

The effect of local universe constraints on halo abundance and clustering

Maxwell L. Hutt^{1*}, Harry Desmond^{1,2,3†}, Julien Devriendt¹ and Adrianne Slyz¹

¹*Astrophysics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford, OX1 3RH, UK*

²*McWilliams Center for Cosmology, Department of Physics, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213*

³*Institute of Cosmology & Gravitation, University of Portsmouth, Dennis Sciama Building, Portsmouth, PO1 3FX, UK*

29 March 2022

ABSTRACT

Cosmological N -body simulations of the dark matter component of the universe typically use initial conditions with a fixed power spectrum and random phases of the density field, leading to structure consistent with the local distribution of galaxies only in a statistical sense. It is, however, possible to infer the initial phases which lead to the configuration of galaxies and clusters that we see around us. We analyse the CSiBORG suite of 101 simulations, formed by constraining the density field within 155 Mpc/ h with dark matter particle mass $4.38 \times 10^9 M_\odot$, to quantify the degree to which constraints imposed on 2.65 Mpc/ h scales reduce variance in the halo mass function and halo–halo cross-correlation function on a range of scales. This is achieved by contrasting CSiBORG with a subset of the unconstrained Quijote simulations and expectations for the Λ CDM average. Using the FOF, PHEW and HOP halo finders, we show that the CSiBORG suite beats cosmic variance at large mass scales ($\gtrsim 10^{14} M_\odot/h$), which are most strongly constrained by the initial conditions, and exhibits a significant halo–halo cross-correlation out to ~ 30 Mpc/ h . Moreover, the effect of the constraints percolates down to lower mass objects and to scales below those on which they are imposed. Finally, we develop an algorithm to “twin” halos between realisations and show that approximately 50% of halos with mass greater than $10^{15} M_\odot/h$ can be identified in all realisations of the CSiBORG suite. We make the CSiBORG halo catalogues publicly available for future applications requiring knowledge of the local halo field.

Key words: large-scale structure of the universe – dark matter – galaxies: halos – galaxies: statistics – software: simulations

1 INTRODUCTION

Undetectable except by its gravitational effect, dark matter (DM) is thought to be the dominant matter component of the universe. Maps of large volumes of the universe find galaxies to be arranged in a web-like structure of clusters, filaments and voids, tracing the distribution of the underlying dark matter density field. Understanding the evolution of this density field and its connection to visible tracers are currently key tasks in astrophysics.

Simulations of DM with initial conditions (ICs) based on a fixed power spectrum and random phases of the primordial matter field have been well studied for decades (Press & Schechter 1974; Davis et al. 1985; Lacey & Cole 1993; Eisenstein & Hut 1998; Tinker et al. 2008). Only more recently however have attempts been made to constrain the phases as well (i.e. the full 3D initial density field) so as to produce structures with positions matching those observed in the local universe. This subfield originated with Hoffman & Ribak (1991), who developed a method for constructing

constrained realisations of Gaussian random fields. Early applications of this algorithm developed galaxy count (Kolatt et al. 1996; Bistolas & Hoffman 1998) and peculiar velocity observations (van de Weygaert & Hoffman 2000; Klypin et al. 2003; Kravtsov et al. 2002) as means of deriving the constraints. Subsequent refinements increased the resolution of the simulations and the precision of the gravity model, improved the modelling used to generate the ICs (e.g. the galaxy bias), sampled the ICs rather than generating them in a maximum-likelihood sense, and implemented additional physics. Galaxy groups have also been used for deriving the constraints (Wang et al. 2014). A focus of constrained simulations has been the properties and formation history of the Local Group (Yepes et al. 2009; Sawala et al. 2022; McAlpine et al. 2022), although they are finding use in an increasingly wide range of science applications.

It was realised already a decade ago (Jasche et al. 2010; Jasche & Wandelt 2012, 2013) that deriving the ICs of the local universe is best viewed as a Bayesian inference problem and solved by forward modelling. The method for doing this—which expanded to include full models for galaxy bias, redshift space distortions and observational selection effects—became known as the *Bayesian Origin Reconstruc-*

* E-mail: maxwell.hutt@new.ox.ac.uk

† E-mail: harry.desmond@port.ac.uk

tion from Galaxies (BORG) algorithm, which has been applied to the 2M++ (Lavaux & Hudson 2011; Lavaux & Jasche 2016; Jasche & Lavaux 2019) and SDSS-III/BOSS (Lavaux et al. 2019) galaxy fields. This uses an efficient Hamiltonian Markov Chain Monte Carlo algorithm to sample the ICs at $z = 1000$ as well as the nuisance parameters. Although other constrained simulations have been run using ICs derived with similar techniques (e.g. Wang et al. 2014; Tweed et al. 2017; Wang et al. 2016), there has not before now been a statistical ensemble of such simulations that fully samples the IC posteriors. This is needed to propagate the uncertainties in the IC inference into the low-redshift density field at high resolution, and hence quantify the extent to which our present knowledge of the local universe’s ICs permits us to localise structures of various masses on various scales and reduce cosmic variance in their N -point statistics. Such an ensemble is also necessary to marginalise over the IC uncertainties in post-processing applications.

To fill this gap, we analyse CSiBORG (*Constrained Simulations in BORG*), a suite of constrained simulations with ICs derived from the 2M++ particle mesh BORG reconstruction (Jasche & Lavaux 2019). The 101 realisations of the suite span the converged part of the BORG posterior, allowing us to investigate the effect of the uncertainty it introduces. CSiBORG has been used previously to test galileon modified gravity in Bartlett et al. (2021) and to identify local voids-as-antihalos in Desmond et al. (2022).

Our focus here is on quantifying the effects of the constraints on the variance in halo abundance and clustering (i.e. their uncertainty), as a function of mass and scale, while accounting fully for the BORG IC uncertainties. This is important to understand the exact manner in which constraints on one scale affect structure on others, an important issue for all constrained simulations with ICs specified on a certain scale or set of scales. While the constraints will not affect the finite-volume uncertainty in halo statistics (for fixed box size), they will reduce shot noise on all scales because pinning down larger-scale modes will induce correlations on smaller scales. We assess this by comparing the CSiBORG statistics with those of Quijote (Villaescusa-Navarro et al. 2020), an unconstrained suite with otherwise similar properties. In particular we address the following questions:

- What is the impact of the BORG constraints on the variance in the halo mass function (HMF) between realisations of the local universe?
- How do the constraints propagate into the halo–halo cross-correlation function on a range of scales?
- Which halos should be considered “the same” between the different local-universe realisations, and how does the fraction of matched halos depend on the properties of the halos?
- How do these results depend on the halo-finder used (Parallel HiErarchical Watershed (PHEW, Bleuler et al. 2015), friends-of-friends (FOF, Davis et al. 1985), or HOP (Eisenstein & Hut 1998))?

This paper is structured as follows. In Sec. 2 we describe the CSiBORG and Quijote simulation suites. In Sec. 3 we present the halo statistics that we investigate, describe the halo-finders used and document our method for matching halos between realisations. In Sec. 4 we compare the variance in the HMF and halo–halo cross-correlation function of CSiBORG to

those of Quijote and analytic expectations, and show the results of our matching procedure. In Sec. 5 we discuss systematic uncertainties and summarise our results.

2 SIMULATED DATA: CSiBORG & Quijote

The CSiBORG suite comprises 101 realisations of a 677.7 Mpc/h box centred on the MW with ICs inferred from the 2M++ survey (Lavaux & Jasche 2016), spanning the BORG posterior. Within this simulation volume, discretised onto a 256^3 grid, the BORG algorithm uses a particle-mesh gravity solver (Jasche & Lavaux 2019) to evolve the DM particle field from an initial redshift of $z = 1000$ to the present day. This constrains the ICs on a grid of cell length 2.65 Mpc/h.

