On Robustness in Cosmological Simulations

Marie Gueguen*†

Abstract

The Cold Dark Matter model faces many controversies at small scales, as simulations fail to reproduce the observed properties of dark matter (DM) haloes. Since rival DM models differ on their predictions about the structure of DM haloes, understanding their mass distribution is crucial for determining the nature of dark matter. However, only numerical approaches to determining what a hypothesis implies for the mass distribution are possible. Hence, simulations are a crucial part of evaluating DM rival models, as is assessing their reliability. I argue that robustness analysis is not a sufficient criterion for the trustworthiness of cosmological simulations.

1 Introduction

The standard cosmological model describes a nearly homogeneous early universe, where small density inhomogeneities evolve with time through gravitational collapse to form the large- and small-scale structures we now observe. An essential component of this cosmological model is a mysterious kind of matter called 'dark matter', that only interacts gravitationally. The Cold Dark Matter model (hereafter CDM) [1] predicts with great accuracy important large-scale structure properties but does not fare as well on small scales, where simulations fail to reproduce the observed abundance and demographics of

*To contact the author, please write to: Center for the Philosophy of Science, University of Pittsburgh, 1117 Cathedral of Learning, 4200 Fifth Ave, Pittsburgh 15260; email:mag395@pitt.edu.

†I am very grateful to Chris Smeenk for his support and advice. I would also like to thank Frank van den Bosch, Eric Winsberg, Craig Fox,the participants of the Rotman Summer Institute of Cosmology and an anonymous referee for very helpful feedback and discussion. This paper is based on work done while funded as a graduate student researcher under the John Templeton Foundation grant: "New Directions in Philosophy of Cosmology" (grant number 61048).

[1] 'Cold' means that particles move at non-relativistic speed.

DM haloes' structure. Since rival models such as cold, warm, or self-interacting dark matter agree on large scale, but differ on their predictions about the structure of DM haloes, understanding how mass is distributed in these haloes is crucial for determining the nature of dark matter. At such a scale though, only numerical approaches to determining what a hypothesis implies for the haloes' mass distribution are possible. Non-linear effects related to star formation and gas dynamics make it impossible to determine the mass distribution analytically. Hence, numerical simulations are needed to determine the mass distribution; and these simulations are a crucial part of evaluating the CDM model and various rival hypotheses.

Understanding in which case a simulation can succeed in (dis)confirming a model is, yet, still a challenge in cosmology. At small scale indeed, the CDM seems to do worse than its rivals: simulations predict, for instance, much more satellite galaxies within DM haloes than is actually observed. Prior to 1998, however, the problem was the exact inverse; simulated DM haloes did not present enough substructure compared to observations. When a model is so sensitive to modeling assumptions, what is the conclusion that should be drawn from a mismatch between simulation outcomes and observations? How can we assess whether this 'missing satellite' problem stems from numerical artifacts or constitutes a genuine failed prediction?

In biology and climate science, evaluation of when numerical evidence confirms a model and in what sense this confirmation must be understood has been based on robustness analysis (Levins 1966; Wimsatt 2012; Weisberg 2012)[2] In cosmology, astrophysicists have been relying on a similar methodology, according to which results that resist a change of values of the numerical parameters are considered trustworthy. In this paper, I will argue that robustness is not a sufficient criterion for determining when a prediction is reliable in N-body simulations, for equally robust but mutually exclusive predictions can obtain in N-body simulations. Even more worrying, robustness is sometimes a direct consequence of numerical artifacts.

[2]Although see Parker (2011) for an argument to the effect that robust predictions in climate science models are not sufficient for trustworthiness.

2 Robustness and Convergence Studies

The CDM model faces a number of problems at small scale. For example, it predicts way more substructure in a DM halo of the size of the Milky Way than is actually observed. Only 59 satellite galaxies seem to orbit our Milky Way, whereas several thousands are predicted by this model. Likewise, the density profile drawn from this model by Navarro et al. (1997) (hereafter NFW) predicts a steep, cuspy profile in the central region of the DM halo, with infinite density at the center. Yet, observations favor, especially for dwarf galaxies and low-surface-brightness galaxies, a 'cored' profile, with a flatter density profile as the radius tends toward zero. Observations for dwarf and low-surface-brightness galaxies are especially problematic because these galaxies are mostly made of dark matter, which means that this discrepancy cannot be washed away by adding baryonic physics to the simulations.[3]

