A flexible analytic model of cosmic variance in the first billion years

A. C. Trapp ⓘ★ and Steven R. Furlanetto
Department of Physics and Astronomy, University of California Los Angeles, Los Angeles, CA 90095-1562, USA

ABSTRACT

Cosmic variance is the intrinsic scatter in the number density of galaxies due to fluctuations in the large-scale dark matter density field. In this work, we present a simple analytic model of cosmic variance in the high-redshift Universe ($z \sim 5$–15). We assume that galaxies grow according to the evolution of the halo mass function, which we allow to vary with large-scale environment. Our model produces a reasonable match to the observed ultraviolet (UV) luminosity functions in this era by regulating star formation through stellar feedback and assuming that the UV luminosity function is dominated by recent star formation. We find that cosmic variance in the UV luminosity function is dominated by the variance in the underlying dark matter halo population, and not by differences in halo accretion or the specifics of our stellar feedback model. We also find that cosmic variance dominates over Poisson noise for future high-z surveys except for the brightest sources or at very high redshifts ($z \gtrsim 12$). We provide a linear approximation of cosmic variance for a variety of redshifts, magnitudes, and survey areas through the public PYTHON package GALCV. Finally, we introduce a new method for incorporating priors on cosmic variance into estimates of the galaxy luminosity function and demonstrate that it significantly improves constraints on that important observable.

Key words: methods: data analysis – galaxies: high-redshift – cosmology: theory.

1 INTRODUCTION

Extragalactic astronomy is closing in on arguably the most important era of galaxy evolution: the formation of the first galaxies. These galaxies will allow us to probe the processes that drove the first emergence of complexity in our Universe.

The most fundamental observations for studying this era have been (and will continue to be) deep galaxy surveys. These surveys measure many important features of the galaxy population, most fundamentally the ultraviolet luminosity function (UVLF) of galaxies. The UVLF is a measure of the number of galaxies at each luminosity, and its shape and evolution through cosmic time has important implications for the physics behind galaxy formation and growth, and much more (see e.g. Bouwens et al. 2015; Finkelstein et al. 2015; Livermore, Finkelstein & Lotz 2017; Atek et al. 2018; Oesch et al. 2018; Behroozi et al. 2019). Unfortunately, these deep galaxy surveys will have very small volumes, which will be a key limitation in measuring the UVLF due to the effects of 'cosmic variance': not all regions of the Universe contain the average number of galaxies, and those galaxies did not all grow up in an average environment. We must understand how cosmic variance affects the UVLF in order to inform and correctly interpret future deep galaxy surveys.

Cosmic variance in the UVLF (and other measures) has been modelled in a variety of ways in the past. For example, analytic models typically start with the linear halo bias function and then connect haloes to galaxies with a halo mass–luminosity relation, or by matching abundances (see e.g. Newman & Davis 2002; Somerville et al. 2004; Stark, Loeb & Ellis 2007; Moster et al. 2011). These models conclude that cosmic variance is a significant source of

uncertainty when studying galaxies at high redshifts. However, such models do not allow cosmic variance to affect the halo mass to luminosity connection itself; they assume galaxies are the same in all environments. Also, the linear halo bias function does not accurately predict cosmic variance in extreme environments.

Cosmic variance can also be estimated using mock observations of galaxy simulations. The early implementations of this method (see e.g. Kitzbichler & White 2007; Trenti & Stiavelli 2008) were very powerful, but also assumed galaxies are the same in all environments. Recently, substantial improvements in computing power have allowed for much higher volume *N*-body simulations that also treat star formation in a more complex way (e.g. Bhowmick et al. 2020; Ucci et al. 2020). These studies take into account the difference in environment on individual galaxy growth, and are a major step forward in predicting cosmic variance in the first galaxies. However, these studies (1) are still limited by their volume, as a complete picture of cosmic variance requires extremely large volumes to calculate cosmic variance on all relevant scales and magnitudes; (2) cannot explore how cosmic variance depends on their specific implementation of mass accretion, star formation, feedback, and other parameters, without rerunning simulations many times, which would be prohibitively expensive; and (3) can be limited in their redshift or magnitude ranges.

Simulations lack large volume and flexibility, while existing analytic models rely on linear theory and lack a fully self-consistent connection between cosmic variance and galaxy growth and star formation.

Quantitative interpretations of high-*z* data require corrections for cosmic variance, especially because most planned surveys subtend relatively small volumes. Such corrections are particularly important when multiple independent surveys are combined, as each such survey contains its own (unknown) intrinsic density. The standard

★ E-mail: atrapp@astro.ucla.edu

method to account for cosmic variance when fitting a UVLF, originally developed by Sandage, Tammann & Yahil (1979) and used by e.g. Efstathiou, Ellis & Peterson (1988) and Bouwens et al. (2015), fits a universal *shape* of the UVLF to all fields, ignoring the normalization parameter of the fit in each individual field. After the shape has been optimized, the overall normalization is determined by demanding that it reproduces the correct total number of galaxies across all surveys. This method cannot account for a change in shape of the UVLF between fields, and it does not include a prior for the amount of variance allowed in the normalization parameter of the UVLF.