The density field is well constrained in the volume in which the 2M++ survey has high completeness, a spherical volume with radius ~ 155 Mpc/h around the MW. In this sub-volume the ICs, propagated linearly to $z = 69$, are augmented with white noise on a 2048^3 grid, giving a spatial resolution of 0.33 Mpc/h and a minimum DM particle mass of $3.09 \times 10^9 M_\odot/h$. Surrounding this zoom-in region, there is a buffer region in which the DM particle mass increases for distances further from the MW. This buffer extends approximately 10 Mpc/h in order to smoothly connect the high resolution central region to the outer region with the same grid resolution as that of BORG. The adaptive mesh refinement code RAMSES (Teyssier 2002) is then used to evolve the DM density field to $z = 0$. Only the central augmented region is refined (reaching level 18 by $z = 0$, a spatial resolution of 2.6 kpc/h), so it is only there that we consider halos. The cosmology of the simulations is $T_{\text{CMB}} = 2.728$ K, $\Omega_m = 0.307$, $\Omega_\Lambda = 0.693$, $\Omega_b = 0.04825$, $H_0 = 70.5 \text{ km s}^{-1} \text{ Mpc}^{-1}$, $\sigma_8 = 0.8288$ and $n = 0.9611$.

To quantify the effects of the constraints we compare with simulations from the Quijote suite. Each of these is an N-body simulation evolved forward in time from unconstrained ICs at $z = 127$ to $z = 0$, using the GADGET-III code (Springel et al. 2008). The volume of each Quijote box is 1 (Gpc/h)^3 . We use the 100 realisations with 1024^3 resolution, giving a particle mass of $8.72 \times 10^{10} M_\odot/h$, and their “fiducial cosmology” ($\Omega_m = 0.3175$, $\Omega_\Lambda = 0.6825$, $\Omega_b = 0.049$, $H_0 = 67.11 \text{ km s}^{-1} \text{ Mpc}^{-1}$, $\sigma_8 = 0.834$ and $n = 0.9624$), with the phases of the ICs varying randomly between realisations at fixed power spectrum. This subset is chosen as its resolution is high enough to compare to CSiBORG over a broad mass range.

To isolate the effect of the IC constraints, we must mitigate the impact of the different cosmological parameters used by CSiBORG and Quijote. We do this by running two simulations with unconstrained ICs drawn from a Gaussian density field at $z = 69$ with the same resolution, box size and refined sub-volume as the CSiBORG simulations. One of these has the cosmology of CSiBORG and the other that of Quijote. Again, we consider only halos located in the central 155 Mpc/h “zoom-in” sphere. These boxes link the two simulation suites since they have identical resolution and ICs and differ only in their cosmological parameters.

3 METHOD

In this section we describe the methods used for halofinding, the statistics that we compare between the simulations described above and the procedure for matching halos between realisations.

3.1 Halofinders

We run three halofinders on the CSiBORG boxes and the two additional unconstrained simulations: PHEW, FOF and HOP. For Quijote we use only the FOF algorithm that has already been run on it by Villaescusa-Navarro et al..

The watershed halofinder PHEW was run on-the-fly with RAMSES for the CSiBORG simulations. This algorithm splits the density field into basins by locating watersheds (boundaries between regions which share a common local minimum density via steepest descent). These basins are sequentially merged according to user-defined thresholds on the density of saddle points between basins. We use the standard threshold value from Bleuler et al. of $200\rho_c$ to merge sub-halos into halos.

For FOF we adopt a linking length parameter of $b = 0.2$. This matches the choice made for the Quijote simulations by Villaescusa-Navarro et al.. FOF links together particles within a given “linking distance” of each other, defined as b times the mean inter-particle separation. It is known that FOF tends to connect extraneous particles to halos along nearby filaments of the density field (Eisenstein & Hut 1998), creating more massive structures than other algorithms. This is particularly true for halos which are merging near $z = 0$ as FOF is more likely to cluster them into a single object than other algorithms, an effect similar to the “bridged” halos discussed by Lukić et al. (2009). Warren et al. (2006) also find that FOF halos are biased towards larger masses for halos with low particle number and propose a correction to the particle number of the form $N_{\text{corrected}} = N_0(1 - N_0^{-0.6})$ for this effect, with N_0 the original number of DM particles in the halo. Marginally decreasing the particle number in this way was found to eliminate the effect of resolution on halo mass. Lukić et al. have shown that corrections of this sort must depend on halo concentration as well as particle number. For our purpose of quantifying the variance between realisations, however, this level of precision of the HMF is not necessary and we do not use such corrections.

Finally, we run HOP. This algorithm first generates an estimate of the local density at each particle position using a kernel over a number N_{dens} of neighbouring particles. Then it associates to each particle the highest density particle from its nearest $N_{\text{hop}} - 1$ neighbours. This association is repeated until a local maximum of the density field is reached. All particles above a certain density threshold ρ_t which lead to the same local maximum are then grouped together. We run HOP with parameters $N_{\text{hop}} = N_{\text{dens}} = 20$ and $\rho_t = 80$, which produces similar results to FOF with $b = 0.2$ (see Tweed et al. 2009 for a summary of these parameters in the AdaptaHOP Aubert et al. 2004 algorithm, which reduces to HOP when substructure is neglected).

In order to reduce resolution errors in halo properties we require that halos found by any halofinder must contain at least 100 particles. This corresponds to a minimum halo mass of $3.09 \times 10^{11} M_\odot/h$ for CSiBORG and $8.72 \times 10^{12} M_\odot/h$

for Quijote. However, the lower resolution of the Quijote simulations means that this constraint would severely limit the mass range across which halos are found compared to CSiBORG. Hence, we include halos from the Quijote suite with a minimum of 20 DM particles when it is desirable to extend the comparison with CSiBORG to lower masses.

3.2 The halo mass function and halo–halo cross-correlation function

We compare halo distributions via two main summary statistics: the HMF, defined by

$$n(M) \equiv \frac{dn}{d \log M} \quad (1)$$

where log is the base-10 logarithm throughout, and the two-point halo–halo cross-correlation function between pairs of simulation boxes. Denoting the halo density field by $\rho(\mathbf{x})$ and its mean by $\bar{\rho}$, we measure the cross-correlation

$$\xi_{ij}^{(hh)}(\mathbf{r}) = \frac{1}{V} \int d^3\mathbf{x} \delta_i(\mathbf{x}) \delta_j(\mathbf{x} - \mathbf{r}) \quad (2)$$

where the subscripts label the pair of simulation boxes being compared and

$$\delta_i(\mathbf{x}) \equiv \frac{\rho_i(\mathbf{x})}{\bar{\rho}_i} - 1 \quad (3)$$

is the halo overdensity field. Assuming an isotropic universe, $\xi(\mathbf{r}) = \xi(r)$ and we can estimate the cross-correlation by counting pairs of halos separated by distances which fall into a given set of radial distance bins and adopt the Landy-Szalay estimator (Landy & Szalay 1993)

$$\hat{\xi}_{ij}^{(hh)}(r) = \frac{D_i D_j - D_i R - D_j R + RR}{RR} \quad (4)$$

where D_i represents the catalogue of data points (halos) in simulation i and R is a catalogue of random points. That is, $D_i D_j(r)$ is calculated as the number of pairs of data points with separation, r , falling in a given distance bin, and similarly for the other terms. We use 100 times as many random points as data points so that the uncertainty in the estimator itself (which is at the Poisson level) is negligible compared to the variance between realisations. We increase the statistical significance by averaging the cross-correlation function over all pairs of realisations in a given suite (CSiBORG or Quijote), and denote the averaged quantity simply as $\langle \xi(r) \rangle$.

