UC Irvine

UC Irvine Electronic Theses and Dissertations

Title

Challenges in the Epistemology of Large-Scale Simulation

Permalink

https://escholarship.org/uc/item/2fv5g2c9

Author

Kadowaki, Kevin

Publication Date

2022

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library University of California

UNIVERSITY OF CALIFORNIA,
IRVINE

Challenges in the Epistemology of Large-Scale Simulation

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Logic & Philosophy of Science

by

Kevin Kadowaki

Dissertation Committee:
Professor James Owen Weatherall, Chair
Chancellor's Professor Jeffrey A. Barrett
Professor Cailin O'Connor
Professor Manoj Kaplinghat

2022

© 2022 Kevin Kadowaki

DEDICATION

To all the teachers and mentors who have guided me along my way,
and in memory of James Harrington—
without whom I would not have set down this road.

TABLE OF CONTENTS

Page

LIST OF FIGURES **iv**

LIST OF TABLES **v**

ACKNOWLEDGEMENTS **vi**

VITA **vii**

ABSTRACT OF THE DISSERTATION **viii**

INTRODUCTION **1**

0.1 Background 1

0.2 Outline 3

1 A Critical Analysis of V&V **6**

1.1 Separability—Descriptive vs. Prescriptive 8

1.2 Verification & Validation 10

1.3 Independent Validation & Simulation Development 14

1.4 Verification & the Method of Manufactured Solutions 16

1.5 Conclusion 26

2 Simulation Verification in Practice **28**

2.1 A Survey of Galaxy Formation Simulation Codes 29

2.2 Fluid-Mixing Instabilities and Test Development 36

2.3 Leveraging both Physics and Numerics 41

2.4 Conclusion 46

3 Simulation and Adequacy-for-Purpose **48**

3.1 The Aims of Adequacy-for-Purpose 49

3.2 A Challenge for Adequacy-for-Purpose 53

3.3 An Ecosystem of Adequate-for-Purpose Models 56

3.4 Conclusion 60

Concluding Remarks **62**

Bibliography **64**

LIST OF FIGURES

Page

1.1 An expanded layout of the various components in the V&V framework. . . . 13

1 A Critical Analysis of V&V

LIST OF TABLES

Page

2 Simulation Verification in Practice

2.1 Hydrodynamics tests. Unless otherwise indicated, the test results as run by a particular code is recorded in the paper indicated at the top of each respective column column. The * citation indicates that a different test setup was cited. 31

2.2 Magnetohydrodynamics tests. As in Table 2.1, unless otherwise specified, the test results as run by a particular code is recorded in the paper indicated at the top of each respective column column. Each test is based on the setup given in the paper cited in the first column, with the exception of the MHD shocktube category: for those marked with *, the cited test was performed *instead*; for those marked with †, the cited test was performed *in addition*. . . 33

2.3 Self-gravity tests. 34

2.4 Common magnetohydrodynamics tests. 35

ACKNOWLEDGEMENTS

I have come this far only with the assistance of many mentors. First and foremost, I would like to thank my advisor, Jim Weatherall, as his support, encouragement, feedback, and guidance enabled me to write this dissertation and prepared me to to be a philosopher of science. I also owe many thanks to Cailin O'Connor, Jeffrey Barrett, and Manoj Kaplinghat, as their feedback and guidance have substantially improved my efforts herein.

I would also like to thank other philosophers and physicists who have supported me at various points in my academic career. I am particularly grateful to J.B. Manchak, Chris Smeenk, Eric Winsberg, and Wendy Parker for their feedback and guidance as I began my research career, and to the faculty of the Irvine LPS department for creating a learning environment where I could pursue the kind of philosophy of science that I wanted to pursue. I am thankful for the encouragement that Hugh Miller, Julie Ward, James Murphy, and James Harrington provided in my undergraduate career, as I do not believe I would have pursued philosophy without their inspiration. And finally, I would like to thank Anne Smith, Robert McNees, and Jesus Pando for instilling in me a love for physics and the skills I needed to become a philosopher of physics.

Though there are too many individuals to name here, I am grateful to all the friends and peers who have accompanied and assisted me along my academic journey. In particular, I am forever indebted to James Keene, for his steadfast friendship and for commiserating with me in my existential moments; Kino Zhao, for being a role model of a principled and pragmatic academic; and Greg Lauro, for providing me with a steady diet of board games. (Without these individuals, I doubt I would have made it through the pandemic lockdowns.)

In my capacity as a graduate worker, I am grateful for the solidarity of thousands of UAW 2865 members, past and present; in particular, I would like to thank Jeremiah Lawson, Kavitha Iyengar, and Mike Miller for their tireless dedication to the collective good.

I would like to thank the administrative staff at UCI, particularly Patty Jones and Dan Paley, for the behind-the-scenes support they provided along my journey. I also owe a special debt of gratitude to Travis LaCroix for his LaTeX template.

Finally, I would like to thank my parents, as their love and support made all of this possible.

The work in this dissertation was made possible in part through the support of grant #61048 from the John Templeton Foundation. The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the John Templeton Foundation.

VITA

Kevin Kadowaki

Ph.D in Logic & Philosophy of Science **2022**
University of California, Irvine *Irvine, California*

M.A. in Philosophy **2020**
University of California, Irvine *Irvine, California*

M.S. in Physics **2016**
DePaul University *Chicago, Illinois*

B.A. in Philosophy **2014**
Loyola University Chicago *Chicago, Illinois*

B.S. in Theoretical Physics/Applied Mathematics **2014**
Loyola University Chicago *Chicago, Illinois*

PUBLICATIONS

"A Note on Saari's Treatment of Rotation Cuve Analaysis **2018**
The Astrophysical Journal

SELECTED PRESENTATIONS

"Verification, Validation, etc." **Nov 2021**
Philosophy of Science Biannual Meeting

"On Rotation Curve Analysis" **July 2018**
Foundations of Physics

ABSTRACT OF THE DISSERTATION

Challenges in the Epistemology of Large-Scale Simulation

By

Kevin Kadowaki

Doctor of Philosophy in Logic & Philosophy of Science

University of California, Irvine, 2022

Professor James Owen Weatherall, Chair

Contemporary astrophysics and cosmology, like other areas of science that study remote or difficult to manipulate subjects, must often rely on sophisticated computer simulations. These simulations can enable investigations that would otherwise be impossible, but they also come with costs: the numerical methods used raise the spectre of numerical errors, which in turn threatens the reliability of the simulations themselves. And while some sources of numerical error are well-understood and can, in some circumstances, be accounted for, the highly nonlinear nature of these simulations makes it exceedingly difficult to categorically rule out all possible sources of error. In this dissertation, I develop and defend an epistemic framework for thinking about these simulations which is informed by scientific practice in astrophysics and cosmology and corrects deficiencies in previous philosophical accounts.

Introduction

0.1 Background

Ordinary experimental and observational techniques require scientists to have a substantial degree of access to the subject of investigation. This may involve bringing the target system into the laboratory or creating the target system (or an analog model thereof) in the laboratory, so that the system can be observed in a controlled environment subject to deliberate manipulations by the experimenter. If these options are impossible, impractical, or unethical, scientists can observe the target system or its traces in its natural environment, accounting for exogenous factors as carefully and completely as possible. These techniques may be highly indirect and the methods themselves may require additional justification, but in any case scientists aim to ensure that they have observed and probed the salient details of the phenomenon, and (ideally) proportion their confidence in their reported conclusions to the degree to which they are justified in believing that their methods have granted them undistorted access to the target phenomenon.

Where access to the target system on the basis of experiment and observation is incomplete, simulation techniques may be necessary—for example, the systems of interest in astrophysics and cosmology are often systems of precisely this kind, largely because of the scales and complexity involved. Consider, for example, the process of galaxy formation. Galaxies are on the order of ∼10-100 kpc, and form on cosmological timescales. It goes without saying that galaxies cannot be brought into a laboratory, and neither can we construct fully robust analog models in a laboratory—the diversity and exoticness of components such as dark matter and supermassive black holes cannot be replicated in a laboratory setting. The cosmological timescales also prevents effective timeseries observation, as even a hypothetical observer with a telescope trained on a galaxy for the entirety of human history would account for barely a snapshot of that galaxy's lifecycle. And finally, the processes involved in galaxy formation are manifold and nonlinear, and themselves interact in highly nonlinear ways, which ensures that simple analytical models are out of the question.[1] These factors together make it impossible to solve for an analytical solution to a galaxy model, or even construct semi-analytical solutions with enough reliability to compare against empirical observation.

Simulations, on the other hand, can numerically solve for the outcome of these various components and their interaction, bypassing the need for analytical solutions, to produce predictions which (in principle) can be compared to observations—thus making even very complex systems amenable to study. With the appropriate resolution schemes, simulations can also represent large systems and evolve a process over massive timescales, circumventing the problems of scale. Thus, modulo various epistemic issues to be be raised shortly, simulations can provide a workaround for some of these problems that make normal scientific methods intractable for certain kinds of systems.

Continuing with our example from astrophysics, a full galaxy formation simulation requires a number of layers. (The following is only intended to be sufficient to motivate the philosophical

[1] I should emphasize that none of these factors—size, timespan, or complexity—is alone indicative of the necessity of simulation, as many systems that are larger, or longer-lived, or more complex have been the subject of successful scientific study without the need for simulation methods. The ΛCDM model of the universe, for instance, obviously represents a system much larger and longer-lived than any galaxy—but this is made tractable by the cosmological principles of homogeneity and isotropy, which allows for the application of equilibrium thermodynamics. Sociologists and economists typically study systems that are highly distributed and complex, but the phenomena of interest happen on more human timescales. This is also not to imply that large and long-lived systems are the only extremes of scale relevant here—the molecular resolution of some chemical reactions, which are obviously small and can happen on timescales too small to be probed effectively, can also require simulation methods.

issues described in the following section, and thus it is not intended to be an exhaustive description.) First, there is underlying dark matter, which is essentially a course-grained N-body gravitational simulation—already a nonlinear phenomenon. Next, a hydrodynamical simulation, representing ordinary matter, self-interacts both gravitationally and via pressure, and couples gravitationally to the dark matter component. Finally, there are a number of subgrid physical effects that must be accounted for, including star formation, supernovae, magnetic fields, etc. These effects are not only nonlinear—by definition, they are the product of baryonic physics that occurs on scales smaller than the simulation resolution, and thus must be instantiated in the simulation as separate components.

Of course, the scientific opportunities of simulation do not come without costs. The numerical methods used—which allow the scientist to bypass intractable analytic solutions—raise the spectre of numerical errors. Indeed, where the simulations are complex, some unphysical numerical errors may be inevitable, and the simulationist's task may be a matter of selecting the numerical methods which are associated with acceptable types of errors.[2] And while some sources of numerical error are well-understood and can, in some circumstances, be accounted for, the high degree of complexity in simulations makes it exceedingly difficult to rule out all possible sources of numerical error as a categorical matter. The very complexities that necessitate simulation are the source of potential barriers to its trustworthiness.

0.2 Outline

As these methods are becoming increasingly necessary in these fields and other scientific fields, philosophers of science have engaged with the novel epistemological issues that arise in these contexts—questions such as how simulations can provide us with knowledge, under what conditions trust in simulations is warranted, and even whether and to what extent these

[2] For example, see Hopkins (2015), Table 1, for an array of hydrodynamic methods and the known numerical errors associated with them.

questions are qualitatively distinct from those raised by more traditional scientific methods. In what follows, I will focus on the central epistemic question: Under what circumstances can we be confident that a simulation is providing us with knowledge about the target system it purports to represent?

In particular, many common intuitions about the proper structure of simulation epistemology are captured by the Verification & Validation (V&V) framework. Briefly, the V&V framework prescribes a careful separation between different simulation justification techniques—segregating the analysis of numerical errors from any comparisons between the simulation and real-world data—to prevent errors in one part of the simulation from spoiling the analysis in other parts. Winsberg (who dubs this notion the "separability thesis") has argued that—at least in highly complex simulations such as those used in climate science—V&V are not always so cleanly divided.

In the first chapter of my dissertation, I introduce the V&V framework and distinguish between two readings of the separability thesis: a descriptive claim, which merely states that V&V can be separated in all simulations of interest; and as a prescriptive claim, according to which V&V must be separated to achieve epistemically justified results. Crucially, arguments addressed to the former claim may have no bearing on the latter, and vice-versa; I clarify that while I reject both, I will address the former in the first chapter and the latter in the second and third chapters, respectively.

In the remainder of the first chapter, I argue against the descriptive separability thesis. I begin by showing that some of the conceptual distinctions drawn by V&V are insufficient to capture epistemically salient details of how simulations are justified, and I supplement these with additional distinctions. I expand Winsberg's argument against independent verification; in particular, where Winsberg merely asserts that the mathematical arguments that can be given in favor of independent verification are weak, I analyze the available methods in light of my supplemental distinctions to show how and why they are weak. I also assess Winsberg's argument against independent validation, showing that it implicitly relies on his argument against independent verification.

In the second chapter, I turn to a close examination of one major class of methods for simulation justification: verification tests. I begin by surveying a representative sample of galaxy formation simulation codes from the last two decades, showing that they do not display the pattern of verification test accumulation one would expect if the standard practice of the field followed the V&V framework. By closely examining the literature on a subset of these tests, I demonstrate that the development processes for these tests belie some of the tacit assumptions built into the V&V account—in particular, these tests are designed to permit researchers to probe the space of possible simulation codes, and not just to confirm the numerical fidelity of a particular code. Based on these observations, I argue that the V&V framework is needlessly overcautious in its separability prescriptions; in the appropriate circumstances, simulationists can and should allow the numerical and physical aspects of simulation justification to support one another.

The third and final chapter of my dissertation gives a more general explication of this epistemic framework. Drawing on the adequacy-for-purpose framework, I characterize the problem of model assessment under conditions of scarce empirical evidence. I argue that, while a single simulation may not suffice under these conditions, a suitable collection of simulation codes may be used in concert to advance a community's scientific understanding of a target phenomena and provide a foundation for the progressive development of more adequate models.

Chapter 1

A Critical Analysis of V&V

The literature on epistemology of simulation is sometimes framed in terms of *verification* and *validation* (V&V), concepts imported from the engineering literature that correspond to different possible ways in which a simulation can fail to adequately represent the target system under study. Under this framework, a simulation is *verified* to the extent that we are confident that the numerical methods employed in the simulation faithfully approximate the intended theoretical model; *validation*, on the other hand, despite being used ambiguously in the literature, generally indicates success of the simulation in representing the target phenomenon in question.

This division of the sanctioning of simulations into aspects seems to suggest a natural epistemic ordering of these activities: specifically, the notion that before one can compare the results of a simulation to observational or experimental data or draw any strong conclusions from this comparison about the adequacy of the model used to construct the simulation, we must first be confident that the simulation adequately implements the model in question. Hence, verification would seem to be both separable from validation and a precondition for validation, as otherwise numerical errors might compensate for errors in the choice of model—thereby providing “the right answer for the wrong reason” (Morrison, 2015, 259). However, upon examination of scientific practice, this picture is not so clear. In particular, Winsberg (who dubs this notion the “separability thesis” (Winsberg, 2018, 156)) has argued that—at least in highly complex simulations such as those used in climate science—V&V are not always so cleanly divided. Morrison (2015), Beisbart (2019a), and Jebeile and Ardourel (2019), in turn, have objected to Winsberg’s arguments against the separability thesis. In this chapter, I aim to further clarify the terms of this debate and defend and expand Winsberg’s thesis.