Given that the predictions of the CDM model are not drawn from first principles, but from analytical fits to DM-only simulations, assessing the extent to which these two problems challenge this model is as difficult as it is urgent. How can we assess whether the discrepancy between the simulated systems and the observed ones stems from the physical model or from an erroneous code, whether the simulated outcome is altered by numerical artifacts or constitutes a genuine failed prediction? Astrophysicists rely on robustness analysis to decide when the outcome of a simulation is trustworthy. Robustness analysis has been first introduced by Levins (1966), as a way to assess the reliability of models in population biology in the absence of a background theory providing analytically soluble equations. According to Levins, a method is needed to evaluate the impact of the simplifications upon which models are based and to determine "whether a result depends on the essentials of the model or on the details of the simplifying assumptions" (1966, 423). This role is played by robustness analysis: by addressing the same problem with a diversity of models based on different 'lies', one can test whether these models agree on their predictions. Such agreement is taken to confirm the independence of the models from their characteristic simplifications: "Hence our truth is the intersection of independent lies" (1966, 423). Levins's account rely on a form of eliminative reasoning, according to which each model must exclude a given possibility. However, **?** have shown that this reasoning only yields valid inferences when an exhaustive

[3] A more exhaustive review of all the controversies arising at small scales for the CDM model can be found in Weinberg et al. (2015).

set of all possible models is examined. Such an exhaustivity is yet excessively difficult to achieve. Wimsatt further fleshed out this methodology by suggesting a four-step procedure for robustness analysis:

- To analyze a *variety* of *independent* derivation or measurement processes.
- To look for things which are *invariant* over the results of these processes.
- To determine the *scope* of the processes across which they are invariant and the *conditions* on which their invariance depends.
- To analyze and explain any relevant *failures of invariance* (2012, 62).

For Wimsatt, the purpose of robustness analysis is first to distinguish the "reliable from the unreliable"; second, to show the invariance of that which reliability (e.g., a prediction) is scrutinized over different models, in order to build confidence in their independence from these; and finally to determine the scope of this invariance. Wimsatt does not justify the robustness of a property *via* eliminative reasoning but through its overdetermination by independent models. Supposing that this independence can be qualified, such an account does not require to consider all possible models; but only models independent 'in an appropriate way'. In what follows, I will consider any procedure satisfying these Wimsattian features an instance of robustness analysis.[4]

Robustness analysis, in astrophysics, takes the form of 'convergence studies'. The idea is to determine whether unconstrained numerical parameters impact the outcome of simulations, by systematically varying their value and defining the value range under which the structure of the simulated halo remains unaffected by such variations. In that case, the halo is deemed 'appropriately resolved'. Such a procedure satisfies the characteristic features of robustness analysis listed above, and will thus be considered as such throughout this paper.

[4]More recently, Weisberg (2012) has suggested a procedure to establish robust theorems that consists in examining a group of models, searching for a 'robust property'; then finding the core structure giving rise to this property. However, the lack of modularity of simulations in cosmology undermines attempts to find the common structure responsible for the robust property, and the subsequent formulations of a robust theorem. I will therefore leave aside Weisberg's proposal.

The influential convergence study of Power et al. (2003) has contributed to set up the parametrization of N-body simulations for the last fifteen years. Their methodology consists in, first, simulating a large cosmological volume with low resolution, tracking the structure growth seeded by primordial density fluctuations. Then, they zoomed-in on some targeted haloes and re-simulated them at higher resolution. Based on this sample of haloes, several hundred of simulations (typically with a resolution of 32^3 particles) were run, allowing to survey the parameter space by varying the numerical parameters and draw preliminary convergence results. Finally the convergence criteria[5] were confirmed with another series of simulation of higher mass resolution–a series of run with 64^3 particles, and a few (given how expensive they are) with 128^3 and 256^3 particles. If, in the region of the parameter space defined by the convergence criteria, the predictions remain the same despite the increased resolution, then one can trust that they are independent of the numerical parameters' values. The convergence of the simulated mass profile is supposed to warrant its independence from numerical parameters and that it is not affected by artifacts.