The HMF is calculated in each box using 20 bins spaced evenly in $\log(M/(M_\odot/h))$. Using a larger number of bins has no impact on the results at the low-mass end but decreases the visibility of trends at the high-mass end where shot noise variance is large as there are fewer halos. For each suite we then find the mean and standard deviation of the HMF, as well as the correlation coefficients between the halo counts in each mass bin (normalised covariance), over all the realisations.

The variance of statistics calculated from N -body simulations, often called “cosmic variance”, has several origins (see Colombi et al. 2000; Colombi et al. 1994; Moster et al. 2011 and references therein). The most important are (a) shot noise due to the finite number of objects in the simulation, (b) finite-volume uncertainty due to underlying density modes on scales larger than the simulation volume, and (c) edge effect uncertainty, which is related to the geometry of the survey in

that objects near its edge carry less weight than those far from the edge (Szapudi & Colombi 1996). The latter two sources of variance are negligible compared to shot noise at scales approaching homogeneity (~ 500 Mpc/h; Schneider et al. 2016). The ICs of CSiBORG will not influence finite-volume or edge-effect uncertainties over the entire simulation box, but the density field constraints will reduce shot noise. By reducing variance in the density modes at and above the IC scale (~ 2.65 Mpc/h), the constraints do however induce correlations on smaller scales which effectively reduce finite-volume variance for them too. We will quantify the extent to which constraints determine structure both within and beyond the range of scales on which they are directly imposed.

Since the full Quijote volume of 1 Gpc/h is well beyond the homogeneity scale, those boxes are not subject to such large finite-volume uncertainties and are therefore dominated by shot noise. However, given that the refined 155 Mpc/h region of the CSiBORG simulations is far below it, we cannot use the same method to isolate the shot noise contribution to the CSiBORG HMF variance. We proceed by using the full Quijote volume to verify that the shot noise uncertainty is Poissonian, allowing us to infer the shot noise contribution within the refined region.

The logic is as follows. First, we measure the standard deviation σ_{155} of the HMF of both the CSiBORG and Quijote simulations within $V_{155} = 4\pi(155 \text{ Mpc/h})^3/3$, such that finite-volume and edge-effect uncertainties are held constant. Second, we measure the Quijote HMF in the full $V_{1000} = (1 \text{ Gpc/h})^3$ volume for all the Quijote simulations and calculate its standard deviation σ_{1000} . This allows us to confirm that the HMF measured in the smaller volume is indeed consistent with that found over the whole box, and to estimate the isolated shot noise uncertainty. We can then verify that the shot noise uncertainty is Poissonian by computing the expected Poisson variance in V_{155} , namely $(N_{\text{halo}})^{-1/2}$, and comparing it with the inferred variance from the full Quijote volume, $\sigma_{1000}\sqrt{V_{155}/V_{1000}}$. N_{halo} is the average number of halos in the 155 Mpc/h region over all realisations. If these two measurements of the shot noise are consistent then we can be confident that the shot noise error in V_{155} is given by $(N_{\text{halo}})^{-1/2}$ and we can assume this for CSiBORG also. Thus we isolate the shot noise contribution to both the Quijote and CSiBORG HMF variances and can identify whether the constraints on CSiBORG decrease the variance below the Poissonian expectation.

We compare the HMF to a standard function from the literature, Warren et al. (2006, hereafter W06). We choose this particular analytic form because it was originally fitted to the results of a FOF halofinder on a wide range of unconstrained N -body simulations. Other works (Despali et al. 2015) have shown that improved parametrisations and halo definitions lead to HMFs which are $\sim 30\%$ discrepant from W06 at large halo masses. We do not, therefore, expect this curve to be fully compatible with our results, but it does allow us to compare the CSiBORG and Quijote HMFs to a standard often seen in the literature.

3.3 Matching halos between realisations

We now describe a method for identifying halos that are sufficiently similar in mass and location between all realisations of a simulation suite to be called the “same” object. The halos

that are then present in *all* realisations are particularly robust and should potentially be prioritised in post-processing applications. For this, we use the FOF halo catalogues of CSiBORG and consider halos with masses greater than $10^{13} M_{\odot}/h$ to enable comparison with Quijote over the same mass range. The procedure is as follows:

- (i) Select a reference simulation and record the positions of all halos in it.
- (ii) Select a halo in the reference simulation and calculate its size according to $R_{200} = (3M/(4\pi \cdot 200\rho_{\text{crit}}))^{1/3}$.
- (iii) In all other simulations, locate all halos within a “search radius” of NR_{200} of the position of the reference halo selected in the previous step and select the one most similar in mass. If there are any simulations in which *no* halos are located within NR_{200} of the reference halo position, discard that halo as it is not consistently present in all simulations. This results in a list of the halos which are “matched” to the reference halo in all the other simulations. Our fiducial value for N is 5.
- (iv) Calculate the variance in the mass and position (i.e. distance from the reference halo position) of the matched halos across all simulation boxes. This gives a measure of the success of the matching for this particular halo in the reference simulation.
- (v) Repeat from step (ii) until all the halos in the reference simulation have been processed.
- (vi) Repeat from step (i) until all simulation boxes have been used as the reference simulation.

We show the result of this procedure in Fig. 7 by plotting the variance in the mass and position of all the matched halos as a function of the reference halo mass, with the expectation that larger halos will be more reliably matched between simulations because they are better constrained by the CSiBORG ICs. We also run this procedure on the Quijote suite, in order to determine the proportion of halos one would expect to find suitable matches for in the absence of IC constraints.

We choose a search radius proportional to the virial radius of the reference halo to alleviate bias between halos of different sizes. That is, a given halo is subject to variations in location between realisations on the scale of its size due to fluctuations in the modes of the density field with wavelengths of the same order. Hence, a massive (and therefore large) halo is expected to fluctuate more in position than a low mass one. Therefore, choosing a fixed search radius would introduce a strong bias in favour of less massive halos. The exact ratio of the virial and search radii, N , is somewhat arbitrary but does not qualitatively affect the salient results: the variance in position and mass of the matched halos, the difference in the number of matched halos in CSiBORG and Quijote, and the dependence of this number on mass. We implement the procedure for $N = 3, 5, 10$ to verify this. Note that the search radius around a reference halo with mass $10^{13} M_{\odot}/h$ is $5R_{200} \approx 1.8$ Mpc/h whereas for a very massive $10^{15} M_{\odot}/h$ halo, it is approximately 8 Mpc/h. Hence, for the smallest halos being considered, we expect far greater variance in the matching since the search radius is below the IC constrained scale. Nevertheless, identifying more successful matches at low masses in CSiBORG than Quijote is a strong indicator of the induced correlation on scales below those set by the CSiBORG constraints.

There is, however, a subtle residual bias in the matching

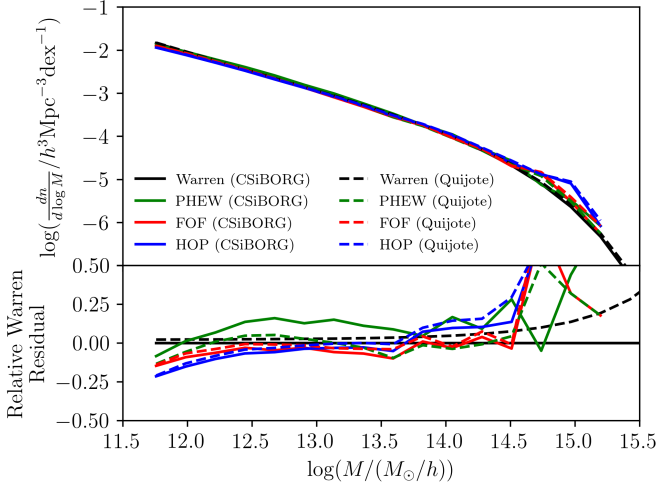


Figure 1. *Upper panel:* HMF of the pair of simulations with identical unconstrained ICs but differing in their cosmological parameters calculated with different halo finders. The solid lines use the CSiBORG cosmology and the dashed lines the Quijote cosmology. (Note that none of these boxes are part of the actual CSiBORG or Quijote suite; the labels “Quijote” and “CSiBORG” refer merely to the cosmology.) The black curves show the HMF from W06. *Lower panel:* Relative residual from W06 using the CSiBORG cosmology.