In this paper, we use a simple, flexible analytic model of high-z galaxies to study the effects of cosmic variance on galaxy surveys. We begin in Section 2 with a close examination of large-scale variations in the dark matter halo population in the context of excursion set models of halo formation. This provides the fundamental basis for the cosmic variance of the galaxy population, and with it we capture some non-linear aspects of these fluctuations. Next, in Section 3, we describe a 'minimalist' model of galaxy evolution (Furlanetto et al. 2017) that fits observed luminosity functions reasonably well but is sufficiently flexible to examine how a large range of assumptions about the physics of these sources affects cosmic variance. We use this model to determine which of the many uncertain parameters of galaxy formation have the most impact on cosmic variance, and we account for changes in galaxy growth and star formation in different environments. We combine our treatments of dark matter haloes and galaxy physics in Section 4, where we also provide a linear approximation to the cosmic variance of galaxies as a function of redshift and absolute magnitude. Unlike other such functions derived from simulations, our results apply across any mass or redshift, and we quantify how uncertainties in galaxy evolution parameters affect the results.

In Section 5, we then describe the importance of our cosmic variance results for future surveys of high-z galaxies with the *James Webb Space Telescope* (*JWST*) and the *Nancy Grace Roman Space Telescope* (hereafter, the *Roman Space Telescope*; Spergel et al. 2015; Akeson et al. 2019; Dore et al. 2019). We show how cosmic variance limits inferences about the average UVLF of the Universe. Additionally, we introduce a method that fully incorporates cosmic variance into UVLF estimates, essentially treating our estimates for cosmic variance as a prior on the measurements. Most commonly, UVLF estimates allow for an arbitrary amount of cosmic variance between fields, by ignoring the normalization of the UVLF in each field (e.g. Bouwens et al. 2015; Finkelstein et al. 2015, though see Livermore et al. 2017 for a contrasting case). We show that our method provides tighter constraints in mock surveys. Finally, in Section 6, we summarize our results.

We take the following cosmological parameters: $\Omega_{\rm m} = 0.308$, $\Omega_\Lambda = 0.692$, $\Omega_{\rm b} = 0.0484$, $h = 0.678$, $\sigma_8 = 0.815$, and $n_{\rm s} = 0.968$, consistent with recent Planck Collaboration XIII results (Planck Collaboration XIII 2016). We give all distances in comoving units.

2 DARK MATTER HALOES

We follow the methods described in Furlanetto et al. (2017) to model dark matter haloes. In this section, we give a brief summary of those methods and also describe some additions.

2.1 Conditional halo mass function

We define the dark matter halo mass function as $n_{\rm h}(m, z)\,{\rm d}m$: the comoving number density of dark matter haloes between masses (m,

$m + {\rm d}m$) at redshift z. By convention,

$$n_{\rm h}(m, z) = f(\sigma) \frac{\bar{\rho}}{m} {\rm d} \ln \frac{(1/\sigma)}{{\rm d}m}, \tag{1}$$

where $\bar{\rho}$ is the comoving average matter density, $\sigma(m, z)$ is the linear rms fluctuation of the matter density field at redshift z smoothed over a spherical region of mass m [see Section 2.3 for the calculation of $\sigma(m, z)$], and $f(\sigma)$ is a dimensionless function that modifies the shape of the mass function. Following Furlanetto et al. (2017), we use $f_{\rm Trac}(\sigma)$ from a fit to the average mass function of a high-z cosmological simulation (Trac, Cen & Mansfield 2015):

$$f_{\rm Trac}(\sigma) = 0.150 \left[1 + \left(\frac{\sigma}{2.54} \right)^a \right] {\rm e}^{b/\sigma^2}, \tag{2}$$

with $a = -1.36$ and $b = -1.14$.

The key aspect of this model is allowing the halo mass function to depend on its environment. This environmental dependence is introduced via the conditional mass function (CMF). The CMF $n_{\rm cond}(m, z, \delta_{\rm b}, R)$ describes the number density of haloes in a spherical region of mass M, with a corresponding Lagrangian radius[1] $R^3 = 3M/(4\pi\bar{\rho})$, and relative density $\delta_{\rm b} = (\rho - \bar{\rho})/\bar{\rho}$, where ρ is the linearly extrapolated matter density in that region at redshift z. The CMF is what adds cosmic variance into the model.

We determine $n_{\rm cond}(m, z, \delta_{\rm b}, R)$ using a coordinate transfer method described in Tramonte et al. (2017) that we will call 'ν-scaling' applied to equation (2),[2]

$$\sigma^2(m, z) \rightarrow [\sigma^2(m, z) - \sigma^2(M, z)] \left[\frac{\delta_{\rm crit}}{\delta_{\rm crit} - \delta_{\rm b}} \right]^2, \tag{3}$$

where M is the mass corresponding to the region R, and $\delta_{\rm crit} \approx 1.69$ is the linear halo collapse threshold (see Loeb & Furlanetto 2013, equation 3.13). The resulting CMF is Lagrangian in that it assumes all regions of fixed mass have the same volume. To convert into a real-space (Eulerian) CMF, we calculate the real-space radius $R_{\rm e}$ of a region of mass M and density $\delta_{\rm b}$ assuming spherical collapse: $R_{\rm e} = R/(1 + \delta_{\rm r})^{1/3}$, where $\delta_{\rm r}$ is the real-space (non-linear) relative density[3] (Mo & White 1996, see Appendix A for more details). Applying this adjustment to the radius of each region results in an Eulerian CMF, $n_{\rm cond}(m, z, \delta_{\rm b}, R_{\rm e}) = n_{\rm cond}(m, z, \delta_{\rm b}, R)(1 + \delta_{\rm r})$.