In Section 1.1, I examine the “separability” thesis as framed by Winsberg and others, and I argue that this thesis can be read in two distinct ways—as either a descriptive or prescriptiive claim. Crucially, arguments addressed to the descriptive reading of this claim may have no bearing on the prescriptive claim, and vice-versa; I address the former in this chapter, and the latter in the following chapters. In Section 1.2, I argue that some of the conceptual distinctions drawn in the current literature on epistemology of simulation are insufficient to capture many of the epistemically salient details about how simulations are sanctioned. As Beisbart (2019b) has already pointed out, philosophers and scientists have used the term “validation” in distinct ways, and I suspect that this has previously led evaluators of the V&V framework to talk past one another. But while Beisbart’s efforts to clarify things have made some progress, I introduce some finer-grained distinctions that will allow us to survey the question of separability with greater clarity. In Section 1.3, I assess Winsberg’s argument against the possibility of independent validation and Beisbart’s response; in particular, I show that while Beisbart is technically correct with regard to the argument at hand, his response implicitly assumes the possibility of independent verification. In Section 1.4, I expand Winsberg’s argument regarding the impossibility of independent verification. In particular, where Winsberg merely alleges that the mathematical arguments that can be given in support of model verification are weak, I show how and why they are weak. I outline the two possible approaches to verification—deductive and inductive—and show that the tools available within the V&V framework will not in general suffice to carry either of these strategies out. In Section 1.5, I discuss how the observations and arguments in this chapter anticipate the following two chapters.

1.1 Separability—Descriptive vs. Prescriptive

Before beginning my primary investigation, I must clarify two different senses in which the "separability thesis" and the rejection thereof can be read.

Some advocates of the V&V framework make their case on epistemically prescriptive grounds; that is to say, they argue that the separated and sequential process of verification and validation is necessary for a simulation to achieve its intended epistemic goal. Oberkampf and Roy (2010), broadly cited as a canonical textbook on V&V, describes V&V as being of "pivotal importance" to the "credibility of scientific computing" (14), and takes care to detail the sequential ordering of the prescribed activities. Morrison similarly insists that the separation of verification from validation is crucial for either of these methods to be a worthwhile undertaking (Morrison, 2015, 266-9). I will call this the *prescriptive* separability thesis: that the processes of verification and validation *must* be separated if we are to achieve epistemically sound results.

Contrast this with how Winsberg frames his arguments against V&V:

> The reason that climate scientists *cannot* genuinely verify and validate their models separately has to do with many of the features of climate simulation models... (Winsberg, 2018, 157, emphasis added)

In this passage, Winsberg is not directly arguing that the prescriptive separability thesis is wrong, as this would require him to demonstrate that there are epistemically robust alternatives to V&V. Rather, he is arguing on the basis of examples that fully separated V&V is practically impossibles in at least some cases. Likewise, Beisbart's response to Winsberg is primarily concerned with arguing that V&V can be separated both conceptually and in practice.[1] Call this the *descriptive* separability thesis: that the processes of verification and validation *can* be separated and successfully achieved in all real contexts of interest.[2]

Importantly, while we (that is, philosophers) have imported the language of normativity into epistemology, there is one important difference between ethics and epistemology: in epistemology, "ought" does not imply "can". As such, a successful argument against the descriptive thesis does not cut against the prescriptive thesis, nor does the prescriptive thesis imply the descriptive thesis; in principle, Winsberg may be descriptively correct while the proponents of V&V are prescriptively correct. Of course, in this unfortunate case, we would be forced to conclude that simulation methods are epistemically unsound past a certain threshold of complexity.

Ultimately, I reject both the prescriptive and the descriptive separability theses—that is, I argue that certain kinds of highly complex simulations will be impossible to sanction via strict V&V procedures, and I independently believe that simulations can be adequately sanctioned by other, less strict methods. The remainder of this chapter will address the descriptive claim, as this is the framing that concerns Winsberg and Beisbart's arguments. In the following chapters, I will develop my case against the prescriptive claim by detailing alternative methods to those found in the V&V toolkit.

[1] This is despite the fact that Beisbart's article is entitled "*Should* Verification and Validation be Separated?" (emphasis added)

[2] I am not trying to suggest that Winsberg does not also reject the prescriptive separability thesis—here, I am just pointing out that many of his arguments cut against the descriptive thesis only, and that this distinction is important.

1.2 Verification & Validation

In this section, I will first introduce some of the standard terminology from the V&V framework; after this, I will introduce some additional distinctions necessary for my arguments in the following two sections.

I will refer to the object that a simulation is trying to simulate as the *target system*. Following Oberkampf and Roy (2010), I will distinguish between the *conceptual* or *theoretical model*, which refers to the mathematical equations (and attendant modeling assumptions) that represent the target system, and the *computational model*, which refers to the discretized instantiation of the conceptual model on a computer. Because the equations of the conceptual model are not in general analytically tractable, the consequences of these equations must be ascertained by numerical methods. The simulationist's attempt to use a simulation to represent the target system can fail in two distinct ways: the theoretical model could fail to represent the target system, or the computational model could fail to faithfully instantiate the conceptual model (to the requisite degree of accuracy). Formally, *verification* refers to the process of confirming that a computational model faithfully instantiates a given conceptual model, independent of any relations to the target system.[3]

As Beisbart (2019b) shows, the term *validation* has been used quite liberally; in particular, by a thorough analysis of various definitions and usages in the literature he presents the definition

> Validation of a computer simulation is a ________(3b) ________(3a) evaluation of ________(1) following the standards ________(2) with cogency ________(4) and using ________(5). (Beisbart, 2019b, 64)

[3] There are more fine-grained distinctions that can be drawn—e.g., by distinguishing between the abstract computational model and the actual coded instantiation of the computational model, verification can be divided into *code verification* and *solution verification*. However, these will not concern us here, as code verification is primarily a matter of software quality assurance and does not present the same kinds of epistemic quandaries as the question of the abstract computational model.

where each of the lacuna has a number of possible fillers. Some of these lacuna are more consequential than others—in particular, he points out that in lacuna (1) could be filled by either the conceptual or computational model! He adopts the terminology *conceptual model validation* and *computational model validation* to prevent possible confusion with respect to this ambiguity; I will adopt this terminology as well. For the purposes of this chapter, the remaining possible variations on the definition of validation will not be relevant to the discussion.

However, even with these distinctions made, there is still an important ambiguity rooted in the fact that, colloquially, "simulation" may refer either to an overarching simulation code, designed to accept a range of parameters and initial conditions—or to a particular execution of that code with parameters and initial conditions defined. Certainly, to this philosopher's ear, the terms "model" and "system" are both suggestive of an individuated token, which might suggest that these terms should be read narrowly. However, this is *not* the way these terms are generally used by proponents of V&V; with a few ambiguous exceptions, discussion and references to concepts such as the "domain of applicability" indicate that these objects refer to entire classes of possible simulation runs (Oberkampf and Roy, 2010, 22). Following their terminology, I will construe the above-defined terms more precisely (see Figure 1):

- the *target system* is a *class* of possible physical configurations, grouped by the expectation that they are all instances of roughly the same phenomenon governed by roughly the same theoretical background.

For example, a climate model simulation may be designed to represent the target system consisting of many possible initial configurations of energy on an earth-like sphere; a galaxy formation simulation may be designed to represent the target system consisting of many possible initial configurations of primordial gas and dark matter with various parameter weightings to subgrid processes. Each of these initial configurations corresponds to a particular choice of initial conditions and parameters, but crucially these should not be taken to be identical—initial conditions and parameters are elements of either the conceptual model or computational model (depending on whether they are discretized), and merely serve as the theoretical or numerical representations of the possible initial configurations. These initial configurations, if instantiated in reality, will evolve according to whatever laws of physics govern the system.

- the *conceptual model* consists of the actual *equations* that represent our understanding of the shared background theory of the target system, as well as a class of initial conditions and parameters that represent possible initial configurations of the target system.

The *background theory* consists of the well-established general physical theories that we believe to be relevant in the case of the target system—e.g., one believes the theory of gravity will be relevant when modeling the solar system. Of course, these physical theories, while general in application, do not always prescribe precisely how they are to be applied, and therefore additional *modeling assumptions* must usually be made to construct the corresponding equations. Crucially, the conceptual model is more general than a particular choice of initial conditions and parameters, though it may specify a set of mathematical representations to serve as initial conditions and underwrite a mapping from these to the set of initial configurations. Each of these initial conditions, combined with the equations of the conceptual model, maps to a *solution*; while these solutions are not analytically obtainable in general, in principle this solution is the conceptual model's representation of the evolution of the initial configuration corresponding to that initial condition.

- the *computational model* is the chosen discretization of the conceptual model—and just as the conceptual model maps initial conditions to solutions, the computerized model maps discretized initial conditions to numerical solutions.

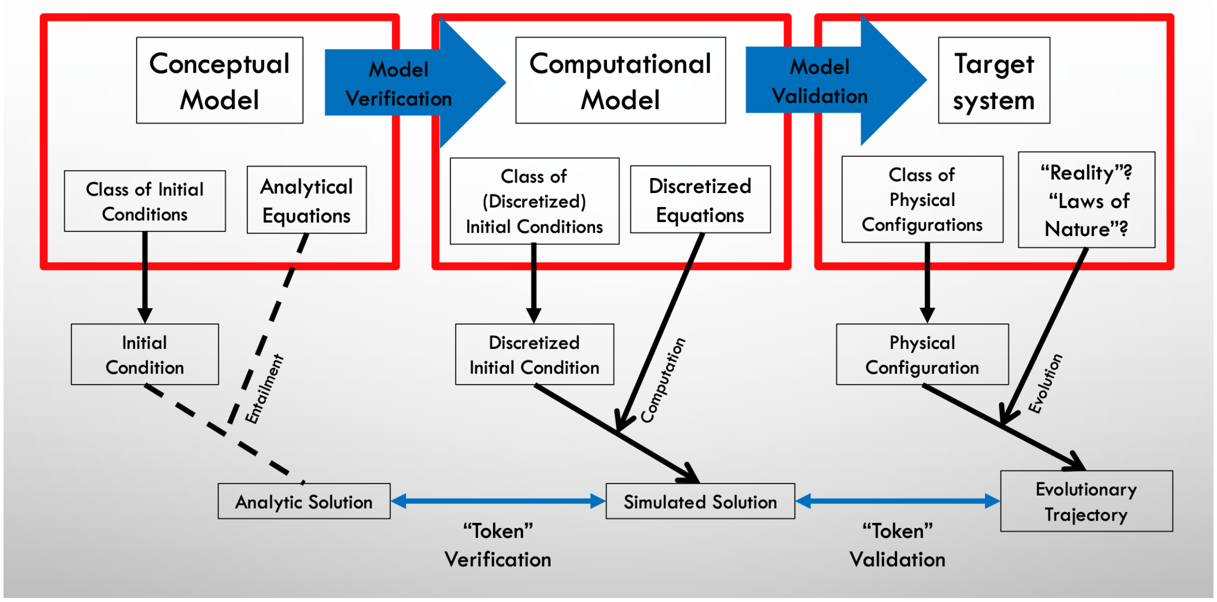

Figure 1.1: An expanded layout of the various components in the V&V framework.

These precisifications have important consequences for the way we think about verification and validation in practical contexts—in some circumstances, one may be confident that an individual token simulation run adequately represents some individual physical system, yet lack confidence in the simulation code as a whole. Thus, in keeping with the above distinction:

- Let *verification* or *model verification* refer to the broader process of confirming that a computational model faithfully instantiates a given conceptual model over the whole range of parameters and initial conditions of interest. This should be distinguished from *token verification*, which I will use to refer to the narrower process of confirming that a computational model faithfully computes the solutions to the equations of the conceptual model with respect to a particular choice of parameters/initial conditions.
- Let *conceptual model validation* and *computational model validation* refer to the broader processes of evaluating the correspondence between the conceptual or computational model and the target system, respectively. Similarly, I will use the terms *conceptual*

> *token validation* to refer to the narrower conception of confirming a correspondence between a particular element of the target system and the solutions to equations of the conceptual model that correspond to that target system's parameters/initial conditions, and *computational token validation* to refer to a correspondence between a particular element of the target system and the output of the computational model under that target system's parameters/initial conditions.

Obviously, verification and validation at the model level are not unrelated to verification and validation at the token level: confidence at the model level is going to warrant confidence at the token level (within some domain of applicability), and, as I will argue shortly, establishing confidence at the model level may need to be inductive on the success of token-verification of many individual simulation runs. However, these distinctions will allow us to get a better grasp on what proponents of V&V could possibly mean when they discuss verification and validation (in my terminology, at the model level), and based on this I will show that there can be scenarios where the separability of V&V breaks down.

1.3 Independent Validation & Simulation Development

In this section, I will assess Winsberg's arguments against the possibility of independent validation, and Beisbart's response.

Winsberg points out that simulations are generally developed not in a linear process, but in a cyclic fashion; on this basis, he argues that any justification of its final product entangles the justification of its parts with one another in a way that is not cleanly separable. Just as the final conceptual model has been influenced by previous discretizations, the final computational model has been influenced by previous attempts to validate the conceptual model; hence, there can be no independent conceptual model validation (Winsberg, 2018, 158).

Beisbart, in turn, responds by arguing that the cyclic development of a simulation is immaterial to the question of whether verification and validation are separable; the simulationist may tweak the conceptual or computational model in the process of trying to craft an optimal simulation, but as long as verification and validation are separable within a single cycle, this fact will not cut against the separability of V&V (Beisbart, 2019b, 1023). Indeed, Beisbart points out that Oberkampf & Roy are perfectly fine with iterative simulation development in practice (Oberkampf and Roy, 2010, 60)

Given only Winsberg's description, it must be admitted that Beisbart is correct; certainly, if we grant his stipulation that at each stage in the cycle we can re-verify and re-validate the models from scratch, it is arguable that the influence on the final choice of model by previous iterations of the cycle is irrelevant. For all the severity of the V&V account, the prescriptive separability thesis concerns only justification—it has little to say about any context of discovery. As such, if the V&V framework has the tools to independently verify a given simulation model without reference to its developmental history, then the fact that some simulations are developed cyclically cannot be said to prove that independent validation is impossible.

However, this line of argument relies on that prior assumption—namely, that the V&V framework has the tools to independently verify a given simulation model. Beisbart suggests that Winsberg has an overly conservative view of what kinds of techniques count as purely mathematical verification methods, and points out that Winsberg has not addressed the Method of Manufactured Solutions (Beisbart, 2019b, 1026, fn. 18). As noted, Winsberg rejects this assumption, but does not provide much elaboration as to *why* the mathematical arguments that can be given for independent verification are weak, and in particular does not address the Method of Manufactured Solutions. In the next section, I will address those arguments to show how and why they are weak.