procedure towards smaller halos since within a given search radius in any realisation, the probability to find a small halo is far greater than that of a large halo simply because there are more small halos than large ones. Consider the case of a very massive reference halo. When searching for a match in another simulation, it is possible that there is no halo of similar mass within the search radius. In this case, the procedure is more likely to match this reference halo to a small halo which is found there. Hence, the mean mass of the matched halos tends to be slightly lower than the mass of the reference halo. In our application, there are negligibly many ($\mathcal{O}(0.1\%)$) halos where the mean mass of matched halos differs by more than a factor of 5 from the reference halo mass and hence we have not imposed any filters on the matched halos. In general, one may reduce this bias by imposing a threshold variance on the mass of the matched halos, or on the ratio of the mean matched halo mass to the reference halo mass.

4 RESULTS

4.1 CSiBORG & Quijote HMFs

We begin our investigation of the HMF by quantifying the shift in it due to the different cosmologies of CSiBORG and Quijote. Fig. 1 shows the HMFs of the two unconstrained simulations performed using the CSiBORG resolution with identical ICs but different cosmologies, as described in Sec. 2, using the three halo finders. We show also a pair of curves using the HMF from W06, evaluated for both cosmologies, as a point of comparison.

It is apparent from this plot that, for a fixed cosmology, PHEW exhibits different behaviour to FOF and HOP: its HMF is $> 10\%$ greater for masses between 10^{12} and $10^{14} M_\odot/h$ and markedly lower at very large masses. PHEW also shows

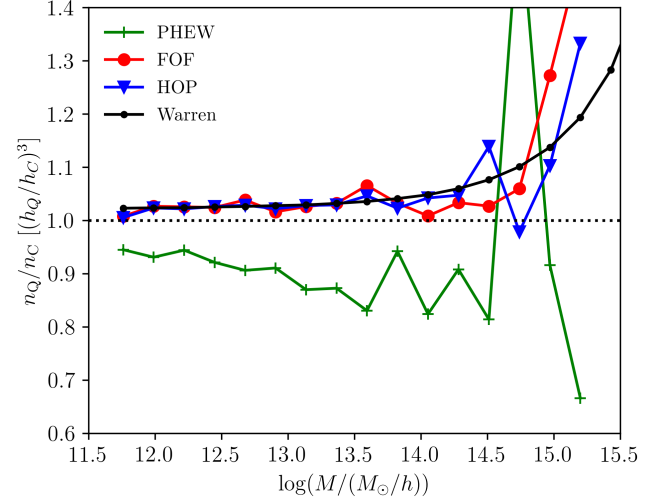


Figure 2. Ratio of HMFs of the Quijote and CSiBORG cosmologies shown in Fig. 1 for the different halo finders. The black curve is the ratio found using the W06 HMF. In the axis label, the subscripts “C” and “Q” refer to the CSiBORG and Quijote cosmologies respectively.

fundamentally different behaviour to the other halo finders between the two cosmologies, as shown in Fig. 2. The PHEW HMF is larger in the CSiBORG cosmology than in the Quijote one whereas the curves from W06, HOP, and FOF have the opposite behaviour. Since this work focuses on the variance between different realisations, we do not investigate further this discrepancy between PHEW and the other halo finders. In light of this result we do, however, regard the FOF and HOP results as more reliable. In Fig. 1, we note also a decline in the HMF from all three halo finders at $M \lesssim 10^{12.5} M_\odot/h$ compared to the fiducial curves. This is due to a variety of numerical effects affecting halos containing fewer than ~ 300 particles, including the starting redshift of the ICs, the order of the Lagrangian Perturbation Theory used to generate these (Hahn et al. 2021; Michaux et al. 2021) and the gravity solver used to run the simulation (e.g. Schneider et al. 2016).

We now show the HMF of the CSiBORG and Quijote suites, averaged over all simulation boxes, alongside the result from W06 evaluated using the CSiBORG cosmology (Fig. 3). The Quijote curve is calculated in a $155 \text{ Mpc}/h$ spherical volume to allow comparison to the CSiBORG “zoom-in” region, although we also show the Quijote HMF of the full $(1 \text{ Gpc}/h)^3$ volume with a dashed grey line and standard deviation band between realisations. The Quijote lines include all halos with ≥ 20 particles to enable comparison to CSiBORG down to lower mass, although we caution that the properties of halos with fewer than 100 particles ($M < 8.72 \times 10^{12} M_\odot/h$ for Quijote) may be unreliable (see, for example, Diemer & Kravtsov 2015).

The Quijote HMF begins to exceed the W06 curve corresponding to the same cosmology considerably for $M \gtrsim 10^{14.5} M_\odot/h$. This behaviour is well documented (Eisenstein & Hut 1998) and is a result of the greatly varying geometries of FOF halos: the algorithm tends to link particles along filaments of the cosmic web, increasing the halo mass in a way which is not observed for algorithms which produce halos with more con-

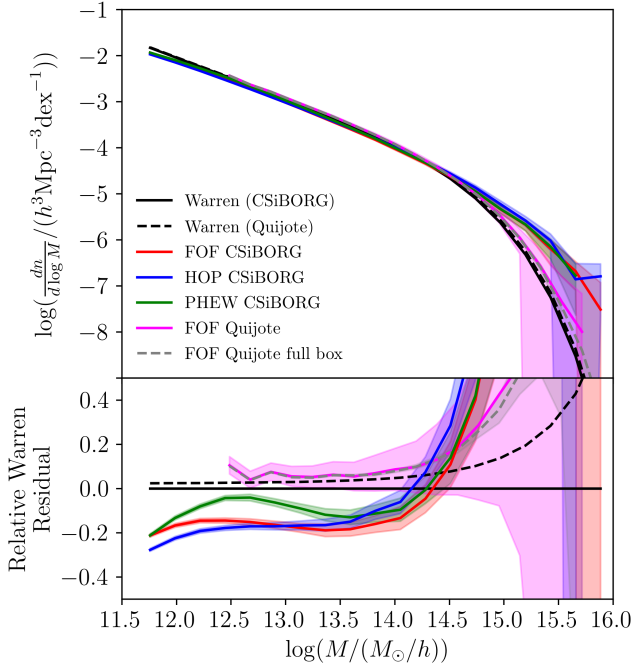


Figure 3. *Upper panel:* Mean HMFs of the CSiBORG suite using all three halofinders and of the Quijote suite using the FOF halofinder both in the full $(1 \text{ Gpc}/h)^3$ volume (grey dashed) and the spherical $155 \text{ Mpc}/h$ volume (pink). The shaded regions around each curve show $\pm 1\sigma$ standard deviation estimated across all the simulations in a suite. The black curve shows the HMF from W06, calculated in the CSiBORG cosmology (solid line) and Quijote cosmology (dashed line). *Lower panel:* Relative residual from W06 using the CSiBORG cosmology.

sistent geometrical structure. We find that the CSiBORG HMFs overshoot W06 even further in this mass range, however, and also fall considerably below it at lower masses. At low masses, all halofinders produce a HMF $\sim 20\%$ below W06 and at the largest masses probed ($\sim 10^{15.7} M_\odot/h$) the CSiBORG HMF is ~ 20 times (1.5σ) larger than that of Quijote. This discrepancy cannot be accounted for by a 5% reduction in the value of Ω_m used to calculate the W06 HMF (which corresponds to the relative underdensity of the CSiBORG volume relative to the cosmic average). The BORG reconstruction of the $2M++$ volume therefore implies more massive structures, and fewer low-mass scale structures, than average. While this is not highly statistically significant at high mass due to cosmic variance (nor in the density field itself; McAlpine et al. 2022), it is at lower masses. These results may indicate an issue with the $2M++$ BORG ICs (J. Jasche & G. Lavaux, priv. comm.) which does not, however, appreciably affect the variance between realisations, the focus of our work here.