3 Against Convergence

3.1 Convergence is not sufficient

Confidence in predictions about the abundance of DM subhaloes extracted is usually explained by the fact that they seem not affected by an increase of resolution above 50-100 particles per subhalo.[6] As mentioned above however, this prediction is very sensitive to modelling assumptions. Up to the close of the last century, simulations were suffering from an 'overmerging' problem, in that not enough substructure was predicted to match the observations. Several culprits had been proposed back then, with no consensus on the cause of the subhaloes disruption, but with an agreement that it was a *numeri-*

[5]That convergence should be reached in a time $t \leq 1.7\tau_r$, with τ_r the relaxation time is an example of such a convergence criterion, since the influence of artificial collisions between particles seems to produce a core profile after $1.7\tau_r$.

[6]See for instance **?**

cal problem. Moore et al. (1996) blamed inadequate force softening.[7] Carlberg (1994), on the other hand, argued that a low mass resolution could cause two-body heating and artificially enhance matter disruption. Since this problem was superseded by a 'missing satellite' problem as the resolution of simulations increased, no definitive conclusion was drawn about the cause of the overmerging. Yet, as shown by van den Bosch et al., subhaloes disruption is still "extremely prevalent in modern simulations, with [...] $\sim$ 65 percent of all subhaloes accreted around $z = 1$ [...] disrupted by $z = 0$" (2017, 2). The question thus arises: are N-body simulations still suffering from overmerging, or can they be considered reliable, based on convergence studies? Is the subhaloes disruption a physical mechanism or the result of numerical artifacts?

This question is addressed by van den Bosch and Ogiya (2018), in a paper also aiming at gaining a better understanding of the non-linear effects of tidal stripping on subhaloes.[8] Tidal processes are very difficult to describe, as the stripping of matter causes the subhalo remnant to fall out of virial equilibrium[9], and then to re-virialize by expanding, thereby provoking more stripping of matter, and to fall out of equilibrium once again. No analytical theory is available to describe such complicated processes of de- and re-virialization: only simulations can tell how the density distribution of the subhalo is affected by tidal stripping. It is thus crucial to disentangle what pertains to this stripping mechanism and what artificially results from an inadequate parametrization in simulations.

In order to do so, the authors came up with an idealized scenario where the *physical* hypothesis of tidal

[7]Due to limits in computational power, real DM particles are substituted in simulations by heavy particles. Hence, the gravitational force can generate very large, unphysical accelerations when two particles get very close to each other. Force softening is used to smooth the gravitational potential and suppress these accelerations below a typical distance–the 'softening length'.

[8]'Tidal stripping' refers to the escape of matter due to the tidal forces exerted by the host halo on the subhalo. Beyond some limit, the tidal forces exerted by the host overcome the gravitational force bounding the subhalo together, resulting in its dislocation.

[9]The virial theorem applied to celestial bodies states that the total energy of a system is equal to half its gravitational potential energy. If this equality does not hold, the system is either collapsing– the gravitational potential energy exceeds the kinetic energy– or expanding. A system for which this equality applies is considered in *virial equilibrium*.

stripping could be tested against the *numerical* hypothesis that matter disruption is mostly caused by inadequate force softening. They focused on the tidal evolution of an isolated subhalo on a circular orbit, given that, in such a scenario, only the host halo's tidal field and the numerical parameters can impact the subhalo disruption. Its isolation forbids high-speed encounters with other subhaloes causing more matter to disrupt; and the circular orbit excludes a fast pericentric passage that would deform the orbiting body, cause internal heating and thus more mass loss. The goal is to isolate two possible causes of the disruption and find the culprit by looking at the evolution of the bound mass of subhaloes when varying the strength of the tidal field and the force softening.

Figure 4 in van den Bosch and Ogiya (2018, 4071) suggests that increasing the strength of the tidal field by decreasing the orbital radius does not lead to the disruption of subhaloes, except in the extreme case where the orbital radius is chosen so that $r_{orb}/r_{vir,h} = 0.1$.[10] However, at $r_{orb}/r_{vir,h} = 0.1$, the subhalo totally disrupts after ~ 13 Gyr. This results concords with the results of the Millenium and the Bolshoi simulations according to which $r_{orb}/r_{vir,h} = 0.1$ is the typical radius at which subhaloes undergo disruption. Now, if this prediction is a *physical* prediction, as it is taken to be by most astrophysicists, it should not be significantly affected by variations in numerical parameters.