Thus, Winsberg's observations about the cyclic development of simulation should not be taken as a *proof* of the inseparability of verification and validation—they should be seen as a *consequence* of the inseparability of verification and validation, and suggestive of how simulations can be justified even when they do not abide by the strict V&V separation prescription. But given that the success or failure of this picture seems to turn on whether one presumes that independent verification is possible in the first place, I will next turn to an evaluation of the methods suggested by Beisbart and others.

1.4 Verification & the Method of Manufactured Solutions

Winsberg notes that the mathematical arguments that can be given in support of independent model verification are generally quite weak, and on this basis he concludes that independent model verification is impossible for sufficiently complex simulations (Winsberg, 2010, 159). I agree with Winsberg's conclusion here, but to meet the above challenge we need to examine why these arguments are weak and what it means for a simulation to be "sufficiently complex."

Because model verification addresses the relationship between the conceptual and computational models, the claim that a simulation has been model-verified is a general claim about the numerical reliability of that simulation over a range of possible parameter/initial condition choices. But as the accuracy of a numerical implementation of a given system of differential equations may differ with different parameter/initial condition choices, this poses a problem—we need to build confidence in the computational model over this whole range, but in typical contexts we cannot check its performance with respect to each individual token. If we confine ourselves to purely mathematical methods—i.e., if we abide by the separability principle and refrain from any validation activities—we can approach this problem with either a deductive or inductive strategy for error estimation. Unfortunately, both of these strategies are untenable in the case of highly complex simulations.

First, consider a deductive strategy. We might attempt to prove the adequacy of a computational model by showing that, for any parameter/initial condition choice within the domain of interest, the numerical errors are bounded within tolerance levels—or, perhaps with some additional qualifiers, can be made to fall within tolerance levels. These kinds of methods are analogous to a proof of a theorem, and thus would be powerful means of verification if they could be consistently applied. However, as Oberkampf & Roy note, these kinds of *a priori* methods for error estimation are generally inapplicable to all but the simplest of problems—and even for those simple cases where they are technically possible, the error bounds that can be achieved are often not within the requisite tolerance levels (Oberkampf and Roy, 2010, 297).

Consider instead an inductive strategy. Obviously, if we approach the process of model verification from below—i.e., by token verification—we cannot individually verify all tokens in the desired range. But as with any inductive exercise, we can nonetheless attempt to build confidence that our simulation is model-verified by token-verifying an increasing numbers of tokens. The practice of benchmarking simulations against known exact solutions is an exercise of precisely this kind; where exact solutions for simple initial conditions exist, we can confirm whether or not a particular execution of the computational model is faithful to the corresponding element of the conceptual model by a straightforward comparison.

Of course, for an inductive inference to be well-supported, it is not enough to simply accumulate verified tokens—the subset of tested tokens needs to be *representative* of the broader population about which the general claim is made. And while benchmarking against exact solutions may provide a bit of grist for the inductive mill, in highly complex simulations the subset of initial conditions simple enough to have analytically tractable solutions is generally small and highly unrepresentative of the larger class of interest (Winsberg, 2010, 21-4)—after all, simulation methods are generally needed *precisely because* the equations of interest are not analytically tractable in most scenarios.

The crucial question, then, is whether additional techniques will suffice to make up the difference. Morrison (2015) and Jebeile and Ardourel (2019) suggest that such techniques exist, but only the latter provides a particular example. I will first address the Method of Manufactured Solutions, as cited by Beisbart (2019a) and Jebeile and Ardourel (2019) and described by Roache (2019), as this technique seems to come closest to achieving its goal. After this, I will make some general comments on the collection of other techniques, as suggested by Morrison.

The Method of Manufactured Solutions The Method of Manufactured Solutions (MMS) is a technique for generating exact solutions to a set of differential equations that are related to our original differential equations of interest (Roache, 2019). Suppose, for instance, that we have a differential equation

$$L(u) \equiv \frac{\partial u}{\partial x}\frac{\partial u}{\partial y} - \frac{\partial^2 u}{\partial x^2} = 0,$$

and that we would like to know the accuracy of a computational model of this differential equation. We can choose an arbitrary differentiable test function of the relevant independent variables, $T(x, y)$, and use this to generate a function $Q(x, y) = L(T(x, y))$. For example, if we choose $T(x, y) = x^2 + y^2$, then

$$Q(x, y) = L(T(x, y)) = \frac{\partial T}{\partial x}\frac{\partial T}{\partial y} - \frac{\partial^2 T}{\partial x^2} = 4xy - 2.$$

On this basis, we can construct a new differential equation, $L(u) = Q(x, y)$, where $Q(x, y)$ acts as a general source term in our original differential equation. In our example, $L(u) =$

$Q(x, y)$ yields

$$\frac{\partial u}{\partial x}\frac{\partial u}{\partial y} - \frac{\partial^2 u}{\partial x^2} = 4xy - 2.$$

The clever trick here is to notice that our test function $T(x, y)$ is (by construction!) an exact solution to this new differentiable equation—and, thus, we can perform an error analysis of $L(u) = Q(x, y)$ using $T(x, y)$ as a benchmark. Moreover, the error associated with our computation of the solution $T(x, y)$ will be generated by a combined discretization of $Q(x, y)$ and our original $L(u)$, which means that we can (at least in principle) infer information about the portion of error caused by $L(u)$. In general, the process requires one to perform an order-of-convergence analysis to confirm the error associated with our stipulated solution $T(x, y)$ decreases with increasing resolution. Specifically, if T^{Δ} is a discrete solution associated with discretization step size Δ that we are using to approximate the exact solution T^{exact}, then it will be said to have order of convergence p just in case the error $E_{L(u)=Q}$ goes as

$$E_{L(u)=Q} = T^{\Delta} - T^{\text{exact}} = C\Delta^{p} + \text{higher order terms}$$

for some constant C. This would seem to nicely solve the simulationist's problem—if we can just pick arbitrary functions to be solutions, we should be able to sample the space of solutions quite easily!

Problems with MMS However, despite the novelty of MMS, it nonetheless suffers from an important weaknesses that prevents it from being a general solution to our problem.

In fact, the name "Method of Manufactured Solutions" is something of a misnomer. In the above-described process, the solution is not manufactured but stipulated; the real work of manufacturing involves constructing an alternative set of *equations* to fit the stipulated solution.[4] As such, it is not always clear that showing $E_{L(u)=Q} \to 0$ as $\Delta \to 0$ will allow us to infer that the error $E_{L(u)=0}$ associated with a solution $U(x,y)$ to the original equation of interest $L(u) = 0$ will similarly converge; even if we allow the highly questionable assumption that the total error can be factored into contributions from the original $L(u) = 0$ term and the source term Q,

$$E_{L(u)=Q} = E_{L(u)=0} + E_Q,$$

this will only allow us to set bounds on $E_{L(u)=0}$ if E_Q is itself estimable, as

$$|E_{L(u)}| \leq |E_{L(u)=Q}| + |E_Q|.$$

Under some circumstances, this will not itself pose a problem—the MMS equations are not totally unrelated to the original equation of interest, and thus we would expect that a suitable array of test solutions would allow us to test various aspects of the original equation in tangential ways. However, the procedure outlined above requires a smooth input $T(x,y)$—else, the $Q(x,y)$ generated will be non-smooth. Moreover, if $Q(x,y)$ is to be a nice smooth function, the original differential operator $L(u)$ must be composed of nice smooth operators—and this makes it difficult to represent physical processes that have a sharp cutoffs, as will often be the case when simulationists need to represent subgrid processes. Indeed, it is generally understood that MMS has practical limitations in this regard, and thus Knupp (2002, 44-5) suggests a number of principle guidelines to the choice of manufactured solution, including:

(1) Solutions should be "sufficiently smooth on the problem domain";

(4) Solution derivatives "should be bounded above by a small constant";

[4] A more appropriate name might be "Prescribed Solution Forcing Method" (Dee, 1991), but "Method of Manufactured Solutions" seems to have stuck.

(6) Solutions should be "composed of simple analytic functions"—polynomials, trigonometric functions, exponential functions, etc.

Crucially, if MMS cannot generally handle solutions with discontinuous behaviors, this means it cannot deliver us a full sampling of the space of solutions whenever we are interested in simulations of phenomena with shockwaves, cutoffs, or other abrupt features such as discontinuous subgrid feedback. As simulations are often necessary precisely because the phenomena of interest have these kinds of features, this limitation is quite significant.

Generalizations of MMS Roache (2019, 2009) acknowledges that the above-described method cannot generally handle these kinds of discontinuous phenomena on their face, but nonetheless claims that MMS can be adapted to treat shockwaves and other discontinuous solutions with supplemental considerations. As my central argument above is to suggest that MMS cannot currently handle these kinds of situations, I will briefly comment on these. Of the citations provided, most do not amount to a general method—e.g., the work of J.M. Powers and coauthors (Powers and Stewart, 1992; Powers and Gonthier, 1992; Grismer and Powers, 1996) examine a number of oblique detonation scenarios. This work, while interesting, amounts to showing that some verification techniques can be applied to shockwaves under isolated and highly simplified conditions; we cannot regard this as a general solution to the above-described problem.

Of much greater interest is the attempt to give a generalized MMS method, as found in Woods and Starkey (2015). Woods & Starkey reframe the above-described differential procedure in integral terms; instead of simply taking the original differential equation $L(u) = 0$ and generating the source term by means of $Q = L(T)$, one can rewrite $L(u)$ as an integral equation

$$\int_V L(u) = 0.$$

In particular, by cleverly writing $L(u)$ in a general form consisting of derivatives of conservative flux terms $F_\mu(u)$ and a source term $S(u)$, one can use Stokes' theorem to recast the volume integrals of the flux derivatives as surface integrals,

$$\sum_\mu \left(\int_V \frac{\partial F_\mu(u)}{\partial x^\mu} \right) + \int_V S(u) = 0 \Rightarrow \sum_\mu \left(\oint_{\partial V} F_\mu(u) \right) + \int_V S(u) = 0.$$

Once the equations of interest are recast in this form, a test function T—this time, allowing for at least some discontinuities—can be applied to generate a source term Q as

$$\sum_\mu \left(\oint_{\partial V} F_\mu(T) \right) + \int_V S(T) = \int_V Q,$$

in particular, by choosing the integration domains in such a way as to track the discontinuities in T. But while this method is certainly an interesting extension of the above differential MMS procedure, a number of new problems are implicit in its formulation.

First, the generation of the source term requires the integration of the various flux and source terms from the original equation of interest—and unless the original equation contains terms that can be integrated analytically, these integrations will need to be performed numerically. Thus, in many cases, T will not be an *exact* solution to $L(u) = Q$ as generated by this method. as was the appeal for the differential MMS method.

Second, the result of integrating over the flux and source terms is not itself the bare source Q, but rather various volume integrals of Q. As the integration domains are chosen to be identical to the computational cells, this is not a problem on its face; the process of generating a numerical solution for comparison against T would require this kind of discretization of Q in any case. However, this does mean that the discretization grid for generating a numerical solution is locked-in by the particular choices used to generate Q; that is, unlike the normal case of numerical solution generation from scratch, a discretization grid that traces out the discontinuities in the correct solution must be known ahead of time and employed.

Finally, this constraint on the discretization grid is potentially going to require computationally intractable grid discretizations, at least once one moves beyond the simplest shocks and discontinuities. Certainly, the above described method seems to work naturally if one chooses the discontinuities in T to have simple linear or quasilinear boundaries—especially if these can be read off of T's form. However, if one wants to test the capacities of a simulation to handle more subtle phenomena such as fluid-mixing, this would require a grid discretization that tracks highly chaotic and turbulent boundaries to incredibly high degrees of precision.[5]

Notably, while Woods & Starkey suggest that their method allows them to relax Knupp (2002)'s conditions (1) and (4), they leave intact condition (6)—suggesting that the kinds of test solutions they have in mind are still piecewise solutions limited to representing the simplest of shocks and discontinuities. Other attempts to extend and generalize MMS run into similar problems with the discretization grid (Grier et al., 2015). As such, it may be premature to suggest that MMS can be fully generalized.

Other verification methods From a practical perspective, these problems are sufficient to show that the MMS cannot provide a general assurance that any given simulation can be model-verified using the inductive approach. Importantly, the MMS is not an outlier in this regard—other techniques within the V&V toolkit suffer from identical or analogous limitations. Other methods for exact solution "generation"—such as the Method of Nearby Problems, which can only be applied to a solution amenable to a reasonable spline or curve fit (Oberkampf and Roy, 2010, 236)—are similarly limited to smooth solutions. In general, discretization error estimators require the solution to be in the *asymptotic range*, which is quite difficult to show in cases of nonlinear, hyperbolic, coupled systems of equations and generally requires uniform grid spacing (Oberkampf and Roy, 2010, 317-8).

[5] This is all assuming that one could generate a sufficiently complex T to begin with!

Thus, while these tools are certainly more sophisticated than those we saw above in the deductive approach, they are nonetheless insufficient to establish the general viability of token-verification in cases where the simulation discretization is not over a uniform grid or where the solution is expected to be nonsmooth. This, in turn, implies that these methods cannot reliably deliver a representative sampling of token-verifications in many simulations of interest—and thus, the mathematical arguments that can be given for independent verification are quite weak.

Before moving on to the final section, three remarks are in order.

First, while I have argued that MMS and other purely numerical verification methods are not adequate in many contexts of interest, I would emphasize that this does not mean they are useless. Indeed, while the solutions generated by MMS are not, strictly speaking, solutions to the original differential equations of interest, they may nonetheless be useful tools—and while these do not include solutions with significant discontinuities, they are certainly more expansive than the simple analytically tractable solutions. In principle, one could even disregard the above caveats and force through unsmooth and discontinuous $T(x, y)$ solutions, *if* one is willing to accept the high probability that the MMS will falsely flag many accurate computational models as error-ridden. However, given that this cuts against both the known practical limitations of MMS and the arguments presented above, I suggest that the onus is on proponents of the V&V framework to demonstrate that these techniques are practically viable in highly complex cases.

Second, while the methods described above are not themselves adequate general methods for inductively achieving independent verification (at least in many cases of interest), my argument should not be construed as a conclusive no-go proof that such methods cannot exist. The application of these methods (or some suitable adaptation thereof) to cases of nonsmooth solutions or simulations with nonuniform mesh patterns is an open research problem, and as I noted with respect to the MMS, in some niche cases their proponents have made progress. To this extent, my argument against the descriptive separability thesis is defeasible—*if* it could be shown that some practically feasible method could reliably be expected to assess nonuniform mesh simulations in nonsmooth regimes.

Third, in consideration of some arguments put forward by Jebeile and Ardourel (2019), I must qualify the limits of this defeasibility. In particular, I will focus on two of their arguments: a case study on the use of formal methods in verification, and a survey of simulationists regarding their use of formal methods. Their case study, in which a declarative modeling language Alloy was used to check the consistency of a triangle mesh (11-12), falls short of showing that formal methods can be used to verify a *solution* to some discretized differential equation as solved over this mesh. In particular, simply checking that the mesh is consistent—e.g., that it does not have any crossed edges—does not amount to checking that the discretized differential equations solved on that mesh are accurate. That is, merely showing that formal methods are involved in verification does not suffice to show that formal methods themselves are capable of verification independently.