Tramonte et al. (2017) justify ν-scaling by noting $\delta_{\rm crit}$ enters into $f(\sigma)$ only through the variable $\nu = \delta_{\rm crit}/\sigma$ in previous parametrizations. They then apply the 'standard' coordinate transfer,

$$\delta_{\rm crit} \rightarrow \delta_{\rm crit} - \delta_{\rm b},$$
$$\sigma^2(m, z) \rightarrow \sigma^2(m, z) - \sigma^2(M, z), \tag{4}$$

to the variable ν, giving equation (3). Tramonte et al. (2017) validate this method using an N-body simulation by Tinker et al. (2008), finding that this scaling technique accurately describes the CMF except for the most underdense regions ($\delta_{\rm b} \lesssim -1.5$), where it overestimates halo abundance.

We note that while we consider halo masses down to $m \sim 10^8\,{\rm M}_\odot$, Tramonte et al. (2017) test their prescription only down to a mass of $m \sim 3 \times 10^{10}\,{\rm M}_\odot$, leaving it untested for the lowest masses we

---

[1]Note that the radius R is really a mass scale, as it does not correspond to the real radius of a region except for regions that happen to be at cosmological average density.

[2]We construct the CMF this way because simulations of dark matter haloes (e.g. McBride, Fakhouri & Ma 2009; Goerdt et al. 2015; Trac et al. 2015) do not provide a full CMF.

[3]In practice, $\delta_{\rm r}$ and $\delta_{\rm b}$ are very similar, especially at the redshifts considered in this paper.

consider. We also note that we use a different mass function (Trac et al. 2015), for which this method has not been explicitly tested. Tramonte et al. (2017) also test 'local scaling', a more rigorous method for constructing a CMF developed in Patiri et al. (2006) and expanded in Rubiño-Martín, Betancort-Rijo & Patiri (2008), and find that method produces a slightly better CMF. We do not test our model using local scaling as it cannot be easily applied to the Trac et al. (2015) mass function.

Mo & White (1996) developed a method to linearly approximate their CMF by use of a bias function b_{PS}. We follow the same steps to calculate a linear bias factor b_{Trac} for the Trac et al. (2015) mass function. We first substitute equation (3) into equation (2), and then Taylor expand to linear order about $\delta_b = 0$, and set $\sigma(M, z) = 0$. We then make a linear volume change correction of $(1 + \delta_b)$, giving us

$$b_{Trac} = 1 + \frac{a}{\delta_{crit}} \frac{(\sigma/2.54)^a}{1 + (\sigma/2.54)^a} - \frac{2b}{\sigma^2 \delta_{crit}}. \quad (5)$$

The CMF can then be approximated as

$$n_{cond,lin}(m, z, \delta_b) = n_h(m, z)(1 + b_{Trac}\delta_b). \quad (6)$$

We will use this linear approximation of the CMF to compare to our results when using the full CMF.

For another comparison, we calculate the CMF by scaling the Trac et al. (2015) mass function by the ratio of the conditional to non-conditional Press–Schechter mass functions:

$$\frac{n_{cond}}{n_h} = \frac{n_{PS,cond}}{n_{PS}}, \quad (7)$$

where n_{PS} is defined in Press & Schechter (1974), and represented here as $f_{PS}(\sigma)$ (which is plugged into equation 1),

$$f_{PS}(\sigma) = \sqrt{\frac{2}{\pi}} \nu \, e^{-\nu^2/2}. \quad (8)$$

The Press–Schechter CMF $n_{PS,cond}(m, z, \delta_b, R)$ is obtained with the 'standard' coordinate transfer in equation (4). We then multiply by the same $(1 + \delta_r)$ factor to obtain an Eulerian CMF. We note that this scaling, which was introduced in the high-z context by Barkana & Loeb (2004), is commonly used in analytic, semi-analytic, and seminumeric calculations of galaxy populations at this time (e.g. in Mesinger, Furlanetto & Cen 2011).

At these high redshifts, we ignore the effects of assembly bias (e.g. Gao & White 2007) on the CMF, as it is a small effect compared to the other uncertainties in our model.

2.2 Dark matter density fluctuations

As stated above, $\sigma(M, z)$ is the linear rms fluctuation of the matter density field at redshift z smoothed over a spherical region of mass M (and corresponding R). Thus, the probability distribution of dark matter density for a given scale R and redshift z, $p(\delta_b|R, z)$, is by definition equal to a zero-mean Gaussian with variance $\sigma^2(M, z)$. However, galaxy surveys measure Eulerian volumes, so to make predictions for them we must convert this distribution to that system. A fixed Eulerian volume will correspond to a range of masses, because each has a different density. In Appendix A, we convert the probability distribution of densities at fixed region mass $p(\delta_b|R, z)$ to fixed real-space volume $p(\delta_b|R_e, z)$. While $p(\delta_b|R, z)$ is a Gaussian, $p(\delta_b|R_e, z)$ is closer to an inverse Gaussian. Fortunately, these two distributions are very similar to one another at the region sizes and redshifts we consider in this paper.