Now, look at the impact of varying force softening and mass resolution (van den Bosch and Ogiya 2018, figure 7, 4074). According to Power et al. (2003) for instance, the optimal value for the force softening ϵ_{opt} ranges between 0.02 and 0.06. Since a force softening ϵ too large results in smaller density distribution at small radii, and since less dense systems are more exposed to tidal stripping, for $\epsilon > \epsilon_{opt}$ one expects enhanced stripping and matter disruption. Likewise, a small softening is more exposed to two-body relaxation effects, that flatten the central density profile and enhance disruption. Thus, force softening should have no other consequence than increasing the mass loss. Surprisingly however, the simulations show the opposite: for $\epsilon < \epsilon_{opt}$, the bound remnants are larger and survive longer. One can suspect then that the physical disruption is very sensitive to the value assigned to ϵ.

[10] $r_{orb}/r_{vir,h} = 0.1$ is the orbital radius of the subhalo expressed in units of the host halo's virial radius, r_{orb} the distance between the centres-of-mass of the host and subhalo and $r_{vir,h}$ the virial radius, defined as the radius inside of which the average density is Δ_{vir}=97 times the critical density for closure.

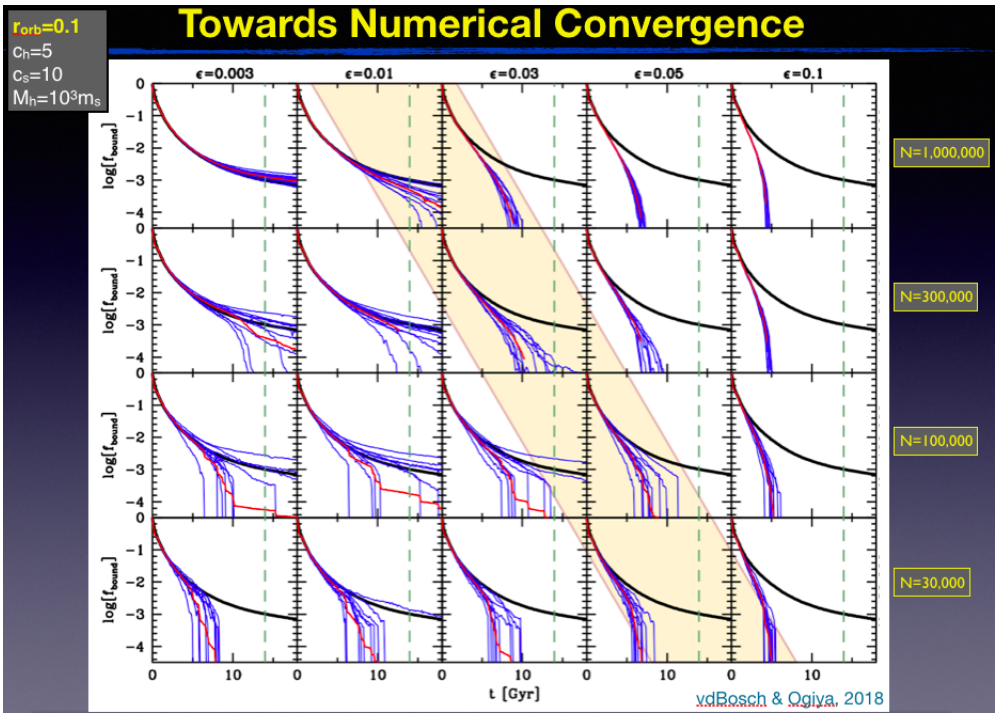

Figure 1: Bound fraction of mass as function of time for simulations with different mass resolution N_p and different force softening ϵ. The black line shows the 'converged' results of a simulation with $N_p = 10^7$ and $\epsilon = 0.003$; the blue line the results from 10 simulations; the red line their average. @copyright: F. van den Bosch, private correspondence. For a detailed discussion of this figure, See van den Bosch and Ogiya (2018), figure 10, 4077.