Likewise, the cited survey simply relies on self-reported use of formal methods by 21.1% of $n = 283$ individuals in the modeling and simulation community, but this does not distinguish true formal methods of verification (such as MMS) from auxiliary tools (such as Alloy). Moreover, even if we assume they are all instances of the former, merely showing that formal methods have been used in some fields and that they have seen increasing use over the years does not warrant an assumption that "formal methods are going to broaden their scope of application" (10-11, 13-14)—none of the critics of V&V deny that these methods exist, or that they can be used effectively under certain circumstances. In the absence of proof-positive of some reliable general method for implementing independent verification, an argument would need to address the core problems that critics have shown to be limiting factors on the efficacy of methods that currently exist.

1.5 Conclusion

In this chapter, I have examined the technical details of various methods that might be deployed to independently verify a simulation, and I have shown that they will insufficient in many cases of interest. Because of this insufficiency, we should reject the descriptive separability thesis.

In reaching this conclusion, I have drawn a number of distinctions—and I have suggested that a failure to draw these distinctions has been a source of some confusion in the literature. First, I distinguished between the prescriptive and descriptive separability theses, and I contended that they must be argued for (or argued against) independently of one another. Second, I drew a distinction between model-level verification and validation, and token-level verification and validation; by using this distinction to clarify what the possible abstract strategies for verification were, I set up a standard to judge the success or failure of the various methods in the V&V toolkit.

Though I have limited my arguments in this chapter to a critique of the descriptive separability thesis, this model/token distinction is useful in other ways. In particular, when we consider Winsberg's observations about the cyclic development of simulations (Winsberg, 2018, 158-9) in light of this distinction, it becomes much easier to see how the processes of verification and validation are interwoven in practice. In the early iterations of the cycle, even a failure to model-verify the whole simulation code does not imply that no tokens were successfully token-verified—nor does a corresponding failure to computational-model-validate the code imply that no tokens were computational-token-validated. Rather, the successes and failures of the simulation at the token level can guide further development cycles, and later iterations of the simulation will draw on these considerations. These, in turn, will be a basis for justifying highly complex simulations when they cannot be verified and validated in the clean, separable way prescribed by the V&V framework. In Chapter 2, I develop a case study of this process; in Chapter 3, I give a more abstract characterization of this process in terms of the adequacy-for-purpose framework.

Chapter 2

Simulation Verification in Practice

Winsberg has argued that the prescription for strict separation between V&V is not followed—and indeed *cannot* be followed—as a matter of actual practice in cases of highly complex simulations (Winsberg, 2010, 2018). In the previous chapter, I defended the latter claim by a more detailed analysis of specific V&V methods; in this chapter, I will present further evidence showing that the prescription goes largely unheeded in the context of astrophysical magnetohydrodynamics (MHD) simulations. But as I noted in Section 1.1, even if I have successfully shown that simulationists *cannot* strictly separate these activities, we still must contend with the possibility that this has fatal epistemic consequences for simulation methods—after all, this strict separation is generally prescribed as a bulwark against an allegedly severe and systematic epistemic risk. In other words, it remains to be shown that the prescriptive separability claim is incorrect—that despite the fact that they do not follow strict V&V prescription, the methods that simulationists *do* use can mitigate this risk. In what follows, I will argue that a careful examination of the development of simulation codes and verification tests allows us to develop just such an alternative account.

In section 2.1, I present the survey of a range of representative MHD simulation codes and the various tests that were proffered in the literature to support and characterize them. In section 2.2, I examine a particular class of tests associated with the phenomenon of fluid-mixing instabilities, the circumstances under which this phenomenon became a concerning source of error, and the simulationists' response to these developments; on the basis of these and other considerations, I will argue that this approach to complex simulation verification is more exploratory and piecemeal than philosophers have supposed. In section 2.3, I examine some of the details of the purpose and implementation of these tests, and I argue that the mathematical and physical aspects of complex simulation evaluation cannot be neatly disentangled—and, in some cases, *should* not be disentangled.

2.1 A Survey of Galaxy Formation Simulation Codes

The primary codes examined for the present survey were Flash (Fryxell et al., 2000), Ramses (Teyssier, 2002), Gadget-2 (Springel, 2005), Athena (Stone et al., 2008), Arepo (Springel, 2010), and Gizmo (Hopkins, 2015). These simulations were chosen to span a range of years and MHD code types, focusing on simulations that were particularly influential and that had a substantive literature. Athena, for instance, uses a static grid-based Eulerian method; Flash and Ramses are also stationary grid-based methods, but use Adaptive Mesh Refinement (AMR) to refine the grid in places. Gadget-2 is a particular implementation of Smooth Particle Hydrodynamics (SPH), a Lagrangian method. Arepo combines elements of the AMR and SPH methods to create a "moving-mesh" code which allows for tessellation without stationary grid boundaries. Gizmo is similar to Arepo in that it combines advantages of the SPH and AMR methods, but it is roughly described as "meshless", as it involves a kind of tessellation akin to Arepo, but allows for a smoothing and blurring of the boundaries according to a kernel function.[1]

While some of the official public release versions of these codes included routines for tests not reported in the literature, the survey generally only looked to tests that were reported in published papers. This was for three reasons. First, I am primarily interested in tests that were considered important enough to be on display and described in some detail in the method papers presenting the code. Second, I am also interested in the analysis of the code's performance on particular tests; simply including a routine in the code suite does not indicate the significance of the test vis-à-vis particular kinds of error or whether the result of the routine measured up to some standard. Third, particular routines may have been included in either the initial or subsequent versions of the code; the papers, being timestamped, provide a better gauge of when tests were performed (or at least considered important enough to publish).

The two exceptions to this are Flash and Athena. Flash includes a bare minimum of tests in its initial release paper but provides many more tests and has an extensive amount of useful documentation in the User Guide (Flash User Guide). This user guide is also available in various editions corresponding to different release versions of Flash, spanning version 1.0 from October 1999 to the most recent version 4.6.2 in October 2019; this allows us to track when the various test problems were introduced. A brief overview of this sequence will be discussed below as well. Athena includes a few additional fluid-mixing instability tests on a (now partially-defunct) webpage, and given my focus on these tests in section 2, I have chosen to include them as well. Given that at least one fluid-mixing test was included in the methods paper (the Rayleigh-Taylor instability test), and given the timeline to be described

[1] Technically, Gizmo is able to facilitate a number of sub-methods, including "traditional" SPH. The new methods of interest here are the Meshless Finite-Volume and Meshless Finite-Mass described in Hopkins (2015).

	FLASH	RAMSES	GADGET-2	ATHENA	AREPO	GIZMO
	Fryxell et al	Teyssier	Springel	Stone et al	Springel	Hopkins
	2000	2002	2005	2008	2010	2015
One-dimensional wave[a]				✓	✓	✓
Sod shocktube[b]	✓	✓	✓	✓	✓	✓
Interacting blast waves[c]	✓			✓	✓	✓
Sedov-Taylor point explosion[d]	✓	✓			✓	✓
Noh problem[e]				✓	✓	✓
Gresho vortex[f]					✓	✓
Driven turbulence					✓[n]	✓
Keplerian disks					✓[o]	✓
Kelvin-Helmholtz				✓[p]	✓[q]	✓[r]
Rayleigh-Taylor[g]				✓	✓	✓*[s]
"Blob" test[h]						✓
"Square" test[i]						✓
Implosion[g]				✓		
Shu & Osher shocktube[j]				✓		
Forced AMR jump	✓	✓[t]				
Advection problem	✓					
Wind tunnel with step[k]	✓					
Strong shock[l]			✓			
Double Mach reflection[c]				✓		
Einfeldt strong rarefaction[m]				✓		
Moving boundary					✓	

[a] Stone et al. (2008) [b] Sod (1978) [c] Woodward and Colella (1984) [d] Sedov (1959) [e] Noh (1987) [f] Gresho and Chan (1990) [g] Liska and Wendroff (2003) [h] Agertz et al. (2007) [i] Heß and Springel (2010) [j] Shu and Osher (1989) [k] Emery (1968) [l] Klein et al. (1994) [m] Einfeldt et al. (1991) [n] Bauer and Springel (2012) [o] Pakmor et al. (2016) [p] ath (2008) [q] Robertson et al. (2010) [r] McNally et al. (2012) [s] Abel (2011) [t] Khokhlov (1998)

Table 2.1: Hydrodynamics tests. Unless otherwise indicated, the test results as run by a particular code is recorded in the paper indicated at the top of each respective column column. The * citation indicates that a different test setup was cited.

in the next section, it is likely that the other fluid-mixing tests were performed around that time.

An overview of the various tests found in the initial documentation papers can be found in Table 2.1 (hydrodynamic tests), Table 2.2 (magnetohydrodynamics tests), and Table 2.3 (self-gravity tests) (FLASH is omitted from Table 2; for an overview of those MHD tests that were eventually included, see Table 2.4). Table 2.4 tracks the inclusion of tests over time in selected editions of the FLASH user guide. Based on the data laid out in the various tables, we can make a number of preliminary observations, some of which I will expand on in later sections.

Among those tests that are common to multiple codes, it is clear that there is a general accumulation of hydrodynamics tests as time progresses, with later-developed codes including far more tests than earlier codes. In many cases, the later codes will cite to examples of the test as implemented in earlier codes, both among those surveyed here and elsewhere. While tests are not all consistent, where possible I have cited both the original paper that described or designed the test and indicated where authors used variants. As I will discuss in the next section, in some cases the appearance of a new test is a clear response to reported concerns about a particular source of error, especially where that source of error was a problem in prior codes and not particularly well-tracked by previously cited tests. In other circumstances, the overarching purpose for adding a new test is unclear—i.e., it may or may not be redundant with respect to the rest of the collection. This accumulation is also apparent in the history of the Flash simulation, where many of the tests added in the two decades since its initial release overlap with the other surveyed codes and several even track with the times that they were introduced.

Where tests are not common among codes, they can roughly be divided into two categories. Some tests are unique to a particular code because they are generally inapplicable to other code types, which is to say they are tailored to test for numerical errors to which other code types are not susceptible. For example, Flash and Ramses both include unique tests of circumstances where the adaptive mesh refinement algorithm is forced to make sharp jumps in spatial resolution—these tests are obviously not applicable in the absence of AMR.

Other tests are not tailored in this manner, although this does not mean that they all serve disparate purposes; in some cases, different tests are probing the same kinds of phenomena, even while the setups and initial conditions are different. This is particularly unsurprising in the case of the myriad unique tests with full self-gravity, as there are few examples of problems

	Ramses	Gadget-2	Athena	Arepo	Gizmo
	Fromang 2006	Dolag 2009	Stone et al 2008	Pakmor et al 2011	Hopkins & Raives 2016
MHD waves[a]			✓		✓
MHD shocktube[b]	✓*[i]	✓	✓	✓*[j]	✓†[k]
Orszag-Tang vortex[c]	✓	✓	✓	✓	✓
MHD rotor[d]		✓	✓		
Current sheet[e,f]	✓		✓[l]		✓
Loop advection[e]	✓		✓	✓	✓
Blast wave[d,g]		✓	✓	✓	✓
Magneto-rotational instabilities	✓				✓[m]
Kelvin-Helmholtz instability			✓[n]		✓
Rayleigh-Taylor instability			✓		✓
Circularly polarized Alfven waves[h]			✓		

[a]Stone et al. (2008) [b]Brio and Wu (1988) [c]Orszag and Tang (1979) [d]Balsara and Spicer (1999) [e]Gardiner and Stone (2005) [f]Hawley and Stone (1995) [g]Londrillo and Del Zanna (2000) [h]Tóth (2000) [i]Torrilhon (2003) [j]Keppens (2004) [k]Tóth (2000) [l]Beckwith and Stone (2011) [m]Guan and Gammie (2008) [n]ath (2008)

Table 2.2: Magnetohydrodynamics tests. As in Table 2.1, unless otherwise specified, the test results as run by a particular code is recorded in the paper indicated at the top of each respective column column. Each test is based on the setup given in the paper cited in the first column, with the exception of the MHD shocktube category: for those marked with *, the cited test was performed *instead*; for those marked with †, the cited test was performed *in addition.*

with self-gravity where analytic solutions exist. Here, the broad aim is to simulate scenarios that are more "realistic" than the other highly simplified tests (albeit still fairly simple!), and consequently in these cases there is less emphasis placed on measuring the code's performance against a straightforward rigorous quantitative standards such as analytic solutions. Further examination of multi-group code-comparison projects also shows that these projects are not always a straightforward exercise, often requiring a great deal of technical elaboration before comparisons can be drawn—and moreover, the various desiderata for these kinds of cross-code comparisons are often in tension with one another (Gueguen, 2021). The fact that these tests are used in a more qualitative way likely accounts for the fact that they do not display the same pattern of accumulation evident among the simpler hydrodynamics tests.

There are also some tests that are *prima facie* relevant to other codes, at least on the basis of the description provided—e.g., both Athena and Gizmo deploy a selective application of

	Flash	Ramses	Gadget-2	Athena	Arepo	Gizmo
	Fryxell et al	Teyssier	Springel	Stone et al	Springel	Hopkins
	2000	2002	2005	2008	2010	2015
Zeldovich pancake[a]		✓			✓	✓
Santa Barbara cluster[b]			✓		✓	✓
Evrard collapse[c]			✓		✓	✓
Simple acceleration		✓				
ΛCDM acceleration		✓				
Spherical infall[d]		✓				
Isothermal collapse[e]			✓			
DM Clustering[f]			✓			
Galaxy collision					✓	
Galaxy disks						✓

[a] Zel'Dovich (1970) [b] Frenk et al. (1999) [c] Evrard (1988) [d] Bertschinger (1985) [e] Burkert and Bodenheimer (1993) [f] Heitmann et al. (2005)

Table 2.3: Self-gravity tests.

two Riemann solvers, including one (the Roe solver) that can give unphysical results if applied incorrectly, but only Athena presents the Einfeldt strong rarefaction test to establish that this will not cause a problem. This may simply be an indication that the problem is no longer of particular concern, or that the Roe solver was tested in Gizmo but the test was not considered important enough to be included in the methods paper.

Additionally, some tests that are common among the various codes are nonetheless used for purposes that do not entirely overlap between codes. The most clear example of this is the distinct use of some common tests by stationary grid codes to test for artificial symmetry breaking along grid axes—e.g., the various shocktubes and blast waves are used in SPH and non-stationary grid codes to test their abilities to handle shocks and contact discontinuities, but in stationary grid codes they can be run both aligned and inclined to the static grid to test for artificial symmetry breaking along grid lines.

