2010.