Consider now simulations at the orbital radius $r_{orb}/r_{vir,h} = 0.1$ when the mass resolution and the force softening are increased together. If you look for a region in the parameter space where the instability caused by mass resolution is kept under control[11] and a 'converged'[12] fraction of bound mass $f_{bound}(t)$ can be found, the upper-left corner of figure 2 gives you the closest result to such an ideal. But there, subhaloes do not disrupt after 13 Gyrs. On the contrary: a large fraction of bound mass survives. Note the contrast between this result at large N_p and small ϵ and the ones observed along the yellow-

[11]The sensitivity to discreteness noise is characterized by the authors using the variance σ_{logf} in $log(f_{bound})$.

[12]'Converged' here means that "(i) no significant changes occur when N_p is increased further, and (ii) the standard deviation in f_{bound} after one Hubble time is sufficiently small (i.e., $\sigma_{logf} \leq 0.05$)" (van den Bosch and Ogiya 2018, 4076).

shaded band, which corresponds to the scaling between mass resolution and force softening defined by Power et al. (2003). This scaling is obeyed by most of the state-of-the-art simulations since 2003. If one focuses on the red lines, i.e., on the averaged results of simulations, the bound fraction of mass appears converged: the red line prediction stays more or less the same despite increasing the mass and the force resolution, which is precisely why it is considered as robust by Power et al.'s standards. Yet, this red line indicates a full disruption of the subhalo between 5 and 8 Gyrs. Thus, simulations converge on predicting that subhaloes fully disrupt after 8 Gyrs but also on predicting that they survive after 13 Gyrs. What should alert the philosopher here is that convergence alone will not tell which one of these results is the correct one, or whether one of them is. Given the range of the optimal softening defined by Power et al. (2003), increasing the mass resolution and the force softening gives good agreement among simulations and confidence in the prediction that subhaloes at $r_{orb}/r_{vir,h} = 0.1$ will not survive after one Hubble time. But if another region of the parameter space is scrutinized, convergence is found on another prediction, that contradicts the former. As summarized by the authors of this study, convergence is "not a sufficient condition to guarantee that the results are reliable" (2018, 4067).

3.2 Pseudo-convergence or convergence?

In the previous subsection, I have argued that robustness analysis in the form of convergence studies is not sufficient to exclude numerical artifacts and that robust predictions cannot be considered reliable on such grounds. Here, I push this thought further and argue that convergence can *result* from artifacts, based on the work of Anton Baushev on the cusp-core problem.

In N-body simulations, real DM is represented by a limited number of heavy test bodies, to make the computational task tractable while preserving the averaged density of the system. However, whereas dark matter is collisionless, heavy test bodies undergo collisional effects, which in turn affect the density profile of DM haloes. These collisional effects are characterized by the relaxation time $\tau_r = \frac{N(r)}{8ln\Lambda.\tau_d}$, with N(r) the number of test bodies inside a sphere of radius r, $ln\Lambda$ the Coulomb logarithm and τ_d the characteristic dynamical time of the system at radius r. This is why Power et al. (2003) recommend, as one of their convergence criteria, that convergence be reached in a time $t \leq 1.7\tau_r$.[13]

[13]See footnote 5.

Baushev, however, has shown that cuspy profiles are generated by an intensive energy relaxation. This fact raises important questions. DM haloes undergo violent relaxation when they collapse, as density inhomogeneities create small-scale gravitational fields mediating the exchange of energy among DM particles. But they do so only at the moment of the collapse. The halo, once formed, has a stationary gravitational field. How come then that N-body simulations based on the CDM model predicts cuspy density profiles for formed haloes? If the cuspy profiles stems from collisionality, where does it come from? It is artificial or physical?

Baushev et al. (2017) addresses this question with a methodology very similar to that of van den Bosch and Ogiya (2018). First, they propose an idealized scenario where all sources of collisionality are turned off except the one under scrutiny. Thus, they simulate an isolated halo, to avoid the tidal influence of nearby haloes and the gravitational capture of more mass from the surrounding, potentially leading to a secondary violent relaxation. They chose a halo with a Hernquist density profile that behaves like the NFW profile in the central region but is fully stable. Given that such a halo has a constant gravitational potential field $\phi(r)$, the density and velocity profiles should remain the same and all (the implicit functions of) the integrals of motion such as the specific energy $\omega = \phi(r) + \frac{v^2}{2}$, the specific angular momentum $\vec{K}$, and the apocenter distance r_0[14] should be conserved. Thus, any variation of them must be due to numerical effects. Such symptoms of artificial collisions can be tracked in their simplified scenario. Figure 2 of their paper (2017, 6) shows the variations of the integrals of motion as a function of radius over 200 snapshots and confirm that they vary significantly even in one single timestep. Thus, the convergence criterion of Power et al. (2003) by no means guarantees that all sources of numerical artifacts have been excluded.