However, we do find that $p(\delta_b|R_e, z)$ predicts that underdense regions occupy a larger volume fraction of the Universe than
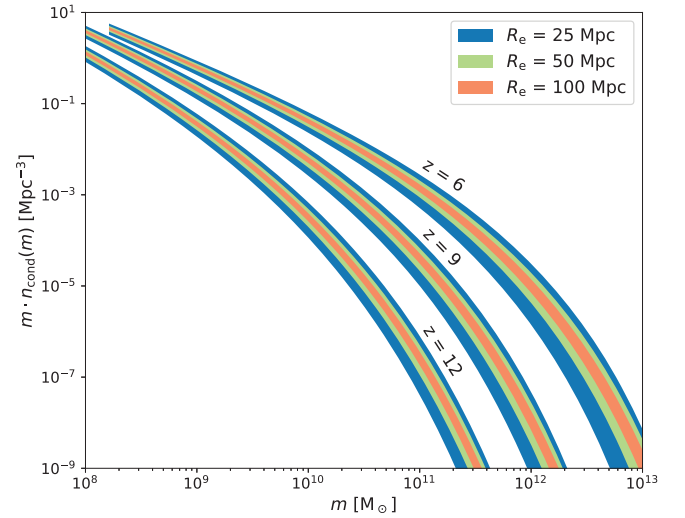

Figure 1. The CMF n_{cond} and its 2σ scatter due to cosmic variance on various scales (identified with their radius R_e) at three redshifts. The scatter increases at high mass and for smaller scales of the Universe (widest shaded area corresponds to the smallest scale).

overdense regions at all scales, by as much as ∼12 per cent when considering very small scales. This result indicates that surveys will be slightly more likely to probe underdense regions (see Appendix A for more details). Using different methods, Muñoz, Trac & Loeb (2010) also found that surveys are more likely to probe an underdense region because of those regions' more rapid cosmic expansion.

With the distribution of densities $p(\delta_b|R_e, z)$ and the CMF $n_{cond}(m, z, \delta_b, R_e)$, we can compute the scatter in halo number density on various scales R_e and redshifts. Fig. 1 shows some example results. Cosmic variance in the mass function is substantial for the haloes in which high-z galaxies form. For example, at redshift $z = 9$ on a 50 Mpc (radius) scale, massive haloes (∼10^{12} M$_\odot$) have a typical relative standard deviation of ∼65 per cent, while haloes at the atomic cooling limit (see Section 3.1) have a relative standard deviation of ∼10 per cent. At fixed halo mass, these relative standard deviations increase at higher redshifts and decrease when considering larger volumes.

2.3 The CMF in realistic survey volumes

The CMF presented in Section 2.1 assumes a spherical region of radius R_e. However, real surveys subtend elongated regions pointing away from the Earth, commonly referred to as pencil beams. Here, we describe a method for building a CMF for a pencil-beam region.

We start with the variance in the dark matter density field σ^2 in a pencil-beam region (following e.g. Newman & Davis 2002; Stark et al. 2007; Muñoz et al. 2010; Robertson 2010). For an arbitrarily shaped volume V,

$$\sigma^2(V) = \frac{F_g(z)}{(2\pi)^3} \int P(\boldsymbol{k})|\hat{W}_V(\boldsymbol{k})|^2 \, d\boldsymbol{k}, \quad (9)$$

where $F_g(z)$ is the growth function (nearly equal to $1/(1 + z)$), $\boldsymbol{k}$ is wave vector, $P(\boldsymbol{k})$ is the power spectrum of dark matter (we use the transfer function from Eisenstein & Hu 1998), and $\hat{W}_V(\boldsymbol{k})$ is the Fourier transform of a real-space top hat in the shape of the region V, normalized such that its integral in real space is equal to unity. In the case of a rectangular pencil-beam volume with side lengths a_x, a_y,

a_z, $\hat{W}_V(\boldsymbol{k}) = \hat{W}(k_x)\hat{W}(k_y)\hat{W}(k_z)$ with $\hat{W}(k_i) = \sin(a_i k_i/2)/(a_i k_i/2)$. When constructing a survey volume, we define a_z as the radial distance corresponding to some Δz centred at z. We define a_x and a_y such that the physical area $a_x * a_y$ at z gives the survey area A as seen from the Earth.

We then make the simple approximation that a pencil-beam region has the same CMF as a (larger) spherical region of radius R_{eff}, such that $\sigma_{\text{sphere}}(R_{\text{eff}}) = \sigma_{\text{PB}}(a_x, a_y, a_z)$. This prescription is analogous to how pencil-beam volumes are treated in other analytic studies of cosmic variance. In such studies, σ_{PB} is multiplied by a halo bias function to find cosmic variance (e.g. Stark et al. 2007; Muñoz et al. 2010; Robertson 2010; Moster et al. 2011). The halo bias functions used in these studies come from CMFs determined assuming spherical regions, so that pencil-beam volumes are treated as spherical volumes with equivalent σ.

2.4 Accretion rates

We now consider how dark matter haloes accrete matter. This accretion will be used in the next section to determine the rate of star formation.

