Simulation Verification in Practice

Kevin Kadowaki

Abstract

With the increased use of simulations as investigative tools in various scientific fields, the question naturally arises as to how these simulations are epistemically justified. One natural approach is to insist that the numerical aspects of simulation justification be performed separately from the physical aspects, but Winsberg (2010) has argued that this is impossible for highly complex simulations. Based on a survey and close examination of a range of astrophysical MHD codes and their attendant literature, I argue that insisting on a strict separation of these aspects of simulation justification is neither epistemically necessary nor advisable.

Introduction

Given constraints on scientists' abilities to observe and experiment, simulations have become a crucial tool for investigating certain kinds of large-scale phenomena. These tools, however, do not come without costs, and naturally philosophers of science have raised a host of epistemic questions as to when simulations can be relied on and how this reliance can be justified. These questions are especially pressing in the case of highly complex simulations, where the efficaciousness of the various methods for sanctioning simulations—code comparisons, convergence tests, benchmarking—is often in question, due to nonlinearities and the sheer size of the simulation. In particular, the rise of simulation highlights the importance of understanding and guarding against the kinds of numerical error introduced by computational methods.

A common and *prima facie* intuitive approach to this problem is to insist that a proper epistemology of simulation will require a separation of the numerical or purely computational aspect of simulation justification from the process of comparing the simulation to real-world target system. The Verification and Validation (V&V) framework captures this intuition, conceptualizing a split between the purely numerical task of ensuring that the computer simulation adequately represents the theoretical model (verification), and the task of comparing the output of the computer simulation to the real-world target phenomenon (validation). Per the V&V account, these separate treatments are required to avert the epistemic risk that errors in one domain may "cancel" errors in the other, leading to false confidence in the adequacy of our scientific theories.

Eric Winsberg has argued that this 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 this paper, I will present further evidence showing that the prescription goes largely unheeded in the context of astrophysical magnetohydrodynamics (MHD) simulations. But even if Winsberg

has 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 methods that simulationists *do* use can mitigate this risk, despite the fact that they do not follow strict V&V prescription. 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 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, I lay out the specifics of the V&V account and show that the survey results are incompatible with this account. To diagnose the problem, 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 argue that this approach to complex simulation verification is more exploratory and piecemeal than philosophers have supposed. In section 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.

1

The survey here concerns *verification tests*, i.e. tests that involve running a simulation with specifically chosen initial conditions and comparing the output to a known analytic solution or some other non-empirical-data metric. Significant discrepancies are then generally taken to indicate some failure of the discretized simulation equations to mimic the original, non-discretized equations—e.g., if a set of hydrodynamic equations naturally conserve energy, but a test of the discretized simulation of these equations shows that energy is not conserved, one can conclude that the numerical methods implemented are the source of the error.

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 which were particularly influential and which 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]

[1]Technically, Gizmo is able to facilitate a number of sub-methods, including "traditional" SPH. The new

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).

	FLASH	RAMSES	GADGET-2	ATHENA	AREPO	GIZMO
	Fryxell et al 2000	Teyssier 2002	Springel 2005	Stone et al 2008	Springel 2010	Hopkins 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](Stone, Stone) [q](Robertson et al., 2010) [r](McNally et al., 2012) [s](Abel, 2011) [t](Khokhlov, 1998)

Table 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.

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

methods of interest here are the Meshless Finite-Volume and Meshless Finite-Mass described in Hopkins 2015.

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 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 1 (hydrodynamic tests), Table 2 (magnetohydrodynamics tests), and Table 3 (self-gravity tests) (FLASH is omitted from Table 2; for an overview of those MHD tests that were eventually included, see Table 4). Table 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 to 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 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 straightfor-

	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](Stone, Stone)

Table 2: Magnetohydrodynamics tests. As in Table 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*.

ward rigorous quantitative standards such as analytic solutions. Further examination of multi-group code-comparison projects also shows that these projects 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 not straightforward side-by-side comparisons, 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 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 include 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

	Flash	Ramses	Gadget-2	Athena	Arepo	Gizmo
	Fryxell et al 2000	Teyssier 2002	Springel 2005	Stone et al 2008	Springel 2010	Hopkins 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 3: Self-gravity tests.