The magnetohydrodynamics tests do not display as clear a pattern of accumulation; unlike the hydrodynamics tests, there seems to be a common core of tests that have been more-or-less consistent over the span of years, with the notable exception of the debut of the MHD Kelvin-Helmholtz and Rayleigh-Taylor instability tests. I speculate that the consistency

	1.0 (1999)	2.0 (2002)	2.5 (2005)	3.3 (2010)	4.6.2 (2019)
Sod shocktube[a]	✓	✓	✓	✓*	✓*
Shu & Osher shocktube[b]		✓	✓		✓
Interacting blast waves[c]	✓	✓	✓	✓	✓
Point explosion[d]	✓	✓	✓	✓	✓
Advection problem	✓	✓	✓		
Isentropic vortex[e]			✓	✓	✓
Noh problem					✓
Wind Tunnel with step	✓	✓	✓	✓	✓
Driven turbulence				✓	✓
Relativistic Sod shocktube				✓	✓
Implosion test				✓	✓
Kelvin-Helmholtz					✓
Brio & Wu shocktube[f]		✓	✓	✓	✓
Orszag-Tang vortex[g]			✓	✓	✓
MHD rotor[h]				✓	✓
Current sheet[i]				✓	✓
Field loop advection[i]				✓	✓
Jeans instability[j]		✓	✓	✓	✓
Homologous dust collapse[k]		✓	✓	✓	✓
Huang-Greengard Poisson test[l]		✓	✓	✓	✓
Maclaurin test[m]				✓	✓
Zeldovich pancake			✓	✓	✓

[a] Sod (1978) [b] Shu and Osher (1989) [c] Woodward and Colella (1984) [d] Sedov (1959) [e] Yee et al. (2000)
[f] Brio and Wu (1988) [g] Orszag and Tang (1979) [h] Balsara and Spicer (1999) [i] Gardiner and Stone (2005)
[j] Jeans (1902) [k] Colgate and White (1966) [l] Huang and Greengard (1999) [m] MacLaurin (1801)

Table 2.4: Common magnetohydrodynamics tests.

apparent in magnetohydrodynamics tests is a function in part of the influence of J. Stone, who (with coauthors) proposed a systematic suite of MHD test problems as far back as 1992 (Stone et al., 1992) and, together with T. Gardiner, wrote the 2005 paper (Gardiner and Stone, 2005) that is either directly or indirectly (through the Athena method paper (Stone et al., 2008)) cited by all the MHD method papers in question.

Stone et al. (1992) is notable for being a standalone suite of MHD test problems without being connected to a particular code—in particular, this suite is not intended as a comprehensive collection of all known test problems, but rather as a minimal subset of essential tests, each corresponding to a different MHD phenomenon. As the field has progressed significantly since this suite was published, there is reason to believe that the specifics of this paper are out of date with respect to the surveyed code examples and the phenomena of interest. However, insofar as it lays out a rationale, not only for each specific test, but also for the choice the collection of tests as a whole, the paper provides a framework for thinking about how these tests might be understood to collectively underwrite simulations. In particular, while we may not be able to think of this framework as providing absolute sufficiency conditions for the adequacy of a given suite of test problems, this approach may still point us towards a more pragmatic notion of sufficiency, especially with respect to the current state of knowledge in the field. Admittedly, I have been unable to find similarly systematic proposals for test suites of hydrodynamic or self-gravity test problems; however, in anticipation of the argument that I will be making in section 4, I will note that this emphasis on MHD *phenomena* as the guiding principle for test selection suggests an approach to these tests that goes beyond merely numerical considerations.

2.2 Fluid-Mixing Instabilities and Test Development

The V&V framework originated in a number of subfields within the sciences—including computational fluid dynamics, which has some obvious theoretical overlap with the field of astrophysical magnetohydrodynamics (Oberkampf and Roy, 2010). Despite this, with one exception (Calder et al., 2002), the V&V framework is not generally invoked in the field of astrophysical MHD simulations. Nonetheless, I will briefly outline the rationale for the prescriptive separability thesis, which I will then contrast with an examination of the tests as they are found in the above survey.

In our above terminology, simulation is said to be *model-verified* when we are confident that the numerical methods employed in the simulation faithfully approximate the analytical equations that we intend to model; the simulation is said to be *computational model-validated* when the output of the simulation adequately corresponds to the intended phenomenon in the world. Together, these two components form a bridge between the phenomenon in the world and the analytical equations that constitute our attempts to theoretically capture that phenomenon, via the intermediary of the simulation code. Within this framework, the function of verification tests is to determine whether the numerically-implemented code is faithful to the analytical equations of the original model.

The epistemic challenge associated with this task stems from the two-part structure of V&V; in particular, the concern is that numerical errors could "cancel out" errors caused by an inaccurate model, leading to a simulation built on incorrect theory that nonetheless produces an output that corresponds to the phenomenon in question. This epistemic concern underpins the prescription for the sequential ordering for these activities: first model-verification, then model-validation. If the simulationist ensures that the simulation code is free of numerical errors *independently* of any comparisons to the phenomenon, then this should preempt any risk that we might accidentally fall prey to a cancellation of errors (Morrison, 2015, 265); I will refer to this conception of simulation verification as the "strict V&V account."

With this framework in mind, one might then believe that the survey in Section 2.1 raises some serious concerns. As noted, there has been a tendency for later-developed codes to include more tests than earlier-developed codes—this, in turn, would imply either that the new tests are superfluous, or that the old simulations were not adequately model-verified against certain kinds of numerical errors. The former possibility is unlikely, especially where newer tests show that new codes display marked improvement over the performance of prior codes. Thus, it would seem that earlier codes were not sufficiently verified. Moreover, absent some assurances that newer codes have remedied this issue, we have no particular reason to believe that the suite of tests is *now* comprehensive, and that future codes will not employ more tests that reveal shortcomings in our current standard codes. To be epistemically satisfied, it seems as if we should want something like a general account of how the various tests fit together into an overall framework, specifically in a way that provides good evidence that all relevant sources of error are accounted for once-and-for-all.

In the next section, I will argue that such a fully comprehensive, once-and-for-all approach to model-verification is unnecessary, and that the philosophical intuitions motivating the strict V&V account are misleading. To lay the groundwork for this argument, I will begin by discussing a particular class of tests—those concerning fluid-mixing instabilities—in more detail. Then, on the basis of these and other examples, I will argue that these tests as used here do not fit the above philosophical intuitions about simulation verification, and that we should (at least in some cases) think about simulation verification as a more piecemeal, exploratory process.

Fluid-mixing instabilities refer to a class of phenomena arising, naturally, in hydrodynamic contexts at the boundary between fluids of different densities and relative velocities. *Kelvin-Helmholtz* (KH) instabilities arise from a shear velocity between fluids, resulting in a characteristic spiral-wave pattern; *Rayleigh-Taylor* (RT) instabilities occur when a lighter fluid presses against a denser fluid with a relative velocity perpendicular to the interface, resulting in structures described variously as "blobs" or "fingers".[2] In the course of galaxy formation, these instabilities are also subject to magnetic fields, which can suppress the growth of small-scale modes and produce novel behavior if the strength of the magnetic field is in the right regime. The importance of these phenomena have been understood for some time—in particular, the presence of KH instabilities is thought to have a significant impact on the stripping of gas from galaxies via ram pressure, which may account for variations in the properties of galaxies (Close et al., 2013). Chandrasekhar's standard theoretical treatment of these instabilities, both in the presence and absence of magnetic fields, was first published in 1961 (Chandrasekhar, 1961), and numerical studies of the same have been conducted at least since the mid-1990s (Frank et al., 1995; Jun et al., 1995).

Given the importance of these instabilities in galaxy formation processes, one might suppose that the ability of simulations to implement them properly would be an essential concern,

[2] Useful illustrations of both KH and RT instabilities, including time-series snapshots, are available in Springel (2010) and Hopkins (2015).

and that the verification tests performed would reflect this. However, as noted in Tables 2.1 and 2.2, none of the codes prior to Athena (2008) included explicit tests of the KH or RT instabilities in their method papers, and only Flash comments on the incidental appearance of KH instabilities in one of its tests. In addition to the surveyed codes, explicit KH and RT tests are also absent from the pre-2008 method papers for Gasoline (TREE-SPH) (Wadsley et al., 2004), Hydra (AP^{3}M-SPH) (Couchman et al., 1994), and Zeus (lattice finite-difference) (Stone and Norman, 1992). On the other hand, a brief perusal of post-2008 method papers such as rpsph (Abel, 2011), Enzo (AMR) (Bryan et al., 2014), Gasoline2 ("Modern" SPH) (Wadsley et al., 2017), and Phantom ("Modern" SPH) (Price et al., 2018), shows that they all *do* cite tests of these instabilities in various capacities.[3]

This disparity between pre- and post-2008 method papers with respect to their treatment of KH and RT tests can be traced (at least in significant part) to a code comparison project published in late 2007 (uploaded to arXiv in late 2006) by Agertz and other collaborators, including most of the authors of the various simulation codes already discussed (Agertz et al., 2007). In this hydrodynamic test, colloquially referred to as the "blob" test, a dense uniform spherical cloud of gas is placed in a supersonic wind tunnel with periodic boundaries and permitted to evolve, with the expectation that a bow shock will form, followed by dispersion via KH and RT instabilities. The dispersion patterns were compared to analytical approximations for the expected growth rate of perturbations, and the study concluded that, while Eulerian grid-based techniques were generally able to resolve these these instabilities, "traditional" SPH Lagrangian methods tend to suppress them and artificially prevent the mixing and dispersion of the initial gas cloud.

These observations led to a number of discussions and disagreements in the literature regarding the precise nature and sources of these problems. Beyond the normal issues with numerical convergence, the culprits were identified as insufficient mixing of particles at sub-

[3] Technically, Enzo only cites to Agertz et al. (2007), where it was used as one of the sample codes, but nonetheless the test is discussed in the method paper.

grid scales (Wadsley et al., 2008) and artificial surface tension effects at the boundary of regions of different density caused by the specifics of SPH implementation (Price, 2008). Eventually, these considerations lead to other fluid-mixing tests aimed at addressing cited shortcomings with the "blob" test (Robertson et al., 2010; McNally et al., 2012).

Concurrent to and following the development of these tests, a number of new SPH formalisms and codes (so-called "Modern" SPH, in contrast to traditional SPH) have been developed to address these problems and subjected to these tests. The proposals themselves are quite varied, from introducing artificial thermal conductivity terms (Price, 2008), to increasing the number of neighbor particles per computation (Read et al., 2010), to calculating pressure directly instead of deriving it from a discontinuous density (Hopkins, 2013). But the common thread is that now, with the phenomenon established and its causes analyzed, the tests that were developed in response to these have (at least for the time being) become new standards for the field.

What observations can we draw from this narrative? First, it should be apparent that the process described here does not follow the strict V&V account of simulation verification, as some sources of numerical error were not accounted for several generations of simulations. This is not to suggest that simulationists simply had no awareness that this area of their simulations might need more development—while the literature post-2008 certainly set the agenda and was the source for most of the key insights leading to the development of these tests, the problems with SPH were not entirely unknown before then. Indeed, while the specifics of the KH and RT instabilities were rarely referenced explicitly, SPH methods were known to have issues related to mixing and other instabilities at least as early as the 1990s (Morris, 1996; Dilts, 1999), and at least one variant of SPH was designed to address mixing issues as early as 2001 (Ritchie and Thomas, 2001). Despite this, the tests did not generally make appearances in method papers until codes were already reasonably capable of handling them, at least in some regimes. This, in turn, raises a concern that an analogous situation holds in the case of our current codes, with respect to as-of-yet ill-defined or underreported sources of error.

Second, in response to this concern, we should note that these verification tests do not present themselves as obvious or canonical; rather, they are a product of experimentation. Obviously, any insistence that simulationists should have tested for these errors before the tests were developed is practically confused, but there is a deeper theoretical point to be raised against the more abstract epistemic objection: the tests themselves are not simply tests of a simulation's fidelity to physical phenomena, but are also tailored to probe at and attain clarity regarding the nature of particular vulnerabilities in specific code types. Hence, the tests for KH and RT instabilities are not just looking to reproduce the expected physics, but are also made specifically to expose the unphysical numerics associated with SPH tests as well. By itself, this may not satisfy a proponent of the strict V&V perspective, but it does suggest that these tests serve a purpose much broader than mere "verification" that numerical error is within tolerance levels for a given simulation—they are also giving simulationists tools to explore the space of simulation code types. I will discuss this in greater detail in the next section, but for now it is enough to note that this means that verification tests are doing far more than "verification" as strictly defined—and, indeed, the development of these tests is just as crucial to the progress of the field as the development of the simulation codes themselves.

2.3 Leveraging both Physics and Numerics

Of course, while it may be suggestive, the narrative from the previous section does not show that this piecemeal and exploratory approach to simulation verification is epistemically sound. Certainly there is no sense in which these tests provide a patchwork cover of all possible situations wherein numerical error might arise, and thus they would fail to satisfy philosophers who stress the importance of complete inductive verification upfront, per the strict V&V account. One might suppose that the above approach is simply the best that can be done, given the constraints of complexity and the current state of knowledge in the field, but even this would imply that the simulationists in question should be doing more to give more thorough accounts of how their tests fit together into the best-available suite given these constraints. In any case, I do not believe such an account would be particularly satisfactory. In this section, I want to argue that the approach taken by the surveyed astrophysical MHD codes is not just epistemically benign (at least in principle), but that limiting simulationists to the strict V&V approach would be an error of outsized caution. Specifically, I will argue that the risks incurred by simulationists are not radically different from those found in ordinary (i.e., non-simulation based) methods of scientific inquiry.

From the strict V&V perspective, the risk of physical and numerical errors "cancelling" each other out leads to the prescription that the model-verification and computational model-validation of simulations should be distinct and sequential—that is to say, that model-verification should be (strictly speaking) a purely numerical/mathematical affair, and that any evaluations in terms of physics should be confined to the model-validation phase. Of course, even in this case it would be permissible for a simulationist to incidentally cast verification tests in physical terms, e.g., in terms of specific physical initial conditions, but this would just be a convenience. But as I suggested above, in practice verification tests are not simply convenient numerical exercises designed to check for generic numerical error. Rather, the tests serve as windows into the physics of the simulation, breaking down the distinction between physics and numerics and providing simulationists with a number of epistemic leverage points that would be obscured if we were to force them to regard verification tests as merely numerical in nature.