We also find differences between the results from different halofinders. The three halofinders produce consistent results at large masses whereas at the low mass end there is a 7σ difference between the HMFs of FOF and HOP. Despite being close to the FOF result at the smallest halo masses, in the intermediate mass range the PHEW HMF has a characteristic local maximum in the Warren residual reminiscent of the trend found in the unconstrained simulations in Fig. 1. We also note that at large masses, the HOP HMF rises consistently

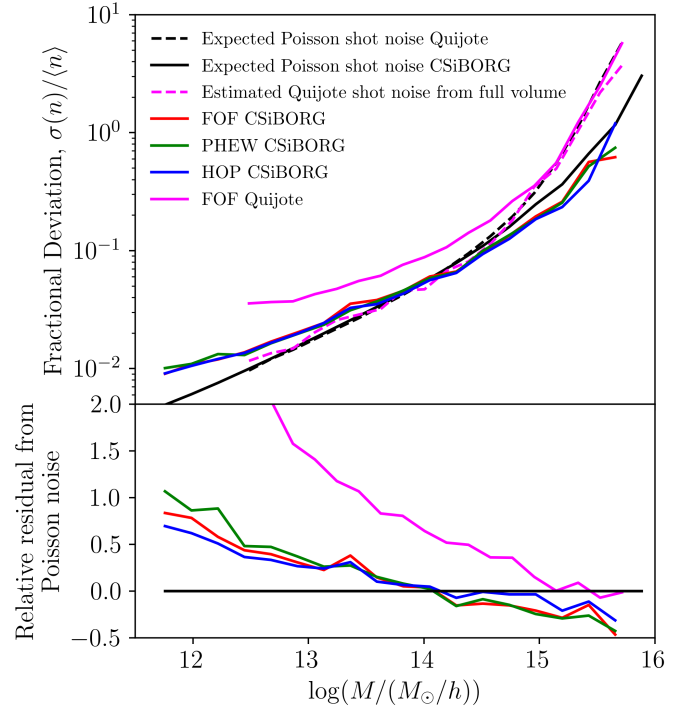


Figure 4. *Upper panel:* Fractional standard deviation of the HMF of the CSiBORG and Quijote suites across all simulations in the central $155 \text{ Mpc}/h$ spherical volume. The dashed pink line is the shot noise contribution to the standard deviation in the $155 \text{ Mpc}/h$ sphere estimated by normalising the standard deviation measured over the entire Quijote volume (see Sec. 3.2). The expected Poisson shot noise calculated in the $155 \text{ Mpc}/h$ sphere for the Quijote HMF is shown with a dashed black line, and for the CSiBORG FOF HMF with a solid black line. *Lower panel:* Relative residuals of the fractional standard deviation with respect to the Poissonian shot noise expectation for the corresponding HMF.

above those of FOF and PHEW (which agree very closely above $10^{14} M_\odot/h$). This behaviour is not seen to such an extent in the unconstrained simulations shown in Fig. 1 and the halofinder comparison project of Knebe et al. (2011) found the opposite behaviour, with the FOF HMF approximately 30% greater than that of HOP at the bright end. This could indicate that the relative behaviour of the halofinders is a function of the ICs.

4.2 Variance of the HMF

We show in Fig. 4 the fractional standard deviation of the HMF across all simulation boxes of the Quijote and CSiBORG suites (for all halofinders) using the same $155 \text{ Mpc}/h$ spherical volume for all simulations. We again include all Quijote FOF halos down to 20 particles to increase the range of comparison. The fractional deviation of the Quijote HMF is approximately a factor of 2 larger than that of CSiBORG across the entire mass range. At the lowest masses, below $1.30 \times 10^{13} M_\odot$, the fractional deviation of the Quijote suite is up to $2.6\times$ larger than that of CSiBORG due to the poorer identification of halos with < 100 particles.

We compare to the shot noise variance of the Quijote HMF in the $155 \text{ Mpc}/h$ volume, calculated in two separate ways.

First we calculate the expected fractional standard deviation due to Poissonian shot noise, $\langle N_{\text{halo}} \rangle^{-1/2}$ (using the **Quijote** HMF to estimate $\langle N_{\text{halo}} \rangle$) with a dashed black line. Second, as discussed in Sec. 3.2, we estimate the shot noise contribution in the 155 Mpc/h volume by normalising the standard deviation of the HMF calculated using the entire **Quijote** volume across all simulations. This is shown with a dashed pink line. These two methods align closely across all halo masses, which gives confidence that the excess deviation in the **Quijote** HMF calculated in the 155 Mpc/h volume above the shot noise contribution is, indeed, due to finite-volume effects, and also that the shot noise error is Poissonian. This also confirms that the dominant source of error at large scales is shot noise since the Poissonian shot noise uncertainty (dashed black) and the measured standard deviation in the 155 Mpc/h volume (solid pink) are negligibly different above $10^{15} M_{\odot}/h$. We note in particular that since the **Quijote** simulations are unconstrained, their fractional deviation is never lower than shot noise uncertainty.

We also show the expected Poissonian shot noise uncertainty in the **CSiBORG** HMF, $\langle N_{\text{halo}} \rangle^{-1/2}$, by using the **FOF** HMF to estimate $\langle N_{\text{halo}} \rangle$, with a solid black line. The lower panel of the figure shows the residual of the measured deviation of the HMF (from a given halo-finder) with respect to the corresponding Poissonian shot noise uncertainty. For example, the blue line shows the residual of the measured variance of the **HOP** HMF with respect to $\langle N_{\text{halo}} \rangle^{-1/2}$ where $\langle N_{\text{halo}} \rangle$ is estimated using the **HOP** HMF, and similarly for the other curves. We find a stark difference between the ‘excess’ variance above Poisson shot noise at low masses, which is due to residual finite-volume variance in the relatively small volume used. Since all HMFs are measured in the same region, the finite-volume contribution to the uncertainty is similar between the **Quijote** and **CSiBORG** simulations (they are not exactly equal as the largest wavelengths of the density field modes differ between the **Quijote** and **CSiBORG** suite) and therefore the ~ 2 -fold reduction in variance across all masses is a result of the IC constraints in **CSiBORG** alone.

The true constraining power of the **CSiBORG** ICs is best evidenced at large masses, $M \gtrsim 10^{14} M_{\odot}/h$, where the HMFs measured by all three halo-finders have variance below that of the Poissonian expectation. At these halo masses, uncertainty is dominated by shot noise and the **CSiBORG** variance beats Poisson by as much as 50%.

4.3 Covariance of halo masses

Fig. 5 shows the full correlation coefficient matrix of halo counts in **CSiBORG** and **Quijote**, again going down to 20 particles in the latter. This is the covariance matrix normalised by the standard deviation and hence measures correlations rather than the magnitude of the scatter:

$$\frac{\text{cov}(N_i, N_j)}{\sigma_{N_i} \sigma_{N_j}} = \frac{\langle (N_i - \langle N_i \rangle)(N_j - \langle N_j \rangle) \rangle}{\sigma_{N_i} \sigma_{N_j}}, \quad (5)$$

where N_i is the number of halos in the i^{th} mass bin and the averages are taken over all realisations. We use the **FOF** halo-finder for both suites for consistency, although the results for **HOP** and **PHEW** are extremely similar for **CSiBORG**.