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 debut of the MHD Kelvin-Helmholtz and Rayleigh-Taylor instability tests. I speculate that the consistency apparent in magnetohydrodynamics tests is a function in part of the influence of J. Stone, who (with coauthors) proposed a systematic suite of test 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 his 2008 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 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 3, 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.

	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 4: Tests included in various editions of the FLASH user guide.

2

In the philosophical literature, the concept of simulation verification has been heavily influenced by the Verification & Validation (V&V) framework, which itself originated in a number of subfields within the sciences (Oberkampf and Roy, 2010)—including computational fluid dynamics, which has some obvious theoretical overlap with the field of astrophysical magnetohydrodynamics. 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 V&V framework to motivate a philosophical perspective on the proper approach to simulation verification, which I will then contrast with an examination of the tests as they are found in the above survey.

Within the V&V framework, a simulation is said to be *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 *validated* when the output of the simulation adequately corresponds to the phenomena in the world.[2] Together, these two components form a bridge between the phenomenon in the world

[2]As Beisbart (2019) has shown, there is some ambiguity regarding the use of the term "validation" in the literature. Here, I will be using the term to refer to what he distinguishes as *computational model validation*, and for our purposes other distinctions are not relevant.

and the analytical equations that constitute our attempts to theoretically capture that phenomenon, via the intermediary of the simulation code. Crucially, this means that verification and validation refer to correspondences over a range of simulation runs—see, e.g., various definitions of "validation" surveyed in (Beisbart, 2019), where notions such as "domain of applicability" implicitly make clear that these concepts are not simply correspondences with respect to an individual system. 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 concern is compounded in highly complex simulations such as the ones at issue here, as the nonlinear regimes at issue make it difficult to assess whether an effect is numerical or physical. Ultimately, this epistemic concern has led some philosophers to stress the importance of a sequential ordering for these activities: first verification, then validation. If the simulationist ensures that the simulation code is free of numerical errors *independently* of any comparisons to the phenomena, then this should preempt any risk that we might accidentally fall prey to the 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 1 raises some serious concerns. As noted in the previous section, 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 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 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".[3] 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, and that the verification tests performed would reflect this. However, as noted in Tables 1 and 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 to tests of these instabilities in various capacities.[4]

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-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 imple-

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

[4]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.

mentation (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 is incompatible with a strict V&V account of simulation verification. 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 numerical fidelity, 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.

3

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 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 isolation. 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 their 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 verification and validation of simulations should be distinct and sequential—that is to say, that verification should be (strictly speaking) a purely numerical/mathematical affair, and that any evaluations in terms of physics should be confined to the 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, 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.[5]

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, e.g., 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

[5]This criticism should be distinguished from another prominent critique of V&V, by Oreskes et al. (1994). Oreskes and collaborators argue that verification is (strictly speaking) impossible given that real-world systems are not closed systems, and advocate instead for a model of confirmation by degrees. I am not unsympathetic to the spirit of this position. However, my argument does not commit to their abstact hypothetico-deductivist picture of confirmation, and moreover aims to give concrete picture of how confidence-by-degrees is achieved in practice—and address the particular concerns about underdetermination that can be raised in by proponents of V&V.

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.

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 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.

Conclusion

In this paper, 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.

References

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

Agertz, O., B. Moore, J. Stadel, D. Potter, F. Miniati, J. Read, L. Mayer, A. Gawryszczak, A. Kravtsov, Å. Nordlund, et al. (2007). Fundamental differences between sph and grid methods. *Monthly Notices of the Royal Astronomical Society 380*(3), 963–978.

Balsara, D. S. and D. 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, A. and V. Springel (2012). Subsonic turbulence in smoothed particle hydrodynamics and moving-mesh simulations. *Monthly Notices of the Royal Astronomical Society 423*(3), 2558–2578.

Beckwith, K. and J. M. Stone (2011). A second-order godunov method for multi-dimensional relativistic magnetohydrodynamics. *The Astrophysical Journal Supplement Series 193*(1), 6.

Beisbart, C. (2019). What is validation of computer simulations? toward a clarification of the concept of validation and of related notions. In C. Beisbart and N. Saam (Eds.), *Computer Simulation Validation*, pp. 35–67. Springer.