In general, the tests provide the simulationist with a sense of the physical phenomena represented because simulationists can interpret and understand mathematical equations in terms of the physical phenomena they represent. In other words, simulationists are not simply checking to see if a given equation produces numerical error by means of comparison to an analytical solution, though that is a useful benchmark if it exists. Rather, terms in the simulation equations have physical significance, *including* terms that are artifacts of the discretization of the original continuous equations. In the case of fluid-mixing instabilities, for instance, the shortcomings of the traditional SPH methods were not simply referred to as "numerical errors"—the error term was specifically characterized as an "artificial surface tension" that became non-negligible in the presence of a steep density gradient (Price, 2008). Where "fictions" such as artificial viscosity or artificial thermal conductivity terms are introduced, their justification is not cached out in numerical terms, but as appropriate physical phenomena whose inclusion will negate the influence of some other (spurious) error term, *because that error term behaves like a counteracting physical phenomenon.* Thus, on the one hand, the simulationist's preexisting physical intuitions about the appropriate behavior for the simulated system can serve to detect deviations that, upon investigation, may be determined to be numerical aberrations; on the other hand, the verification tests themselves enable the simulationist to develop this insight into the ways in which the simulation is functionally different from the corresponding real system.[4]

Moreover, this insight into the physical significance of these numerical terms allows the simulationist to partition the space of possible simulation scenarios in a manner that is far more salient for the purposes of extracting scientifically useful confidence estimates. If, e.g., a simulationist wanted to know whether a particular simulation code is likely to give reliable results when they simulate a galaxy with a particular range of properties, estimates of performance in terms of the generic categories of "numerical error"—round-off error, truncation error, etc.—are not going to be particularly useful. But an understanding of the kinds of *physical* phenomena for which this code is particularly error-prone lends itself more naturally

[4] Obviously, even if the simulation is "functionally different" from the corresponding target system, e.g. by inclusion of useful fictions, the goal of the simulationist is to ensure that these differences end up reproducing the correct behavior as an output.

to judgements of this form. These judgements can even take a more granular form, where different aspects of a simulation could be gauged more or less reliable based on the strengths of the simulation code—e.g., a simulationist would presumably be somewhat hesitant to draw strong conclusions about aspects of galaxy formation that rely on KH or RT instability mixing on the basis of a traditional SPH code.

But most importantly, this physical intuition allows for a kind of feedback loop, akin to the normal process of scientific discovery: we do our best to model complex systems by means of approximations, which in turn helps us understand how other, more subtle factors play an important role in the system; learning how to characterize and integrate these more subtle factors gives us a better, more robust model; and the process repeats. In this case, however, the object under investigation is not just the target system—we are also investigating the space of simulation code types, and experimenting with different ways to flesh out its properties by experimenting with various kinds of verification tests.

Of course, this approach is not foolproof. There will always exist the possibility that the simulationist is radically wrong about the adequacy of their simulation, that they have failed to account for some important phenomena. But this risk, while real, need not warrant wholesale skepticism of simulationist methods or embrace of the strict V&V account. In fact, this risk is analogous to the underdetermination risks incurred in the process of ordinary scientific inquiry—namely, that our theory might be incorrect or woefully incomplete, and that it only seems correct because some unaccounted-for causal factor is "cancelling out" the inadequacy of our theory. If we are going to regard this risk as defeasible in the context of the familiar methods of scientific inquiry, we should at least grant the possibility that the simulationist's risk is similarly benign.

Here, the proponent of the strict V&V approach may level an objection: namely, that the risks associated with simulation numerics "cancelling" other errors are potentially systematic in a way that the ordinary scientific risks of theory underdetermination by evidence are not. In the case of ordinary scientific theorizing, we regard this risk as defeasible because we have no reason to believe that the phenomena are conspiring to subvert our theorizing; even if we make mistakes given a limited set of data, we are confident that with enough rigorous testing we will eventually find a part of the domain where the inadequacies of the theory are apparent. In the case of simulation, however, one might worry that the risk may stem from a *systematic* collision between the numerical and physical errors, obfuscated by the complexities of the simulation—and if this is the case, further investigation will not allow us to self-correct, as continued exploration of the domain will not generally break this systematic confluence.

This objection makes some sense if we understand verification tests merely as straightforward tests of numerical fidelity. However, as I have tried to show, many verification tests are *not* of this simple character—by developing new kinds of tests to better understand the way simulation *codes* work, simulationists are simultaneously exploring the domain of possible real-world systems and probing the space of simulation code types. A particular verification test may be inadequate to the task of detecting or understanding certain kinds of errors—indeed, some argued in the literature that the original "blob" test proposed by Agertz et al. gave us a distorted picture of SPH's undermixing problem—but simulationists are not limited to a set of pre-defined tools. In the same way that we (defeasibly) expect that rigorous testing renders the risk of conspiracy tolerable in ordinary scientific contexts, the careful and targeted development of verification tests—in conjunction with the usual exploration of the domain of real systems—can mitigate the risk of conspiracy in the context of simulation.

With these considerations in mind, I would suggest that the best framework for thinking about these tests is as a collective network of tests roughly indexed to *phenomena*, specifically phenomena that, in the simulationist's estimation given the current state of knowledge in the field, are significant causal factors in the system under study. Under this picture, a simulation will be sufficiently (though defeasibly) verified just in case it produces tolerable results according to the full range of tests—which are themselves subject to scrutiny and modification as simulationists develop better understandings of how these tests probe their codes. This more pragmatic notion of sufficiency rejects the strict V&V insistence that simulations need to be verified against all sources of numerical error up front, but in exchange requires the simulationist to be sensitive to the various strengths and weaknesses of the code they are using—a sensitivity acquired in part by means of these tests, but also by general use of the code, and by familiarity with other codes and their strengths and weaknesses. In the next Chapter, I will outline a more general framework in which this type of practice can be understood.

2.4 Conclusion

In this chapter, I have presented a survey of the verification tests used in selected MHD codes, and drawn lessons about simulation justification on the basis of this real-world scientific practice. Notably, the pattern observed does not fit with the V&V framework's prescriptions, and a careful examination of the development and deployment of these tests shows that they serve epistemic functions beyond simply checking for numerical errors—they can be used to probe the differences between different code types and come to a deeper understanding of their strengths and weaknesses. By examining the case study of fluid-mixing instability tests, I traced this process in action and showed that the creation of these tests, the subsequent analysis, and the development of improved simulation codes is deeply entangled with our understanding of the underlying *physics*, not merely the numerics.

On the basis of this survey and case study, I argued that this process of improving our understanding of the target phenomena and the space of simulation code types can be understood to follow a pattern of incremental improvement similar to ordinary scientific theories in ordinary experimental contexts. I also addressed the a skeptical objection that might be leveled by those convinced by the strict V&V approach—in particular, given this expanded understanding of how verification tests can inform our investigations, we can be reasonably confident that we are not exposing ourself to any severe underdetermination risks.

This wider understanding of the role of verification tests also has significant implications for how we characterize the role of the simulationist—in particular, the simulationist's knowledge of simulation methods and techniques is not merely *instrumental* for the goal of learning about the target phenomenon, because the simulationist's understanding of the target phenomenon is developed in tandem with their knowledge of simulation methods and techniques. This entanglement suggests that merely reproducing some target phenomenon by simulation is not sufficient for a full understanding of that phenomenon—the simulationist must also understand the principles by which the different specifics of the various code types yield this common result.

Chapter 3

Simulation and Adequacy-for-Purpose

The argument in the previous chapter was based on a case study—and this, naturally, leaves a number of questions to be answered: to what extent can the particularities of this case study be generalized? Can we formulate a more general picture of how simulations can help us advance our knowledge in a generic field of study?

To address these questions, I will draw on the resources of the adequacy-for-purpose framework of model assessment. Simulations are, of course, a type of model, and thus they are amenable to this treatment. However, I aim to show that simulations naturally lend themselves to this kind of analysis in a way that draws on the strengths of this framework and reinforces its appeal against other conceptions of modeling. In particular, thus far the focus of this framework has been to show how models can be used as individuated tools to achieve some individuated purpose—in what follows, I will show how these models can be used as tools in concert with one another to develop better and better models.

The remainder of the chapter will proceed as follows. I begin by briefly outlining the adequacy-for-purpose framework, emphasizing some important aspects, and adapting some terminology (Section 3.1). I then characterize the core epistemic challenge that this paper seeks to address—namely, the problem of assessing the adequacy-for-purpose of simulation models where empirical evidence is scarce (Section 3.2). Finally, I argue that a suitable collection of models, considered in the light of the adequacy-for-purpose framework, can provide a satisfactory response to this challenge (Section 3.3).

3.1 The Aims of Adequacy-for-Purpose

In the philosophy of science literature, adequacy-for-purpose has been characterized as a useful framework for resisting the view that models should be judged solely against the ideal of a perfect and complete representation of their intended target—I'll call this the *norm of global accuracy.* In relaxing this demand, it joins a rich tradition that aims to understand the many ways in which models are used as tools and to better account for observed modeling practice (Parker, 2020, and references therein). Before I add my own example to this collection, it will be helpful to recast some of the terminology used by Parker in a manner more salient to the domain of simulation. Thus, in this section I will give a brief overview of the key elements of the adequacy-for-purpose framework and emphasize some subtleties that will be important for later sections.

According to Parker, introducing the consideration of purpose into model assessment leads to the key insight that a whole range of contextual factors must be considered:

> In particular, for a model to be adequate-for-purpose, it must stand in a suitable relationship not just with the representational target T but with a target T, user U, methodology W, circumstances B, and goal P jointly. The model must have features, including but not limited to how it represents target T, such that user U, using the model in way W in circumstances B achieves (or is very likely to achieve) purpose P. (Parker, 2020, 464)

Parker also argues, in light of this more contextual understanding of modeling practice, that the process of assessing adequacy-for-purpose will thus involve a *broader range of considerations* than under the norm of global accuracy. Here, it will be helpful to distinguish two broad ways in which a model M can fail to be adequate-for-purpose, as I will argue that we can neglect one of these in my subsequent analysis. Consider a particular model M, with intended target T.

First, M could be inadequate-for-purpose because it fails to adequately represent the salient aspects of T, relative to the demands of the contextual factors (U, W, B, P)—call this a failure of representational adequacy relative to P. In this regard, assessing M under the norm of global accuracy is more demanding, because by this metric it must be representationally accurate in all scenarios, not just those specified by (U, W, B, P). Or, conversely: if M fails to accurately represent T in any non-U, non-W, non-B, or non-P scenarios, this will not count against its adequacy-for-P—and in this sense, the contextual factors serve to *delimit* the scope of what must be considered in evaluating M, relative to a norm of global accuracy.

Second, M could be inadequate-for-purpose for reasons other than a failure of representational accuracy—e.g., M could be too complex, such that the process of extracting predictions from the model is too cumbersome for the intended purpose. In particular, in these situations M could, in principle, pass muster under the norm of global accuracy, yet still fail to be adequate-for-P; in this regard, the norm of global accuracy can be less demanding than the adequacy-for-purpose framework.

In general, the latter type of inadequacy-for-purpose will be a source of concern only in highly practical affairs—so long as M is representationally adequate for achieving P, additional problems can arise only by some mismatch between the model and the user's ability to access or leverage the needed representational content. In contexts of scientific inquiry, where time is not generally a scarce resource and where users are presumed to have the same baseline competence in managing models, neglecting these possible points of failure is a reasonable idealization. Thus, for the purposes of my discussion in the following sections, I will only be concerned with the former kind of inadequacy-for-purpose.

However, as Parker's general categories of contextual factors (U, W, B, P) do not always cleanly correspond to one or the other type of inadequacy-for-purpose, it will be helpful to introduce more fine-grained categories. Suppose we have some model and an intended target, and suppose that one intends to instantiate this model as a computer simulation. Even if we confine ourselves to a purely scientific context, different individuals will want to use this model for different purposes: to draw conclusions about different kinds of scientific hypotheses, to make predictions about certain systems, etc. Given our idealization—that the model will be adequate-for-purpose just in case it is representationally adequate for that purpose—we need not explicitly enumerate purpose as an independent contextual factor, as the various factors that encode different standards of representational adequacy will suffice. In particular, we will need to specify the methodology employed in constructing the simulation, the salient aspects of the target system, the regime under which our model must succeed, and the error tolerances.

By way of example, consider a simulation fluid dynamics model of the Navier-Stokes equations:

- The methodology employed concerns the details of the numerical method—e.g., the choice between particle and continuum methods, and within these a number of other choices regarding the discretization method and the handling of phenomena such as shocks.

- Some aspects of the target will be important, and other aspects can be neglected. Here, of course, certain kinds of deliberate fictions inserted into the simulation will not be counted against its representational adequacy. But one could also focus on some aspects, such that real non-vital features are left out entirely. In particular, these aspects of the model need not be individuated as spatially isolated components—e.g., even if weak magnetic fields are known to be present throughout the target, a modeler might choose to exclude those force terms from the simulation if they have negligible impact on the processes of interest.

- I take the regime to represent the kinds of initial conditions for which the simulation is intended to be reliable. This may be influenced by consideration of those aspects of the simulation that were idealized or neglected—forgoing any claim to representational adequacy in regimes where one expects that excluded or simplified elements of the simulation will play a non-negligible role.

- Error tolerances may be straightforward measures of accuracy, or involve more complicated tradeoffs between the accuracy of different aspects. In particular, different aspects may require more or less precision—e.g., depending on their needs, a modeler might be satisfied with a course-grained approximation of various turbulence effects without explicitly modeling the small-scale details.

For a particular choice of factors, note that each of these factor specifications can be thought of as a subset of a much larger set of possible factors that might have been chosen—the particular aspects of importance for a given purpose are a subset of the complete set of possibly relevant aspects, and so on. Following Parker, I'll call a specific choice of relevant contextual factors a *problem space*, as these outline the dimensions of a specific problem. I'll call the wider space from which these are drawn the *context space*.

Before moving on to the next section, two remarks are in order.

First, note that these two ways in which a model can fail to be adequate-for-purpose are not always cleanly separable in practice—e.g., one can certainly imagine situations in which it is difficult to discern where the overall problem with adequacy occurs.

Second, one might be skeptical that the adequacy-for-purpose framework is really a necessary starting point for this analysis. After all, by confining my analysis to considerations of representational adequacy, I am not taking advantage of the full scope of the adequacy-for-purpose framework—and one might even argue that I have idealized away its most salient features. Ultimately, I admit this may prove too narrow a construal to fully account for some practical situations—even in the context of simulation, factors unrelated to representational adequacy can prove important, contra my assumptions. My goal, however, is to isolate a challenge to the adequacy-for-purpose account that might not be apparent on its face; thus, while one might feel that the more interesting aspects of adequacy-for-purpose to lie elsewhere, this setup will suffice for the constructive purpose that I want to undertake.

3.2 A Challenge for Adequacy-for-Purpose

We have seen that an adequate-for-purpose model need not aspire to perfectly represent all aspects of the target system. In the previous section, I emphasized how this was less demanding, and ultimately a strength, relative to the norm of global accuracy. If we could completely compare models against their respective target systems (relative to the problem space, of course), not much more would need to be said.

Unfortunately, this kind of complete comparison will often be impossible. For one, the problem space generally covers a continuum of scenarios, and thus checking each of them for adequacy is impossible—this is an especially salient concern in simulation contexts, where empirical data is often quite scarce. But even if one ignores this difficulty or supposes a narrowly construed problem space, note that directly checking the complete model against the target system is a process that presupposes that we have access to all the relevant structural information about the target—and if we have this, the model itself is redundant. As such, we will need some account of how we can come to believe that a model is adequate-for-purpose without direct and complete comparisons. In this section, I will flesh out some of the subtleties with this challenge as it pertains to large-scale simulation.

Given the impossibility or redundancy of completely checking a model in all relevant circumstances, we must consider other ways to build confidence that a model is adequate-for-purpose. Parker suggests a number of possibilities, which can be used in combination: considerations related to the performance of model components, model construction, direct tests of adequacy, and indirect tests of adequacy (Parker, 2020, 467-71). In what follows, I will argue that the last three share a set of common difficulties.[1]

In direct tests of adequacy, one infers that the model is adequate in general by examining its performance within some subset of the problem space; in indirect tests of adequacy, one examines its performance in other contexts outside the problem space.[2] In both of these instances, the modeler is relying on implicit assumptions.