Many simulations provide similar predictions of halo mass accretion rates (e.g. McBride et al. 2009; Fakhouri, Ma & Boylan-Kolchin 2010; van den Bosch et al. 2014; Goerdt et al. 2015; Trac et al. 2015). However, these rates have not been tested at the very high redshifts and very low masses relevant to our model. For this paper, we calculate the accretion rates using the method described in Furlanetto et al. (2017), which is analogous to abundance matching (Vale & Ostriker 2004): haloes maintain a constant number density as they evolve according to the mass function of the region they are in. That is, we require that at any two nearby redshifts z_1 and z_2, a halo has masses $m_1(z_1)$ and $m_2(z_2)$ such that

$$\int_{m_1}^{\infty} \mathrm{d}m\, n_{\text{cond}}(m, z_1, \delta_b, R) = \int_{m_2}^{\infty} \mathrm{d}m\, n_{\text{cond}}(m, z_2, \delta_b, R), \qquad (10)$$

where $n_{\text{cond}}(m, z, \delta_b, R)$ is the Lagrangian CMF from equation (7). We define the accretion rate of haloes $\dot{m}_h$ in a region such that they satisfy equation (10) at all masses over a small redshift interval ($\Delta z \sim 0.1$).

In practice this accretion method means that in a given region, the most massive halo at one time step is also the most massive halo at the next time step, and the same goes for the second and third most massive haloes, etc. In this treatment, accretion is continuous and smooth, increases monotonically with halo mass, and has zero scatter at a fixed mass. This treatment is obviously not entirely correct, but it is in line with our goal of simplicity and has the added benefit maintaining the CMF across cosmic time in a way that conserves mass. Furlanetto et al. (2017) show that accretion rates obtained using this method are similar to the simulation accretion rates mentioned above in the redshift and mass ranges they probe. We do neglect mergers in this model, which will provide an additional source of scatter (see the discussion in Furlanetto et al. 2017).

This method allows haloes in over- and underdense regions to accrete at different rates. However, we find that this is actually a small effect. The shaded areas in Fig. 2 show how the accretion rates depend on large-scale environment. For most masses, redshifts, and scales, the variation in accretion is less than 5 per cent. A 5 per cent difference in accretion can be significant over a Hubble time, but we will show it has a small effect on the UVLF, which is most sensitive to instantaneous star formation.
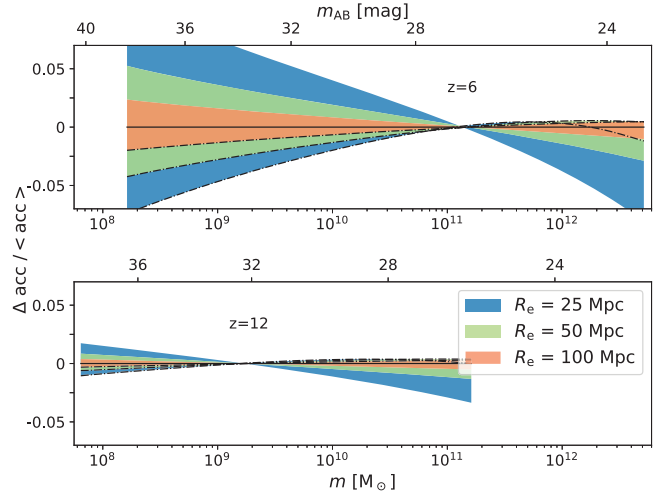

Figure 2. Accretion rate scatter (shaded areas) due to cosmic variance compared to the average accretion rate of haloes (solid black line) for various scales (identified with their radius R_e). Adding cosmic variance to our models changes accretion rates by <5 per cent for most haloes; smaller regions have larger variance in accretion. The dot–dashed lines show the accretion of a 1σ overdense region, and the opposite end of the shaded regions show the 1σ underdense region. Higher mass haloes overaccrete in overdense environments and underaccrete in underdense environments. For lower mass haloes, the opposite is true. The threshold mass where haloes in all environments accrete nearly equally evolves to lower mass with redshift. The top axes show the approximate apparent magnitude of the haloes 'm_{AB}' (we assign haloes their magnitudes in Section 3).

3 FEEDBACK-REGULATED STAR FORMATION

In this section, we transform the mass accretion rates of haloes into UV luminosities. Furlanetto et al. (2017) provide a detailed explanation of our star formation model; we briefly summarize it here. We intentionally choose this simple, 'minimalist' model so as to make our assumptions about the mass–luminosity relation transparent. Estimates of cosmic variance must necessarily account for the many uncertainties about high-z galaxies, and a simple, flexible model allows us to estimate how important the specifics of galaxy formation are for the variance.

3.1 Models of feedback

We assume that haloes only form stars when they exceed a threshold mass m_{min}. This mass corresponds to a halo virial temperature $T_{\text{vir}} = 10^4$ K, when atomic line cooling becomes efficient enough for gas clouds to collapse and fragment for star formation (Loeb & Furlanetto 2013). This mass is typically $m_{\text{min}} \sim 10^8$ M$_\odot$. At the redshifts considered in this paper, haloes at the threshold are always far below the detection limit of next generation telescopes.