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

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

Bryan, G. L., M. L. Norman, B. W. O'Shea, T. Abel, J. H. Wise, M. J. Turk, D. R. Reynolds, D. C. Collins, P. Wang, S. W. Skillman, et al. (2014). Enzo: An adaptive mesh refinement code for astrophysics. *The Astrophysical Journal Supplement Series 211*(2), 19.

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

Calder, A. C., B. Fryxell, T. Plewa, R. Rosner, L. Dursi, V. Weirs, T. Dupont, H. Robey, J. Kane, B. 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, J., J. Pittard, T. Hartquist, and S. 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, S. A. and R. H. White (1966). The hydrodynamic behavior of supernovae explosions. *The Astrophysical Journal 143*, 626.

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

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

Einfeldt, B., C.-D. Munz, P. L. Roe, and B. Sjögreen (1991). On godunov-type methods near low densities. *Journal of computational physics 92*(2), 273–295.

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

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

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

Frenk, C., S. White, P. Bode, J. Bond, G. Bryan, R. Cen, H. Couchman, A. 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. Timmes, M. Zingale, D. Lamb, P. MacNeice, R. Rosner, J. Truran, and H. Tufo (2000). Flash: An adaptive mesh hydrodynamics code for modeling astrophysical thermonuclear flashes. *The Astrophysical Journal Supplement Series 131*(1), 273.

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

Gresho, P. M. and S. 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.

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

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

Flash User Guide. `https://flash.rochester.edu/site/flashcode/user_support/`. Accessed: 2020-11-14.

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

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

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

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

Huang, J. and L. 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, J. H. (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.

Jun, B.-I., M. L. Norman, and J. M. Stone (1995). A numerical study of rayleigh-taylor instability in magnetic fluids. *The Astrophysical Journal 453*, 332.

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

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

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

Liska, R. and B. 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, P. and L. Del Zanna (2000). High-order upwind schemes for multidimensional magnetohydrodynamics. *The Astrophysical Journal 530*(1), 508.

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

McNally, C. P., W. Lyra, and J.-C. Passy (2012). A well-posed kelvin-helmholtz instability test and comparison. *The Astrophysical Journal Supplement Series 201*(2), 18.

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

Morrison, M. (2015). *Reconstructing reality: Models, mathematics, and simulations*. Oxford Studies in Philosophy.

Noh, W. 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, W. L. and C. J. Roy (2010). *Verification and validation in scientific computing*. Cambridge University Press.

Oreskes, N., K. Shrader-Frechette, and K. Belitz (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science 263*(5147), 641–646.

Orszag, S. A. and C.-M. Tang (1979). Small-scale structure of two-dimensional magnetohydrodynamic turbulence. *Journal of Fluid Mechanics 90*(1), 129–143.

Pakmor, R., V. Springel, A. Bauer, P. Mocz, D. J. Munoz, S. T. Ohlmann, K. Schaal, and C. Zhu (2016). Improving the convergence properties of the moving-mesh code arepo. *Monthly Notices of the Royal Astronomical Society 455*(1), 1134–1143.

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

Price, D. J., J. Wurster, T. S. Tricco, C. Nixon, S. Toupin, A. Pettitt, C. Chan, D. Mentiplay, G. Laibe, S. Glover, et al. (2018). Phantom: A smoothed particle hydrodynamics and magnetohydrodynamics code for astrophysics. *Publications of the Astronomical Society of Australia 35*.

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

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

Robertson, B. E., A. V. Kravtsov, N. Y. Gnedin, T. Abel, and D. H. Rudd (2010). Computational eulerian hydrodynamics and galilean invariance. *Monthly Notices of the Royal Astronomical Society 401*(4), 2463–2476.

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

Shu, C.-W. and S. Osher (1989). Efficient implementation of essentially non-oscillatory shock-capturing schemes, ii. In *Upwind and High-Resolution Schemes*, pp. 328–374. Springer.

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

Springel, V. (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, J. M. The athena code test page. `https://www.astro.princeton.edu/~jstone/Athena/tests/`. Accessed: 2020-11-30.

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

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

Stone, J. M. and M. 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, R. (2002). Cosmological hydrodynamics with adaptive mesh refinement-a new high resolution code called ramses. *Astronomy & Astrophysics 385*(1), 337–364.

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

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

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

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

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

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

Winsberg, E. (2018). *Philosophy and Climate Sciencde*. Cambridge University Press.

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

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

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