The dominant effect for the unconstrained **Quijote** suite is that the mass in the central 155 Mpc/h region in which we identify halos varies by ~ 8 per cent between realisations.

Realisations with an upward fluctuation of this mass contain more halos at all masses, introducing a global positive covariance. This is particularly significant for less massive halos which respond more uniformly to the fluctuations in the total mass of the volume as they are much more numerous. There are $\mathcal{O}(1)$ halos in the most massive bins in each realisation and halos with masses in excess of $10^{15} M_{\odot}/h$ make up less than 10 per cent of the total mass in halos. Hence, even in a realisation with an 8 per cent increase in the total mass of the 155 Mpc/h volume, the number of halos in each of these most massive bins is unlikely to vary systematically. This reduces the covariance between the most massive halos in the **Quijote** suite, which, due to statistical fluctuations, is sometimes found to be negative.

Including the IC constraints in the **CSiBORG** suite, the dominant effect is instead that a halo lost from one mass bin in one realisation is likely to be found in a nearby one in any other realisation. The total mass in halos varies by only ~ 2 per cent between realisations in **CSiBORG**, making this effect subdominant to that of halos ‘matching’ between realisations with only a small mass difference. This predominantly affects nearby mass bins and massive halos for which the ICs are most constraining; for widely separated bins, or at lower mass, the correlation may again become positive due to the effect of correlated mass fluctuations (although at a weaker level than in **Quijote** due to the lower amplitude of these fluctuations). These results complement those on the variance of the HMF, and illustrate the extent to which halos may be taken to match among the constrained realisations, discussed in more detail in Sec. 4.5.

4.4 Cross-correlation function

Fig. 6 shows the two-point halo–halo cross-correlation function between halos of mass in excess of $10^{13} M_{\odot}/h$ in the central 155 Mpc/h spherical volume, averaged over all pairs of realisations in the **CSiBORG** and **Quijote** suites. The solid line shows the average over all pairs of realisations, and the shaded band the standard deviation among them. The different simulation boxes of the **Quijote** suite are fully uncorrelated, as expected. The **CSiBORG** suite, however, demonstrates strong positive cross-correlation, increasing on smaller scales. This result is salient as the trend continues down to scales below the IC constrained scale (~ 2.65 Mpc/h). This implies that constraints on large (> 1 Mpc/h) scales induce correlation on smaller ones, despite having considerably larger variance at these (formally) unconstrained small scales. This is because smaller-scale modes reside in larger ones, inheriting their constraints to some extent.

4.5 Matching halos between realisations

We show in Fig. 7 the fractional standard deviation of the distance between the matched and reference halos, and of the mass of matched halos, when we carry out the halo matching procedure described in Sec. 3.3 on **CSiBORG** using the **FOF** halo-finder. We find a clear decrease in the relative variance for massive halos compared to lower mass ones, particularly in the matched halo mass. Averaged over all halos of all simulations, for a search radius of $5R_{200}$, approximately 1.4% of halos are matched throughout all 101 realisations. However,

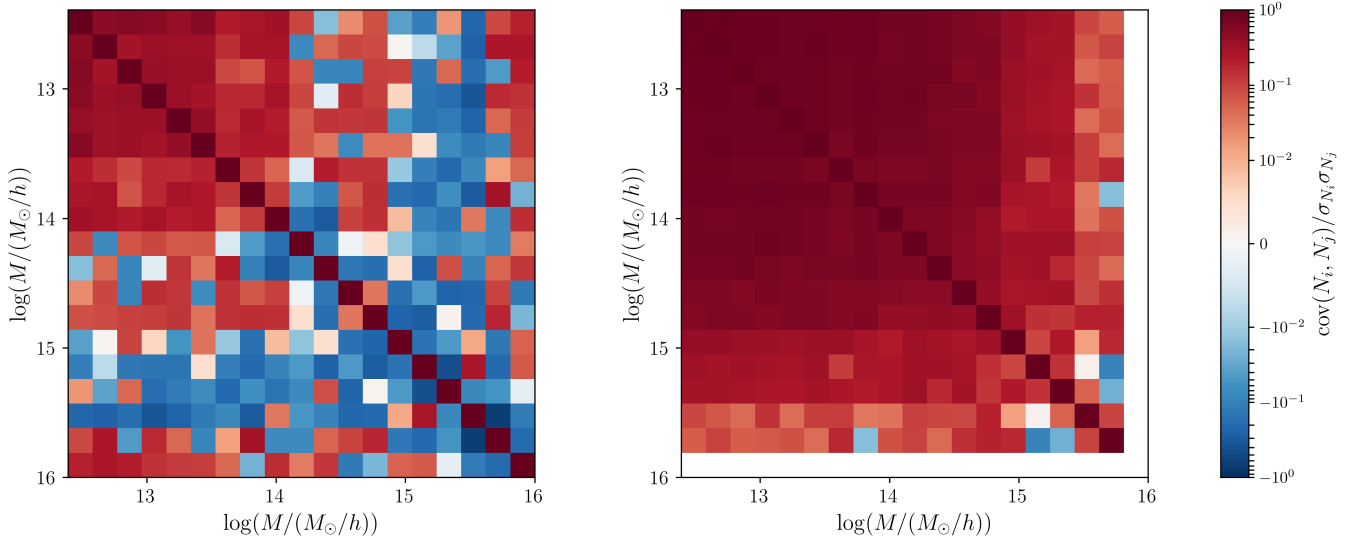


Figure 5. Heatmap of the correlation coefficient matrix of the number of halos in each mass bin for the FOF halofinder in the CSiBORG (left) and Quijote (right) suites. The final row and column on the right are left blank as there are no halos in the final mass bin in the Quijote suite.

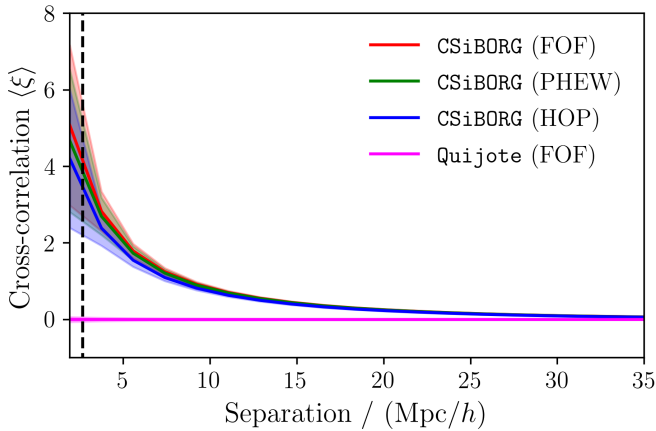


Figure 6. Mean cross-correlation function between all pairs of simulations in the CSiBORG and Quijote suites. Only halos with $M > 10^{13} M_{\odot}/h$ are included. The vertical dashed line shows the IC constrained scale (2.65 Mpc/h). The error bars show $\pm 1\sigma$ deviation measured in the correlation functions between all pairs of realisations.

the fraction of halos which are successfully matched varies with the mass of the reference halo, as seen in Fig. 8. For example, for the intermediate search radius of $5R_{200}$, over 50% of all halos with masses between 10^{15} and $10^{16} M_{\odot}/h$ are matched throughout, whereas only a negligible fraction of those with masses $< 10^{14} M_{\odot}/h$ are present in all realisations. This clearly demonstrates the effectiveness of the constraints for pinning down massive halos.