In the case of direct tests, there is the classic problem of induction: how do we know that the adequacy of a subset of the problem space suffices to establish the adequacy of the whole problem space? By itself, this is not a great cause for concern—as long as the modeler has god reasons to be confident that the tested subset is sufficiently similar to the rest of the problem space. However, this raises a second-order worry: how can we be confident that we have divided the context space into physically salient problem spaces? In setting up this framework, we have not been given any assurances that the chosen problem spaces have sufficient internal self-similarity to support this inference. Indirect tests also rely on these

[1] For sake of space, I will not address the possibility of assessing model components individually in any depth—while there are doubtless some circumstances in which testing the adequacy of individual modular components of the simulation may yield confidence in the overall model, 'fuzzy' modularity generally precludes this in highly complex simulation (Lenhard and Winsberg, 2010).

[2] Considerations of model construction involves asking whether the model "has properties that will facilitate the achievement of the purpose in a context or instance of use" (Parker 2020, 468)—however, given that we could only have knowledge that these properties facilitate the achievement of the purpose by virtue of some direct or indirect test, I will collapse this category into the other two.

implicit assumptions about similarities between contexts—in this case, between different problem spaces, rather than within a given problem space.[3]

Thus, if all attempts to evaluate a model rely on these kinds of similarity assumptions, we must ask whether and how we can have good grounds for believing that a given division of the context space is well-motivated. Certainly, we should not presume that the answer will be obvious, as the history of science provides ample evidence that our attempts to "carve nature at its joints" are prone to error—e.g., before Einstein, scientists failed to realize that $v \ll c$ and $v \approx c$ were relevantly different regimes for kinematics; because of this, the evidence for Newton's theory in the $v \ll c$ regime was presumed to lend support to the inference that Newton's Laws were adequate far outside that regime. The case of special relativity also provides an excellent example of one way that we can come to reevaluate our previous presumptions about how the context space should be divided—that is, we can sometimes find the model to be inadequate via some direct test in a regime that we previously assumed to be unproblematic.

However, this approach will not always be an option. In particular, highly complex simulations often cannot be directly tested against empirical data in many regimes of interest. This is not to say that highly complex simulations cannot be compared to *any* observational evidence—e.g., we can obviously check a given large-scale climate model to see if it reproduces current-day observations within tolerances. However, the purpose of these simulations is not generally limited to reproducing current observations; indeed, a common concern is that these simulations might be "overtuned" to spuriously reproduce these. Rather, we want to know if a model has achieved that state by adequately capturing the processes that drove the real target system—as this will, in turn, warrant further inferences about how the climate will evolve into the future. We may have a good theoretical basis for many of the model components in isolation, but nonlinearities and the prospect of numerical errors make this

[3] Indeed, one could simply recast indirect tests of adequacy as direct tests of adequacy relative to an expanded problem space.

insufficient by itself (Lenhard and Winsberg, 2010).

This, then, is our puzzle: assessing the adequacy of a model seems to rely on assumptions about how the context space should be divided, but in many cases we will be unable to test these assumptions using only the model. In the next section, I will show how a community of modelers can work in tandem to get traction on these assumptions.

A brief aside: the challenge I've outlined in this section is certainly not unique to the adequacy-for-purpose framework, though its explicit invocation of contextual factors makes the problem perspicuous. Indeed, the norm of global accuracy faces all these problems and more—the same basic problem of induction applies, and thus any attempts to make general claims will need to implicitly rely on these kinds of similarity assumptions. Thus, while I have cast this problem in the language of the adequacy-for-purpose framework, it is not a problem with adequacy-for-purpose *per se*; rather, it is a general problem with empirical investigation, and the adequacy-for-purpose framework provides a natural way of both characterizing and confronting this problem.

3.3 An Ecosystem of Adequate-for-Purpose Models

Suppose a community of simulationists has some range of priors about the appropriate way to divide the context space into problem spaces, and the relationship between these problem spaces and purposes they might serve. Suppose, further, that at least some of these priors are flawed in some important respect: perhaps the community believes that capturing some aspects of the target is sufficient to adequately predict some quantity of interest, but in fact a wider set is needed; perhaps the community has failed to grasp that some regime should be further divide into more fine-grained regimes; etc. Under what circumstances will the practices of the community provide an impetus to reevaluate these priors?

As noted in the previous section, direct tests of a single simulation are unlikely to provide many points of leverage when empirical data is scarce. Instead, suppose that the community constructs an ensemble of simulations, each designed on variations to the problem space—using different methodologies, emphasizing different aspects, and intended to hold in different regimes. Crucially, these problem spaces will not generally be disjoint; because they are aiming to simulate similar systems, they will typically overlap with respect to many of the intended aspects and regimes. This overlap, then, provides a basis for assessing these simulations—not relative to the target system, but relative to each other.

As a simple example, imagine that two simulations with different methodologies are intended to model the same aspects of the target system in the same regime. Even if direct comparisons to the target system are impossible, a mismatch between them is evidence that at least one is inadequate—and this, in turn, may lead the community to any number of diagnoses as to the source of the problem: perhaps one of the methodologies is vulnerable to numerical errors in this regime, and thus we must adjust it to compensate; perhaps a correspondence in one part of the regime and a mismatch in another is evidence that this regime should be broken into finer-grained categories; etc. Similarly, a comparison between two simulations that were designed to simulate different but overlapping sets of aspects may provide evidence about the relative importance of various aspects in producing certain kinds of phenomena, show that one methodology has trouble representing some aspect in certain regimes, etc. Moreover, these assessments need not be limited to simple pairwise comparisons; if there is some question as to the source of a discrepancy, other contexts and models may help adjudicate. Once these sources are understood, further models can be developed to account for the underlying problems—and this may involve minor refinements to existing simulations, or a more significant adjustment to the community's conceptual framework.[4]

[4] Peschard (2011) describes a similar process in the context of experiment, and in doing so offers a useful way to think about how investigation can provoke conceptual clarification and innovation. Using a case study from the field of computational fluid dynamics, she argues that experiments can be used as part of a creative and interactive process to parse relevant parameters from artifacts—which, by showing new factors to be relevant, can provoke conceptual change. The above account similarly describes a process for the

In practice, these kinds of comparisons can take a variety of forms, from published code comparisons to more informal tests in code development—as we saw in Chapter 2, even verification tests, which are sometimes thought to be simple benchmarks, can be used to probe differences in simulation methods and clarify the importance of previously-overlooked aspects of the target system. And while the expertise needed to construct effective comparisons is often developed informally—by long experience, by intra-lab apprenticeships and inter-lab collaboration, by conference and workshop discussion—the fruits of this expertise can nonetheless be appraised by looking to the broad trajectory of the field. Insofar as the community manages to progressively refine their collective understanding and avoid conceptual stagnation, we have good reason to believe that successive models are more adequate than previous models.[5]

Thus, while a single simulation may not be enough to investigate the community's priors, a suitable ecosystem of models—not too diverse, not too similar—can provide a basis for investigation and reevaluation. Further models, built with this better understanding in mind, can then serve as a foundation for further investigation and refinement.

Of course, one might object that this method does not suffice to establish a real solution to the problem posed. There are a number of ways this objection might be construed, so I will address each in turn.

First, one might argue that I have only given a how-possibly story—i.e., nothing in my argument shows that a discrepancy will necessarily exist to witness a given inaccuracy in the

adjustment and refinement of the various background concepts that mediate our knowledge of the target phenomena—and while one will not generally be able to "read off" the necessary conceptual change from a given discrepancy, it nonetheless can provide an indication that conceptual change is needed and a framework for pursuing this change in a guided manner.

[5] Smeenk and Gallagher (2020) emphasize the importance of *eliminative reasoning* for validating large-scale simulations against numerical artifacts in astrophysical contexts. My account also evinces a kind of eliminative reasoning, as the goal is to eliminate problems with the background assumptions—though here, I have emphasized the detection aspect of this process, as identifying and characterizing new sources of error will be just as important as accounting for known sources of error. To their account, I would add only that this eliminative reasoning can be useful for conceptual innovation, in addition to numerical error detection.

community's priors. This will likely be true in some ways at a given time, and may be true in some respects in principle. However, the community is not generally limited to improving their understanding of these systems in one way—as long as some avenues for improvement remain open, they can continue to refine their models and priors until they are able to pick up on more subtle factors.

Second, one might argue that this method fails to guarantee that the community will stumble across a discrepancy needed to reevaluate their priors. To this, I must admit that I have no satisfactory answer. However, I must also protest that the same problem is endemic to ordinary experimental methods—we can never be assured that an exploration of an inductive space will be optimal, and thus we can only commit to rigorously testing as best we can.

Third, one might argue that, for a given comparison, the models in question will be either too different to be truly comparable, or too similar to provide the needed leverage to draw useful conclusions. Gueguen (2021) argues that this tension is a problem within code comparisons, and draws on the examples from galaxy simulation code comparisons substantiate this point. While her critique is principally aimed at the much stronger claim that code comparisons can serve as "a method for determining when simulations faithfully track the logical consequences" of the target system, it is still worth pointing out two factors that distinguish the above analysis from the object of her critique.

First, Gueguen's examples (the AQUILA and AGORA code comparisons) were both large-base code comparison projects that compared codes with respect to fairly generic initial conditions; by contrast, the methods I have outlined will often require a carefully targeted approach. Verification tests, for example, could be framed as highly targeted code comparisons—and as we saw in Chapter 2, these have been successful in the past. Similar successes have been found with respect to more scope-limited code comparison projects, as well (Meskhidze, 2022). Gueguen's conclusion is certainly compelling with regard to large-scale code comparisons such as AGORA and AQUILA, as we would have little reason to believe that generically initial conditions will yield a useful discrepancy. However, this does not mean that carefully selected comparisons—especially if motivated and refined by other comparisons—cannot find points of epistemic leverage.

Second, while Gueguen is correct that similar codes may encode similar biases, this only speaks to the limitations of a single comparison. In particular, the method described above does not require that a single comparison be able to ferret out all problems with our priors at once—it is enough that these comparisons can help us refine *some* of our priors.

3.4 Conclusion

In this chapter, I have drawn on the adequacy-for-purpose framework to give an account of how a collection of models can be used to advance our understanding of some target phenomenon, even where a single model alone would be inadequate. In particular, this account generalizes the fluid-mixing instabilities case study from Chapter 2, where the Agertz et al. (2007) comparison between SPH and Eulerian grid methods provoked a series of investigations into the structure of these respective code types, and subsequently led to the development of more adequate simulation codes. Just as these targeted comparisons provided the epistemic leverage needed to refine our understanding of the role of fluid-mixing instabilities in galactic phenomena, careful comparisons among an ecosystem of models can flag areas in need of conceptual refinement or simple model adjustments. And as these improvements to our conceptual understanding of the target phenomenon accrue, simulationists are (by degrees) more justified in relying on these later-developed models—at least for applications within well-studied problem spaces.

Moreover, this account is not merely theoretical—to the extent that this account has already been concretely realized in practice, we can be confident that it will prove practicable in other contexts where an ecosystem of suitable models exist. To be clear, this does not preclude the possibility that other methods could prove similarly useful for advancing our understanding of simulation models or justifying particular simulation codes, and caution is always advisable when trying to apply a framework generalized from one context in a different context. However, insofar as these kinds of conceptual refinements are necessary for progress in science, this account provides a framework for delving into the details of different fields of study and comparing how they address particular difficulties.

Finally, this account suggests a number of connections to broader themes in the philosophy of science—as noted in footnote 4, a similar account of guided conceptual change in the context of experimental methods may prove an interesting point of comparison and contrast between simulation methods and experimental methods.

Concluding Remarks

Throughout this dissertation, I have developed a framework for understanding the epistemology of large-scale simulation that reflects the practice of working scientists: rather than the simplistic two-step V&V process, simulationists develop their simulation models by drawing on a host of resources—analytical benchmarks, empirical data, the ecosystem of models, etc.—as well as their knowledge of both numerical methods and physical systems. These enable the progressive refinement of this knowledge, and (defeasibly) justifies our simulation models in well-probed contexts and regimes. If the above account is correct, as a given field evolves, new kinds of phenomena will be understood and more tests and methods will be developed—both to enable this understanding and on the basis of it.

This account, in turn, suggests a number of practical prescriptions for working simulationists. While I have tried to describe this approach to the epistemology of simulation with enough specificity to allow for these kinds of inferences, many of these will depend on the granular details of particular codes and contexts—in these cases, I leave it to the simulationists themselves to evaluate if any of these apply in the particular. However, I will proffer two broader remarks about the significance of this framework.

First, the ecosystem requires a diversity of models to function properly—and as a failure of the ecosystem has consequences for all individual models, simulationists have a collective interest in ensuring that this diversity within their community is maintained at a healthy level. Of course, this may be easier said than done, given the massive resource commitment required to start up a large-scale simulation code project; however, if the community recognizes this diversity as a common good, it should be more likely to shoulder some of those costs that might otherwise be borne primarily by individuals. Alternatively, this framework also suggests that, in cases where the divide between communities is mostly a matter of disciplinary convention, boundary-crossing between these communities should be encouraged.

Second, the entanglement between physics and numerics in simulation justification raises a number of interesting questions for classical debates within the philosophy of science. E.g., regarding the nature of scientific explanation—can this be accommodated by previous accounts, or does this represent a new type of explanation? Regarding the philosophy of experiment—can similar kinds of justification be given for other unorthodox experimental models, such as analog experiments? Though I have not addressed these questions in this dissertation, these represent interesting and potentially fruitful avenues for future research in the light of a novel conceptual framework.

Bibliography

(2008). The athena code test page. https://www.astro.princeton.edu/~jstone/Athena/tests/. Accessed: 2020-11-30.

(2019). Flash user guide. http://flash.uchicago.edu/~jbgallag/2012/flash4_ug/. Accessed: 2020-11-14.

Abel, Tom (2011). rpsph: a novel smoothed particle hydrodynamics algorithm. *Monthly Notices of the Royal Astronomical Society*, 413(1): 271–285.

Agertz, Oscar, Ben Moore, Joachim Stadel, Doug Potter, Francesco Miniati, Justin Read, Lucio Mayer, Artur Gawryszczak, Andrey Kravtsov, Åke Nordlund, et al. (2007). Fundamental differences between sph and grid methods. *Monthly Notices of the Royal Astronomical Society*, 380(3): 963–978.

Balsara, Dinshaw S and Daniel S Spicer (1999). A staggered mesh algorithm using high order godunov fluxes to ensure solenoidal magnetic fields in magnetohydrodynamic simulations. *Journal of Computational Physics*, 149(2): 270–292.

Bauer, Andreas and Volker Springel (2012). Subsonic turbulence in smoothed particle hydrodynamics and moving-mesh simulations. *Monthly Notices of the Royal Astronomical Society*, 423(3): 2558–2578.

Beckwith, Kris and James M Stone (2011). A second-order godunov method for multidimensional relativistic magnetohydrodynamics. *The Astrophysical Journal Supplement Series*, 193(1): 6.