This preliminary conclusion raises another interesting question: why does the density profile remains stable if the integrals of motion vary? In other words, is the stability of the cuspy profile a mere coincidence, or the result of the numerical effects observed above? If the system is collisional, then it is better modelled by the Fokker-Planck equation, that models dynamical friction and diffusion, than by the collisionless Boltzmann equation usually appealed to. In that case, the diffusion streams created by

[14] r_0 is the maximum distance on which the particle can move off the center and depends only on the integrals of motion: $\omega = \phi(r_0) + K^2/2r_0$.

the particle interactions could contribute in forming a stable density profile. And, as it turns out, the Fokker-Planck equation does have a stationary solution close to the NFW one. If the cusp is a product of this Fokker-Planck diffusion, then artificial collisions should form a downward and an upward stream of particles, with increasing and decreasing r_0 respectively, that compensate each other and thereby explain the stability of the profile. This prediction can be tested by taking two adjacent snapshots S_1 and S_2 and calculating, for an array of radii r the number of particles $\Delta N_+(r)$ of particles which had $r_0 < r$ at S_1 and $r_0 > r$ at S_2 and the number of particles $\Delta N_-(r)$ of particles which had $r_0 > r$ at S_1 and $r_0 < r$ at S_2. Figure 4 in (2017, 9) confirms that the Fokker-Planck streams are present, compensate each other very well outside a given radius, and are important enough to shape the density profile.
In sum, the variations of the integrals of motion unambiguously demonstrate that converged N-body simulations still suffer from artificial collisionality. Convergence fails to identify results independent of artifacts. Moreoever, confidence in the density profile is not warranted by its stability, since this stability is itself produced *by virtue* of numerical artifacts.

3.3 Discussion

Although convergence failed, in the two cases discussed above, in diagnosing artifacts, one could defend robustness by arguing that van den Bosch and Ogiya and Baushev et al.'s studies exemplify the fourth step of Wimsatt's procedure–that they demonstrated the unreliability of simulations based on the breaking down of their robustness.
These two papers cannot, however, be considered instances of robustness analysis, for at least two reasons. Let us consider first the target of van den Bosch and Ogiya (2018). What the authors examined is not the trustworthiness of the simulations outcomes, but that of the convergence criterion itself. The main conclusion of this paper is that "most, if not all, disruption of substructure in N-body simulations is numerical in origin", a conclusion that "questions whether the fact that subhalo mass functions appear to be converged down to 50-100 particles per subhalo implies that results are reliable" (2017, 4084). In other words, the authors showed that, even though convergence was reached, numerical artifacts had not been excluded; and that the results backed up through convergence were still not reliable. What is at stake here is not whether the amount of substructure predicted through N-body simulations

is a reliable prediction, but whether the reason it is taken to be a reliable prediction is a sound one. Still, one could insist that this study makes a negative use of robustness analysis, i.e., that it uses robustness to show that predictions made about the number of satellite galaxies that should be observed in haloes are not reliable, because their apparent robustness breaks down when the resolution is increased further. However, this reconstruction of their argument is misleading: what we see in their paper is not that the appearance of convergence of the results based on Power et al. breaks down, while 'true' convergence is reached for higher resolution. What we see instead is a first instance of convergence for a given range of values, and a second one for another range of values. Asserting that the robustness of the first *breaks down* when increasing the resolution is assuming that we have more reasons to trust the second set of results than the former. However, based on convergence only, we have equally good reasons to accept one result or the other –and, reversely, equally good reasons to reject them. Convergence alone will not tell us which result to trust. But here is where the problem becomes particularly sharp: in the astrophysical context at least, no other criterion is available to back up or supplement the results of robustness analysis. Simulations are used to determine what rival dark matter models tell us about the structure of the universe precisely because non-linear effects related to star formation and gas dynamics make it impossible to determine the distribution of matter in dark matter haloes analytically. Thus, there is no analytical solution against which the results of simulations can be tested. Furthermore, a comparison between observations of 'real' dark matter haloes and simulations is of no help, for at least two reasons. While the similarity between the model and the real world is often used to validate the former, what is under scrutiny here is precisely the pertinence of such a comparison given that we do not know whether the simulated outcome reliably tracks the predictions of the CDM model or crucially depends on the parametrization process. Second, parametrization –or 'calibration', or 'tuning'– is also the process through which one improves the agreement with real world data when other processes of validation are not feasible (Oberkampf 2014, 33). How could then one appeal to the agreement between observations and simulations when trying to circumscribe the impact of calibration on the simulation outcome? But in the absence of any other criterion to assess reliability, how could one conclude anything from the fact that two mutually exclusive but converged results can be found in two different regions of the parameter space?