Gas accreting on to a galaxy can be turned into stars. When stars form, they expel baryons from their host galaxy through radiation pressure, supernovae, or some other process like grain heating (e.g. Faucher-Giguère, Quataert & Hopkins 2013; Hayward & Hopkins 2017; Krumholz et al. 2018). Balancing this stellar feedback with accretion provides a simple estimate of the star formation rate $\dot{m}_*$ of a galaxy via

$$\dot{m}_* = \dot{m}_b - \dot{m}_w, \qquad (11)$$

where $\dot{m}_b$ is the mass accretion rate of the halo times the baryon fraction $\dot{m}_b = [\Omega_b/\Omega_m]\dot{m}_h$, and $\dot{m}_w$ is the rate of baryon loss through

feedback. The fraction of accreting baryons that are converted into stars is defined as $f_* = \dot{m}_*/\dot{m}_b$. Finally, we write the mass ejection rate as a multiple of the star formation rate $\dot{m}_w = \eta(m, z)\, \dot{m}_*$, yielding

$$f_* = \frac{1}{1 + \eta(m, z)}. \tag{12}$$

Many models suggest that massive haloes accrete gas more slowly than our simple argument suggests, because of the heating at the virial shock. Furlanetto et al. (2017) show that virial shock heating only has a modest effect on the results of this model, but we include it because it helps match the observed densities at large luminosities. Faucher-Giguère, Kereš & Ma (2011) show the fraction of gas that can cool on to a galaxy in the presence of a virial shock is[4]

$$f_{\text{shock}} = 0.47 \left(\frac{1+z}{4} \right)^{0.38} \left(\frac{m}{10^{12}\,\text{M}_\odot} \right)^{-0.25}. \tag{13}$$

Because we only include stellar feedback, which limits star formation at small masses, we also impose a maximum efficiency $f_{*,\text{max}}$ that limits star formation when $\eta(m, z) \to 0$ at large halo masses. We impose it in a way that keeps f_* smoothly differentiable. Combining f_{shock} and $f_{*,\text{max}}$ with equation (12) gives

$$f_* = \frac{f_{\text{shock}}}{f_{*,\text{max}}^{-1} + \eta(m, z)}. \tag{14}$$

Finally, we parametrize the strength of stellar feedback, $\eta(m, z)$, as

$$\eta = C \left(\frac{10^{11.5}\,\text{M}_\odot}{m} \right)^{\xi} \left(\frac{9}{1+z} \right)^{\sigma}. \tag{15}$$

For energy-driven supernova feedback: $C = 1$, $\xi = 2/3$, $\sigma = 1$, and $f_{*,\text{max}} = 0.1$. We will also consider a redshift-independent version ($C = 2$, $\xi = 2/3$, $\sigma = 0$, and $f_{*,\text{max}} = 0.1$) and a momentum-driven version ($C = 5$, $\xi = 1/3$, $\sigma = 1/2$, and $f_{*,\text{max}} = 0.2$) for comparison (for more details on η and its parametrization, see Sun & Furlanetto 2016; Furlanetto et al. 2017; Mirocha, Furlanetto & Sun 2017). These alternate parametrizations of η will allow us to test how cosmic variance depends on our galaxy formation model.

This simple model undoubtedly ignores many important elements of galaxy formation, but it suffices to consider a wide range of possible halo mass–luminosity relations. For example, we do not take into account that gas should cycle through the interstellar medium (ISM) before forming stars. However, in the 'bath tub' model of galaxy formation, galaxies evolve towards a quasi-equilibrium state between mass accretion and star formation such that the ISM maintains roughly constant mass (Dekel & Mandelker 2014). Once this equilibrium is reached, our model more accurately describes star formation.

3.2 From star formation to luminosity

We now convert star formation rate to UV luminosity. UV luminosity is a good tracer of star formation because it is produced only by massive, short-lived stars. We take the standard conversion,

$$\dot{m}_* = K_{\text{UV}} \times L_{\text{UV}}, \tag{16}$$

where L_{UV}[5] is the rest-frame continuum (1500–2800 Å)[6] intrinsic luminosity (without extinction). K_{UV} is a conversion from luminosity to star formation rate, and it is dependent on the initial mass function, metallicity, star formation history, binaries, etc. We take $K_{\text{UV}} = 1.15 \times 10^{-28}\,\text{M}_\odot\,\text{yr}^{-1}/(\text{erg s}^{-1}\,\text{Hz}^{-1})$ from Madau & Dickinson (2014). We will show that K_{UV} will not substantially affect our predictions for the relative cosmic variance of the UVLF, even though it can have significant effects on the UVLF itself.

We do ignore dust in our fiducial model, because the extinction in these sources is only poorly constrained. Models suggest that it is modest, and most importantly is not a strong function of halo mass (Mirocha, Mason & Stark 2020). We do however test the effects of dust on our results at $z < 8$ using an empirical dust correction (Vogelsberger et al. 2020, 'Model A'). In order to match the data when applying this dust correction to our energy-driven model, we set $C = 2$ and $f_{*,\text{max}} = 0.3$, making the galaxies intrinsically brighter at fixed halo mass. For simplicity, we ignore f_{shock} in this case, because it also affects the bright end of the luminosity function.

We also ignore scatter in the halo mass–luminosity relation, which would have the effect of flattening out the exponential drop off of the UVLF, as upward scatter in the luminosity has a larger relative effect on the luminosity function in that regime. While we expect this effect to be small (at least on population-level statistics such as the UVLF), we plan to explore it in the future by introducing a scatter in the accretion rates and/or star formation rates.

4 COSMIC VARIANCE IN THE UVLF

In this section, we present the conditional UVLF generated by our model. We provide a fit to the conditional UVLF with a simple Gaussian approximation. We then test the robustness of our results against model choices. Finally, we compare our results to recent works.