To make sure our matching procedure depends on the strength of the IC constraints alone, we apply it also to the Quijote suite. For $N = 5$ we find no matched halos across all simulations at any mass, as we would expect from unconstrained realisations. Hence, the result that 1.4% of all halos

found in a given simulation (and approximately half of the largest halos) are matched successfully throughout the *entire* CSiBORG suite is a statistically significant consequence of the IC constraints. Naturally, unconstrained modes beneath the scale of the constraints lead to structure formation differing between realisations. Our halo matching procedure, as well as the cross-correlation function discussed above, is a strong diagnostic of the strength of the IC inference from BORG, and provides a natural way to quantify the improvement in the strength of the constraints that will be afforded by future inferences. In the public halo catalogues we include a flag that indicates when a halo is matched to a halo in all other CSiBORG realisations when the realisation in question is taken as the reference, using our fiducial matching scheme ($N = 5$).

5 DISCUSSION AND CONCLUSION

We analyse the CSiBORG suite of 101 constrained realisations of the 2M++ volume using ICs drawn from the BORG posterior, identifying halos within 155 Mpc/h of the MW using three halofinders (PHEW, FOF and HOP). We compare the resulting HMFs and halo-halo cross-correlation functions with those of a similar subset of the unconstrained Quijote simulations to quantify the extent to which the local universe constraints reduce variance between the realisations. This indicates the reduction in the uncertainty of halo statistics afforded by the BORG algorithm.

By using the entire Quijote simulation volume to eliminate finite-volume effects, we isolate the shot noise contribution to the variance in the HMF, and show that the IC constraints of CSiBORG lead to a fractional standard deviation which is ~ 2 times lower than that of the unconstrained Quijote suite over all masses, and up to 50% lower than Poisson shot noise uncertainty at large masses $\gtrsim 10^{15} M_{\odot}/h$. The similarity of halos between CSiBORG realisations is reflected in the covariance matrix of halo masses, which is near-uniformly positive

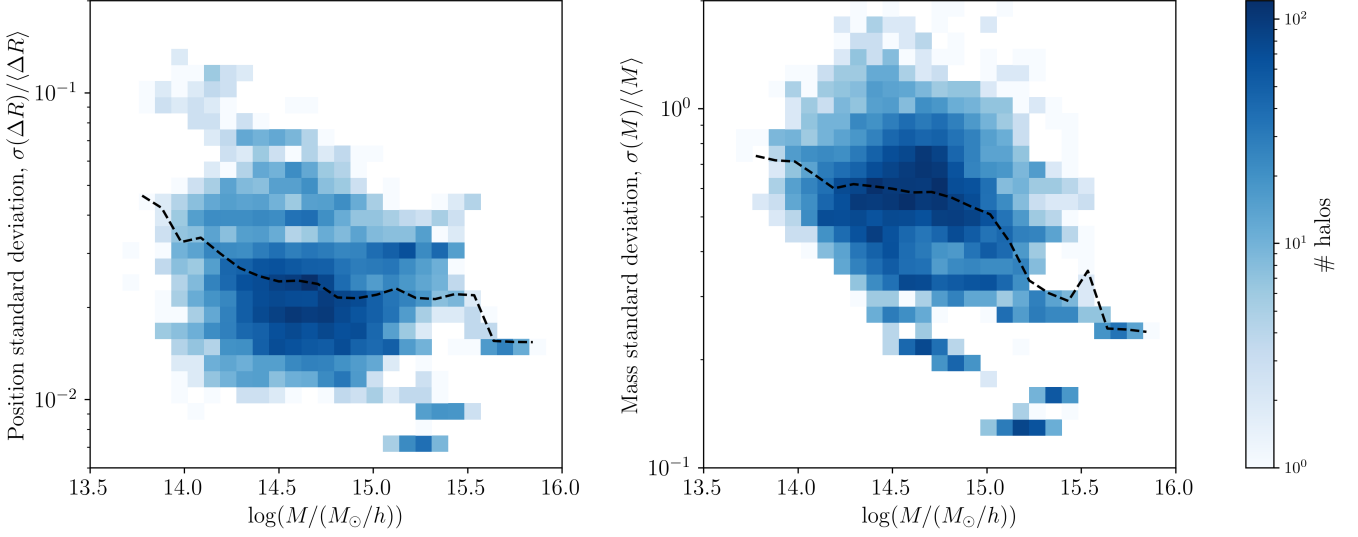


Figure 7. Matched halos distribution for a search radius of $5R_{200}$. *Left panel:* Fractional standard deviation in the position of matched halos with respect to the reference halo position, $\Delta R = |\mathbf{r}_{\text{ref}} - \mathbf{r}|$ where \mathbf{r} is the matched halo position, as a function of the mass of the reference halo. *Right panel:* Fractional standard deviation in the mass of matched halos as a function of the mass of the reference halo. The dashed black lines trace the mean fractional standard deviation of the matched halos in each mass bin.

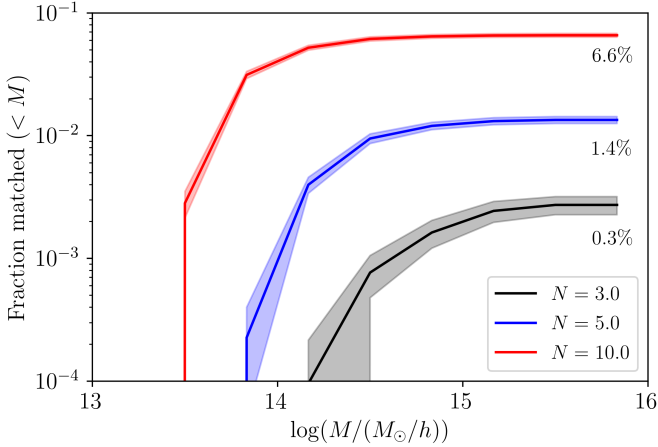


Figure 8. Cumulative fraction of CSiBORG halos matched between all realisations for three different search radii NR_{200} ($N = 3, 5, 10$). The shaded region shows $\pm 1\sigma$ standard deviation measured across all reference realisations. The percentages at the far right by each line indicate the total fraction of halos that are matched throughout the entire suite.

in *Quijote* but typically negative for nearby mass bins in CSiBORG. We also investigate the effect of the constraints on the halo–halo cross-correlation across all pairs of simulation boxes and find positive correlation on all scales, increasing towards short distance. The fact that this trend continues below the IC scale of $2.65 \text{ Mpc}/h$ implies that large-scale constraints increase consistency in structures on much smaller, formally unconstrained scales. That is, variations in the smaller scale modes of the density field are limited by the constraints on the larger modes which they inhabit. Hence, IC constraints imposed on a given set of scales percolate down to introduce

effective constraints, albeit with larger variance, on smaller scales.

We develop a halo matching procedure to identify structures which are particularly well localised throughout the entire CSiBORG simulation suite. The variance in the masses and positions of the matched halos with respect to a given reference halo decreases strongly with increasing mass, indicating that the CSiBORG constraints have greater effect for more massive halos. We demonstrate that the qualitative results of this matching are independent of the details of its implementation, and that our method finds no matches within the unconstrained *Quijote* suite. We publicly release the CSiBORG halo catalogues from the PHEW, FOF and HOP algorithms for analyses requiring knowledge of the local halo field, including its uncertainty given state-of-the-art constraints.

In order to quantify the impact of constrained ICs on the halo population we rely on halofinder algorithms, and our results indicate significant differences between them. In future, the application of 6D algorithms such as *Rockstar* (Behroozi et al. 2012) or *Velociraptor* (Elahi et al. 2019) to suites like CSiBORG and *Quijote* would improve halofinding at the massive end because they are less susceptible to erroneously combining halos that are merging near $z = 0$.

We anticipate that constrained DM-only and hydrodynamical simulations will be of great value to future studies of the local universe. Future galaxy surveys with high completeness out to scales comparable with the homogeneity scale across large parts of the sky will reduce the finite-volume uncertainties that dominate in small simulation volumes. In parallel, the continued development of algorithms like BORG will integrate the ever-improving observational efforts with the numerical techniques required to pin down the distribution of DM on cosmological scales. Incorporating additional information from, for example, peculiar velocities (Boruah et al. 2021) or cosmic shear (Porqueres et al. 2021) would strengthen the constraints and further reduce variance in the

statistics we consider here. An investigation of the effects of ICs imposed on multiple scales would also be illuminating (cf. [Sawala et al. 2022](#) for the Local Group).