Beisbart, Claus (2019a). Should validation and verification be separated strictly? In *Computer Simulation Validation*, pages 1005–1028. Springer International Publishing.

Beisbart, Claus (2019b). What is validation of computer simulations? toward a clarification of the concept of validation and of related notions. In *Computer Simulation Validation*, pages 35–67. Springer International Publishing.

Bertschinger, E (1985). Self-similar secondary infall and accretion in an einstein-de sitter universe. *The Astrophysical Journal Supplement Series*, 58: 39–65.

Brio, Moysey and Cheng Chin Wu (1988). An upwind differencing scheme for the equations of ideal magnetohydrodynamics. *Journal of computational physics*, 75(2): 400–422.

Bryan, Greg L, Michael L Norman, Brian W O'Shea, Tom Abel, John H Wise, Matthew J Turk, Daniel R Reynolds, David C Collins, Peng Wang, Samuel W Skillman, et al. (2014). Enzo: An adaptive mesh refinement code for astrophysics. *The Astrophysical Journal Supplement Series*, 211(2): 19.

Burkert, Andreas and Peter Bodenheimer (1993). Multiple fragmentation in collapsing protostars. *Monthly Notices of the Royal Astronomical Society*, 264(4): 798–806.

Calder, Alan C, Bruce Fryxell, T Plewa, Robert Rosner, LJ Dursi, VG Weirs, T Dupont, HF Robey, JO Kane, BA Remington, et al. (2002). On validating an astrophysical simulation code. *The Astrophysical Journal Supplement Series*, 143(1): 201.

Chandrasekhar, S (1961). Hydrodynamic and hydromagnetic stability oxford univ. *Press (Clarendon) London and New York*.

Close, JL, JM Pittard, TW Hartquist, and SAEG Falle (2013). Ram pressure stripping of the hot gaseous haloes of galaxies using the k-ϵ sub-grid turbulence model. *Monthly Notices of the Royal Astronomical Society*, 436(4): 3021–3030.

Colgate, Stirling A and Richard H White (1966). The hydrodynamic behavior of supernovae explosions. *The Astrophysical Journal*, 143: 626.

Couchman, HMP, PA Thomas, and FR Pearce (1994). Hydra: An adaptive–mesh implementation of pppm–sph. *arXiv preprint astro-ph/9409058*.

Dee, Dick P. (1991). Prescribed solution forcing method for model verification in hydraulic engineering. *Proceedings of the 1991 National Conference on Hydraulic Engineering.*

Dilts, Gary A (1999). Moving-least-squares-particle hydrodynamics—i. consistency and stability. *International Journal for Numerical Methods in Engineering*, 44(8): 1115–1155.

Einfeldt, Bernd, Claus-Dieter Munz, Philip L Roe, and Björn Sjögreen (1991). On godunov-type methods near low densities. *Journal of computational physics*, 92(2): 273–295.

Emery, Ashley F (1968). An evaluation of several differencing methods for inviscid fluid flow problems. *Journal of Computational Physics*, 2(3): 306–331.

Evrard, August E (1988). Beyond n-body-3d cosmological gas dynamics. *Monthly Notices of the Royal Astronomical Society*, 235: 911–934.

Frank, Adam, Thomas W Jones, Dongsu Ryu, and Joseph B Gaalaas (1995). The mhd kelvin-helmholtz instability: A two-dimensional numerical study. *The Astrophysical Journal*, 460: 777.

Frenk, CS, SDM White, P Bode, JR Bond, GL Bryan, R Cen, HMP Couchman, August E Evrard, N Gnedin, A Jenkins, et al. (1999). The santa barbara cluster comparison project: a comparison of cosmological hydrodynamics solutions. *The Astrophysical Journal*, 525(2): 554.

Fryxell, B., K. Olson, P. Ricker, F. X. Timmes, M. Zingale, D. Q. Lamb, P. MacNeice, R. Rosner, J. W. Truran, and H. Tufo (2000). FLASH: An adaptive mesh hydrodynamics code for modeling astrophysical thermonuclear flashes. *The Astrophysical Journal Supplement Series*, 131(1): 273–334.

Gardiner, Thomas A and James M Stone (2005). An unsplit godunov method for ideal mhd via constrained transport. *Journal of Computational Physics*, 205(2): 509–539.

Gresho, Philip M and Stevens T Chan (1990). On the theory of semi-implicit projection methods for viscous incompressible flow and its implementation via a finite element method that also introduces a nearly consistent mass matrix. part 2: Implementation. *International journal for numerical methods in fluids*, 11(5): 621–659.

Grier, Benjamin, Richard Figliola, Edward Alyanak, and José Camberos (2015). Discontinuous solutions using the method of manufactured solutions on finite volume solvers. *AIAA Journal*, 53(8): 2369–2378.

Grismer, M. J. and J. M. Powers (1996). Numerical predictions of oblique detonation stability boundaries. *Shock Waves*, 6(3): 147–156.

Guan, Xiaoyue and Charles F Gammie (2008). Axisymmetric shearing box models of magnetized disks. *The Astrophysical Journal Supplement Series*, 174(1): 145.

Gueguen, Marie (2021). Comparability or diversity: A tension within code comparisons. *British Journal for the Philosophy of Science*, Forthcoming.

Hawley, John F and James M Stone (1995). Mocct: A numerical technique for astrophysical mhd. *Computer Physics Communications*, 89(1-3): 127–148.

Heitmann, Katrin, Paul M Ricker, Michael S Warren, and Salman Habib (2005). Robustness of cosmological simulations. i. large-scale structure. *The Astrophysical Journal Supplement Series*, 160(1): 28.

Heß, Steffen and Volker Springel (2010). Particle hydrodynamics with tessellation techniques. *Monthly Notices of the Royal Astronomical Society*, 406(4): 2289–2311.

Hopkins, Philip F (2013). A general class of lagrangian smoothed particle hydrodynamics methods and implications for fluid mixing problems. *Monthly Notices of the Royal Astronomical Society*, 428(4): 2840–2856.

Hopkins, Philip F. (2015). A new class of accurate, mesh-free hydrodynamic simulation methods. *Monthly Notices of the Royal Astronomical Society*, 450(1): 53–110.

Huang, Jingfang and Leslie Greengard (1999). A fast direct solver for elliptic partial differential equations on adaptively refined meshes. *SIAM Journal on Scientific Computing*, 21(4): 1551–1566.

Jeans, James Hopwood (1902). I. the stability of a spherical nebula. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 199(312-320): 1–53.

Jebeile, Julie and Vincent Ardourel (2019). Verification and validation of simulations against holism. *Minds and Machines*, 29(1): 149–168.

Jun, Byung-Il, Michael L Norman, and James M Stone (1995). A numerical study of rayleigh-taylor instability in magnetic fluids. *The Astrophysical Journal*, 453: 332.

Keppens, Rony (2004). Nonlinear magnetohydrodynamics: numerical concepts. *Fusion science and technology*, 45(2T): 107–114.

Khokhlov, Alexei M (1998). Fully threaded tree algorithms for adaptive refinement fluid dynamics simulations. *Journal of Computational Physics*, 143(2): 519–543.

Klein, Richard I, Christopher F McKee, and Philip Colella (1994). On the hydrodynamic interaction of shock waves with interstellar clouds. 1: Nonradiative shocks in small clouds. *The Astrophysical Journal*, 420: 213–236.

Knupp, Patrick (2002). *Verification of Computer Codes in Computational Science and Engineering.* Chapman and Hall/CRC.

Lenhard, Johannes and Eric Winsberg (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.

Liska, Richard and Burton Wendroff (2003). Comparison of several difference schemes on 1d and 2d test problems for the euler equations. *SIAM Journal on Scientific Computing*, 25(3): 995–1017.

Londrillo, Pasquale and Luca Del Zanna (2000). High-order upwind schemes for multidimensional magnetohydrodynamics. *The Astrophysical Journal*, 530(1): 508.

MacLaurin, Colin (1801). *A Treatise on Fluxions: In Two Volumes*, volume 1. W. Baynes and W. Davis.

McNally, Colin P, Wladimir Lyra, and Jean-Claude Passy (2012). A well-posed kelvin-helmholtz instability test and comparison. *The Astrophysical Journal Supplement Series*, 201(2): 18.

Meskhidze, Helen (2022). (what) do we learn from code comparisons? a case study f self-interacting dark matter implementations. In *Philosophy of Astrophysics: Stars, Simulations, and the Struggle to Determine What Is Out There.* Springer.

Morris, Joseph Peter (1996). A study of the stability properties of smooth particle hydrodynamics. *Publications of the Astronomical Society of Australia*, 13: 97–102.

Morrison, Margaret (2015). *Reconstructing Reality.* Oxford University Press.

Noh, William F (1987). Errors for calculations of strong shocks using an artificial viscosity and an artificial heat flux. *Journal of Computational Physics*, 72(1): 78–120.

Oberkampf, William L. and Christopher J. Roy (2010). *Verification and Validation in Scientific Computing*. Cambridge University Press.

Orszag, Steven A and Cha-Mei Tang (1979). Small-scale structure of two-dimensional magnetohydrodynamic turbulence. *Journal of Fluid Mechanics*, 90(1): 129–143.

Pakmor, Rüdiger, Volker Springel, Andreas Bauer, Philip Mocz, Diego J Munoz, Sebastian T Ohlmann, Kevin Schaal, and Chenchong Zhu (2016). Improving the convergence properties of the moving-mesh code arepo. *Monthly Notices of the Royal Astronomical Society*, 455(1): 1134–1143.

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

Peschard, Isabelle (2011). Modeling and experimenting. In *Models, Simulations, and Representation*. Routledge.

Powers, Joseph M. and Keith A. Gonthier (1992). Reaction zone structure for strong, weak overdriven, and weak underdriven oblique detonations. *Physics of Fluids A: Fluid Dynamics*, 4(9): 2082–2089.

Powers, Joseph M. and D. Scott Stewart (1992). Approximate solutions for oblique detonations in the hypersonic limit. *AIAA Journal*, 30(3): 726–736.

Price, Daniel J (2008). Modelling discontinuities and kelvin–helmholtz instabilities in sph. *Journal of Computational Physics*, 227(24): 10040–10057.

Price, Daniel J, James Wurster, Terrence S Tricco, Chris Nixon, Stéven Toupin, Alex Pettitt, Conrad Chan, Daniel Mentiplay, Guillaume Laibe, Simon Glover, et al. (2018). Phantom: A smoothed particle hydrodynamics and magnetohydrodynamics code for astrophysics. *Publications of the Astronomical Society of Australia*, 35.

Read, JI, T Hayfield, and O Agertz (2010). Resolving mixing in smoothed particle hydrodynamics. *Monthly Notices of the Royal Astronomical Society*, 405(3): 1513–1530.

Ritchie, Benedict W and Peter A Thomas (2001). Multiphase smoothed-particle hydrodynamics. *Monthly Notices of the Royal Astronomical Society*, 323(3): 743–756.

Roache, Patrick J. (2009). *Fundamentals of Verification and Validation*. Hermosa Publishers.

Roache, Patrick J. (2019). The method of manufactured solutions for code verification. In *Simulation Foundations, Methods and Applications*, pages 295–318. Springer International Publishing.

Robertson, Brant E, Andrey V Kravtsov, Nickolay Y Gnedin, Tom Abel, and Douglas H Rudd (2010). Computational eulerian hydrodynamics and galilean invariance. *Monthly Notices of the Royal Astronomical Society*, 401(4): 2463–2476.

Sedov, LI (1959). Similarity and dimensional methods in mechanics (new york: Academic) cahill me and taub ah, 1971. *Commun. Math. Phys*, 21(1).

Shu, Chi-Wang and Stanley Osher (1989). Efficient implementation of essentially non-oscillatory shock-capturing schemes, ii. In *Upwind and High-Resolution Schemes*, pages 328–374. Springer.

Smeenk, Chris and Sarah C. Gallagher (2020). Validating the universe in a box. *Philosophy of Science*, 87(5): 1221–1233.

Sod, Gary A (1978). A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws. *Journal of computational physics*, 27(1): 1–31.

Springel, Volker (2005). The cosmological simulation code gadget-2. *Monthly notices of the royal astronomical society*, 364(4): 1105–1134.

Springel, Volker (2010). E pur si muove: Galilean-invariant cosmological hydrodynamical simulations on a moving mesh. *Monthly Notices of the Royal Astronomical Society*, 401(2): 791–851.

Stone, James M, Thomas A Gardiner, Peter Teuben, John F Hawley, and Jacob B Simon (2008). Athena: a new code for astrophysical mhd. *The Astrophysical Journal Supplement Series*, 178(1): 137.

Stone, James M, John F Hawley, Charles R Evans, and Michael L Norman (1992). A test suite for magnetohydrodynamical simulations. *The Astrophysical Journal*, 388: 415–437.

Stone, James M and Michael L Norman (1992). Zeus-2d: a radiation magnetohydrodynamics code for astrophysical flows in two space dimensions. i-the hydrodynamic algorithms and tests. *The Astrophysical Journal Supplement Series*, 80: 753–790.

Teyssier, Romain (2002). Cosmological hydrodynamics with adaptive mesh refinement-a new high resolution code called ramses. *Astronomy & Astrophysics*, 385(1): 337–364.

Torrilhon, Manuel (2003). Uniqueness conditions for riemann problems of ideal magnetohydrodynamics. *Journal of plasma physics*, 69(3): 253.

Tóth, Gábor (2000). The $\nabla \cdot b = 0$ constraint in shock-capturing magnetohydrodynamics codes. *Journal of Computational Physics*, 161(2): 605–652.

Wadsley, JW, G Veeravalli, and HMP Couchman (2008). On the treatment of entropy mixing in numerical cosmology. *Monthly Notices of the Royal Astronomical Society*, 387(1): 427–438.

Wadsley, James W, Benjamin W Keller, and Thomas R Quinn (2017). Gasoline2: a modern smoothed particle hydrodynamics code. *Monthly Notices of the Royal Astronomical Society*, 471(2): 2357–2369.

Wadsley, James W, Joachim Stadel, and Thomas Quinn (2004). Gasoline: a flexible, parallel implementation of treesph. *New astronomy*, 9(2): 137–158.

Winsberg, Eric (2010). *Science in the age of computer simulation.* University of Chicago Press.

Winsberg, Eric (2018). *Philosophy and Climate Science.* Cambridge University Press.

Woods, C. Nathan and Ryan P. Starkey (2015). Verification of fluid-dynamic codes in the presence of shocks and other discontinuities. *Journal of Computational Physics*, 294: 312–328.

Woodward, Paul and Phillip Colella (1984). The numerical simulation of two-dimensional fluid flow with strong shocks. *Journal of computational physics*, 54(1): 115–173.

Yee, Helen C, Marcel Vinokur, and M Jahed Djomehri (2000). Entropy splitting and numerical dissipation. *Journal of Computational Physics*, 162(1): 33–81.

Zel'Dovich, Ya B (1970). Gravitational instability: An approximate theory for large density perturbations. *Astronomy and astrophysics*, 5: 84–89.