We show the conditional UVLF $\phi_{\text{cond}}(M_{\text{AB}}, z, \delta_b, R_e)$ and its 2σ scatter due to cosmic variance in Fig. 3. As in the CMF, cosmic variance increases with increasing galaxy luminosity and also with increasing redshift. The data points shown in Fig. 3 are from Bouwens et al. (2015, 2016); for a more in-depth analysis of this model's agreement with current data, see Furlanetto et al. (2017).

The mapping from halo mass to luminosity in our model is nearly independent of environment because accretion is also nearly independent of environment (see Fig. 2). Thus, nearly all of the cosmic variance of the UVLF comes directly from the variance in the CMF (see Fig. 1). Similarly, simulations by Lovell et al. (2020) find the star formation rate of a galaxy is independent of the dark matter environment, although they compare star formation rates at fixed stellar mass, not halo mass.

4.1 Calculating cosmic variance

As shown in Fig. 3, the amount of cosmic variance in a given galaxy formation model will depend on luminosity, redshift, and the survey characteristics. In this section, we provide a simple descriptor of cosmic variance across all these parameters. We quantify cosmic variance ε_{cv} as the relative standard deviation of the conditional

---

[4]We require $f_{\text{shock}} \leq 1$, and we smooth the function near where $f_{\text{shock}} \to 1$ in order to ensure that it is smoothly differentiable.

[5]For the remainder of the paper, we will display luminosity as absolute and apparent AB magnitudes (M_{AB} and m_{AB}).

[6]This wavelength range corresponds to H band in the redshift range of $z \approx$ 5–9, and K band for $z \approx$ 8–12.
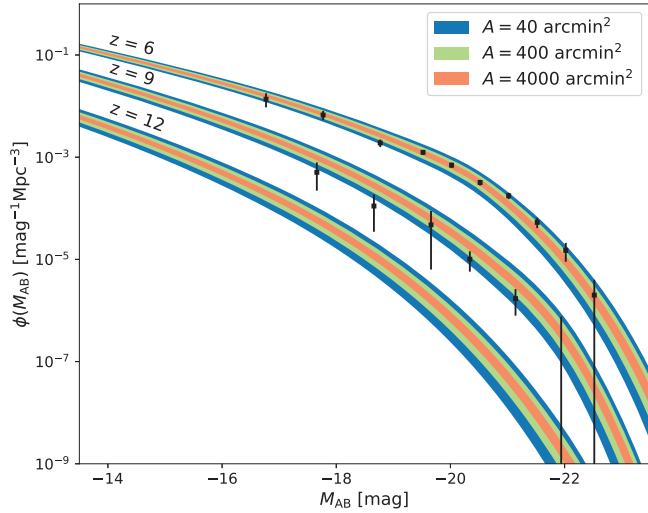
Figure 3. The UVLF and its 2σ scatter from cosmic variance for three selections of survey area at three redshifts ($z = 6$, 9, and 12; $z = 9$ and 12 are off-set in log space by -0.5 and -1 dex, respectively). The scatter in the UVLF increases at the bright end and for smaller survey areas, similar to the CMF in Fig. 1. The survey areas $A = 40$, 400, 4000 arcmin^2 have volumes equivalent to spheres with radii $R_e \approx 29$, 63, 135 Mpc at $z = 6$ (redshift bin $\Delta z = 1$). The data points are from Bouwens et al. (2015, 2016).

UVLF at fixed redshift, apparent magnitude, survey area, and redshift bin width[7]:

$$\varepsilon_{cv}^2 = \frac{\langle \phi_{cond}^2 \rangle - \langle \phi_{cond} \rangle^2}{\langle \phi_{cond} \rangle^2}, \tag{17}$$

with $\langle \phi_{cond}^n \rangle$ defined as

$$\langle \phi_{cond}^n \rangle = \int \phi_{cond}^n(m_{AB}, z, \delta_b, R_e)\, p(\delta_b | R_e, z)\, d\delta_b, \tag{18}$$

where R_e is determined from the survey area and redshift bin width as described in Section 2.3. Fig. 4 shows ε_{cv} as a function of survey area for various redshifts and apparent magnitudes (all with $\Delta z = 1$). This definition of ε_{cv} uses ϕ_{cond} at fixed z, but applies it to the entire volume defined by A and Δz. This approximation breaks down if cosmic variance evolves significantly over the range defined by Δz. Thus, the choice of Δz should be made with care, especially at lower z where ε_{cv} evolves most rapidly in a relative sense (ε_{cv} evolves more rapidly in an absolute sense at high z). Over $\Delta z = 1$, ε_{cv} evolves 10–30 per cent (at $z = 14$ and 5, respectively). Choosing $\Delta z < 0.5$ keeps the change in ε_{cv} below 10 per cent for most cases.[8]

Fig. 4 shows that the relative importance of cosmic variance varies widely across the galaxy population, with a strong dependence on survey parameters. Using a redshift bin width $\Delta z = 1$, ε_{cv} is low at the faint end of the UVLF ($m_{AB} = 32$), ranging from ∼5 per cent at large survey area (1000 arcmin^2) to ∼15 per cent at small survey area (1 arcmin^2) at $z = 6$. As redshift increases, so does ε_{cv}; at $z = 12$ and $m_{AB} = 32$, ε_{cv} ranges from ∼12 per cent at large survey area (1000 arcmin^2) to ∼35 per cent at small survey area (1 arcmin^2).