6 DATA AVAILABILITY

Our CSiBORG halo catalogues for the PHEW, FOF and HOP halofinders are available [here](#). Other data will be made available upon reasonable request.

ACKNOWLEDGEMENTS

We thank Jens Jasche, Guilhem Lavaux and Stephen Stopyra for helpful discussions. We also acknowledge the authors of the Halomod ([Murray et al. 2021](#)) and Corrfunc ([Sinha & Garrison 2020](#)) Python libraries, whose codes were used to generate the fiducial HMF curves and the cross-correlation functions respectively. HD was supported by a Junior Research Fellowship at St John’s College, Oxford, a McWilliams Fellowship at Carnegie Mellon University and a Royal Society University Research Fellowship (grant no. 211046). This work was done within the [Aquila Consortium](#).

This work used the DiRAC Complexity and DiRAC@Durham facilities, operated by the University of Leicester IT Services and Institute for Computational Cosmology, which form part of the STFC DiRAC HPC Facility ([www.dirac.ac.uk](#)). This equipment is funded by BIS National E-Infrastructure capital grants ST/K000373/1, ST/P002293/1, ST/R002371/1 and ST/S002502/1, STFC DiRAC Operations grant ST/K0003259/1, and Durham University and STFC operations grant ST/R000832/1. DiRAC is part of the National E-Infrastructure.

REFERENCES

Aubert D., Pichon C., Colombi S., 2004, *MNRAS*, **352**, 376
 Bartlett D. J., Desmond H., Ferreira P. G., 2021, *Phys. Rev. D*, **103**, 023523
 Behroozi P. S., Wechsler R. H., Wu H.-Y., 2012, *ApJ*, **762**, 109
 Bistolas V., Hoffman Y., 1998, *ApJ*, **492**, 439
 Bleuler A., Teyssier R., Carassou S., Martizzi D., 2015, *Computational Astrophysics and Cosmology*, **2**
 Boruah S. S., Lavaux G., Hudson M. J., 2021, arXiv e-prints, p. [arXiv:2111.15535](#)
 Colombi S., Bouchet F. R., Schaeffer R., 1994, *A&A*, **281**, 301
 Colombi S., Szapudi I., Jenkins A., Colberg J., 2000, *MNRAS*, **313**, 711–724
 Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, *ApJ*, **292**, 371
 Desmond H., Hutt M. L., Devriendt J., Slyz A., 2022, *Monthly Notices of the Royal Astronomical Society: Letters*, **511**, L45
 Despali G., Giocoli C., Angulo R. E., Tormen G., Sheth R. K., Baso G., Moscardini L., 2015, *MNRAS*, **456**, 2486
 Diemer B., Kravtsov A. V., 2015, *ApJ*, **799**, 108
 Eisenstein D. J., Hut P., 1998, *ApJ*, **498**, 137
 Elahi P. J., Cañas R., Poulton R. J. J., Tobar R. J., Willis J. S., Lagos C. d. P., Power C., Robotham A. S. G., 2019, *Publ. Astron. Soc. Australia*, **36**, e021
 Hahn O., Rampf C., Uhlemann C., 2021, *MNRAS*, **503**, 426
 Hoffman Y., Ribak E., 1991, *ApJ*, **380**, L5
 Jasche J., Lavaux G., 2019, *A&A*, **625**, A64
 Jasche J., Wandelt B. D., 2012, *MNRAS*, **425**, 1042

Jasche J., Wandelt B. D., 2013, *MNRAS*, **432**, 894
 Jasche J., Kitaura F. S., Wandelt B. D., Enßlin T. A., 2010, *MNRAS*, **406**, 60
 Klypin A., Hoffman Y., Kravtsov A. V., Gottlöber S., 2003, *ApJ*, **596**, 19
 Knebe A., et al., 2011, *MNRAS*, **415**, 2293–2318
 Kolatt T., Dekel A., Ganon G., Willick J. A., 1996, *ApJ*, **458**, 419
 Kravtsov A. V., Klypin A., Hoffman Y., 2002, *ApJ*, **571**, 563
 Lacey C., Cole S., 1993, *MNRAS*, **262**, 627
 Landy S. D., Szalay A. S., 1993, *ApJ*, **412**, 64
 Lavaux G., Hudson M. J., 2011, *MNRAS*, **416**, 2840
 Lavaux G., Jasche J., 2016, *MNRAS*, **455**, 3169
 Lavaux G., Jasche J., Leclercq F., 2019, arXiv e-prints, p. [arXiv:1909.06396](#)
 Lukić Z., Reed D., Habib S., Heitmann K., 2009, *The Astrophysical Journal*, **692**, 217
 McAlpine S., et al., 2022, *MNRAS*,
 Michaux M., Hahn O., Rampf C., Angulo R. E., 2021, *MNRAS*, **500**, 663
 Moster B. P., Somerville R. S., Newman J. A., Rix H.-W., 2011, *The Astrophysical Journal*, **731**, 113
 Murray S., Diemer B., Chen Z., Neuhold A., Schnapp M., Peruzzi T., Blevins D., Engelman T., 2021, *Astronomy and Computing*, **36**, 100487
 Porqueres N., Heavens A., Mortlock D., Lavaux G., 2021, *MNRAS*, **502**, 3035
 Press W. H., Schechter P., 1974, *ApJ*, **187**, 425
 Sawala T., McAlpine S., Jasche J., Lavaux G., Jenkins A., Johansson P. H., Frenk C. S., 2022, *MNRAS*, **509**, 1432
 Schneider A., et al., 2016, *Journal of Cosmology and Astroparticle Physics*, **2016**, 047–047
 Sinha M., Garrison L. H., 2020, *MNRAS*, **491**, 3022
 Springel V., et al., 2008, *MNRAS*, **391**, 1685
 Szapudi I., Colombi S., 1996, *ApJ*, **470**, 131
 Teyssier R., 2002, *A&A*, **385**, 337
 Tinker J., Kravtsov A. V., Klypin A., Abazajian K., Warren M., Yepes G., Gottlöber S., Holz D. E., 2008, *The Astrophysical Journal*, **688**, 709–728
 Tweed D., Devriendt J., Blaizot J., Colombi S., Slyz A., 2009, *A&A*, **506**, 647
 Tweed D., Yang X., Wang H., Cui W., Zhang Y., Li S., Jing Y. P., Mo H. J., 2017, *ApJ*, **841**, 55
 Villaescusa-Navarro F., et al., 2020, *The Astrophysical Journal Supplement Series*, **250**, 2
 Wang H., Mo H. J., Yang X., Jing Y. P., Lin W. P., 2014, *ApJ*, **794**, 94
 Wang H., et al., 2016, *ApJ*, **831**, 164
 Warren M. S., Abazajian K., Holz D. E., Teodoro L., 2006, *ApJ*, **646**, 881–885
 Yepes G., Martínez-Vaquero L. A., Gottlöber S., Hoffman Y., 2009, in Balazs C., Wang F., eds, *American Institute of Physics Conference Series Vol. 1178*, 5th International Workshop on the Dark Side of the Universe. pp 64–75, [doi:10.1063/1.3264558](#)
 van de Weygaert R., Hoffman Y., 2000, in Courteau S., Willick J., eds, *Astronomical Society of the Pacific Conference Series Vol. 201*, Cosmic Flows Workshop. p. 169 ([arXiv:astro-ph/9909103](#))

This paper has been typeset from a \LaTeX file prepared by the author.