Even more undoubtedly, robustness analysis has no role to play in Baushev et al.'s result. Like that of van den Bosch and Ogiya's, the aim of the paper is not to test the reliability of the NFW density profile *per se*, but to test whether its robustness says anything about its reliability. This goal does not involve any attempt to make the stability of the profile break down. On the contrary, their argument shows that its robustness will *not* break down, because it results from numerical artifacts. Simulations converge because the artificial diffusion streams generated by numerical artifacts compensate each other. The lesson to be drawn from this paper is that robustness by no means says anything about the *physical* nature of the prediction made, and even less about its reliability.

4 Conclusion

I have argued that robustness analysis, in the form of convergence studies, fails to exclude numerical artifacts in N-body simulations and thus to warrant their reliability. Simulations in astrophysics constitute a very exciting opportunity for philosophers to explore the limits of robustness and to suggest possible rivals better suited for cosmological simulations.'*Crucial*' simulations, where a *numerical* hypothesis about the origin of a prediction is tested against a *physical* explanation, used both by van den Bosch and Ogiya and Baushev et al., constitute a possible candidate for this task.

References

Baushev, A., L. del Valle, L. Campusano, A. Escala, R. Muñoz, and G. Palma (2017). Cusps in the center of galaxies: a real conflict with observations or a numerical artefact of cosmological simulations? *Journal of Cosmology and Astroparticle Physics 2017*(05), 042.

Carlberg, R. (1994). Velocity bias in clusters. *The Astrophysical Journal 433*(468).

Levins, R. (1966). The strategy of model building in population biology. *American scientist 54*(4), 421–431.

Moore, B., N. Katz, and G. Lake (1996). On the destruction and over-merging of dark halos in dissipationless n-body simulations. *The Astrophysical Journal 457*, 455–459.

Navarro, J. F., C. S. Frenk, and S. D. White (1997). A universal density profile from hierarchical clustering. *The Astrophysical Journal 490*(2), 493.

Oberkampf, W. L. (2014). Verification and validation in computational simulation. http://www.psfc.mit.edu/ttf/2004/talks/oberkampf.pdf.

Parker, W. S. (2011). When climate models agree: The significance of robust model predictions. *Philosophy of Science 78*(4), 579–600.

Power, C., J. Navarro, A. Jenkins, C. Frenk, S. D. White, V. Springel, J. Stadel, and T. Quinn (2003). The inner structure of ΛCDM haloes—I. A numerical convergence study. *Monthly Notices of the Royal Astronomical Society 338*(1), 14–34.

van den Bosch, F. C. and G. Ogiya (2018). Dark matter substructure in numerical simulations: a tale of discreteness noise, runaway instabilities, and artificial disruption. *Monthly Notices of the Royal Astronomical Society 475*(3), 4066–4087.

van den Bosch, F. C., G. Ogiya, O. Hahn, and A. Burkert (2017). Disruption of dark matter substructure: fact or fiction? *Monthly Notices of the Royal Astronomical Society 474*(3), 3043–3066.

Weinberg, D. H., J. S. Bullock, F. Governato, R. Kuzio de Naray, and A. H. G. Peter (2015). Cold dark matter: Controversies on small scales. *Proceedings of the National Academy of Sciences 112*(40), 12249–12255.

Weisberg, M. (2012). *Simulation and similarity: Using models to understand the world*. Oxford University Press.

Wimsatt, W. C. (2012). Robustness, reliability, and overdetermination (1981). In *Characterizing the Robustness of Science*, pp. 61–87. Springer.