Cosmic variance also increases significantly at the bright end of the UVLF. At $m_{AB} = 26$, ε_{cv} ranges from ∼15 per cent at large survey area (1000 arcmin^2) to ∼40 per cent at small survey area (1 arcmin^2)

---

[7]Our definition of m_{AB} assumes the galaxy is at the specified 'fixed redshift', regardless of redshift bin width.

[8]Our public PYTHON package GALCV can be used to explore the evolution of ε_{cv} over any desired parameter.
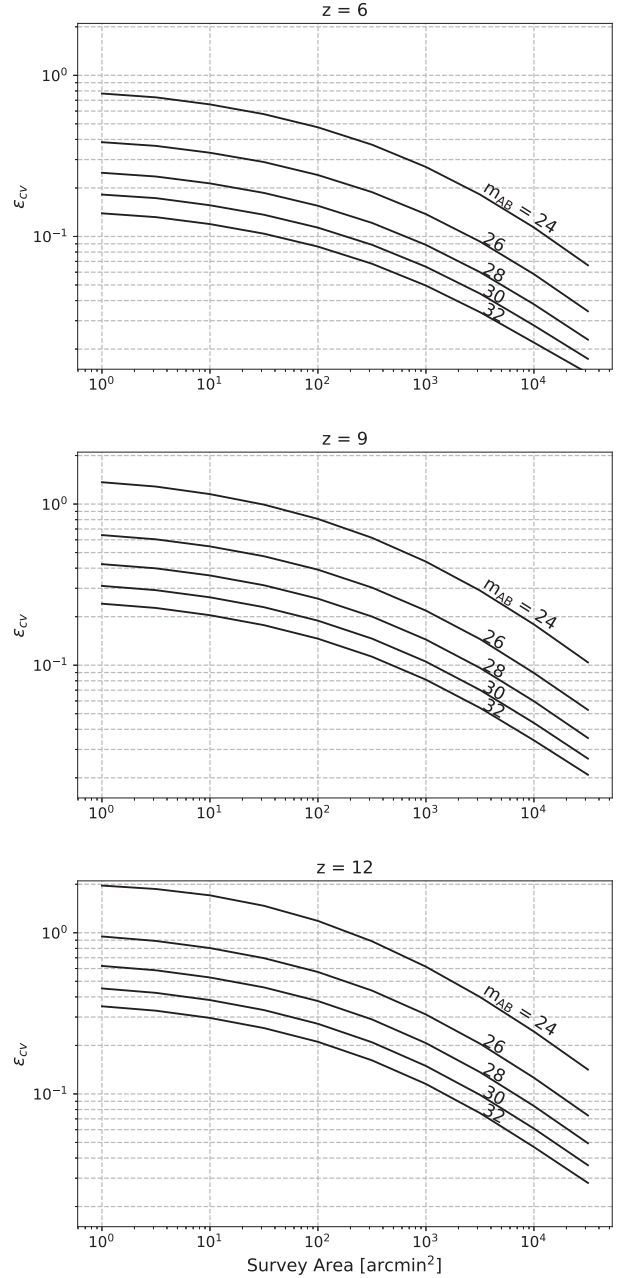
Figure 4. The relative cosmic variance of the UVLF ε_{cv} as a function of survey area for various apparent magnitudes with a redshift bin width of $\Delta z = 1$ (black lines; shown from apparent magnitude $m_{AB} = 32$ on the bottom, decreasing by $\Delta m_{AB} = 2$ towards the top).

at $z = 6$. At $z = 12$ and $m_{AB} = 26$, ε_{cv} ranges from ∼30 per cent at large survey area (1000 arcmin^2) to ∼90 per cent at small survey area (1 arcmin^2).

Cosmic variance flattens out at small survey areas, which is largely due to the effects of the pencil-beam shape of surveys. Even at small survey areas, such a geometry still contains a relatively large range of environments due to its elongated shape. This effect keeps cosmic variance much lower than what one would obtain with a spherical region of the same volume.

We approximate ε_{cv} with a simple functional form; a polynomial in $\log_{10}(\varepsilon_{cv})$ fits well,

$$\log_{10}(\varepsilon_{cv}) \approx \Psi A^\gamma + b, \tag{19}$$

Table 1. Parameters for fits to ε_{cv} (equation 19). We provide ε_{cv} on a much wider range of parameters via our public PYTHON package GALCV (see Data Availability section).

Redshift	App. UV mag.	Fit parameters		
z	m_{AB}	Ψ	γ	b
6	32	-0.223	0.167	-0.608
	30	-0.189	0.184	-0.529
	28	-0.174	0.192	-0.410
	26	-0.161	0.200	-0.235
	24	-0.165	0.199	0.070
9	32	-0.198	0.184	-0.399
	30	-0.190	0.188	-0.298
	28	-0.178	0.194	-0.175
	26	-0.173	0.197	-0.002
	24	-0.197	0.188	0.353
12	32	-0.195	0.188	-0.240
	30	-0.188	0.191	-0.137
	28	-0.184	0.193	-0.001
	26	-0.185	0.193	0.182
	24	-0.202	0.189	0.533

where A is in arcmin^2, and Ψ, γ, and b are fit parameters.[9] Table 1 displays the parameter fits at a selection of redshifts and magnitudes. These fits have a typical/maximum fractional error of 3/5 per cent. We provide ε_{cv} for a wider range of parameters via a public PYTHON package GALCV (see Data Availability section for more details).

With ε_{cv}, we define a linear approximation of the conditional UVLF in a region with density δ_b, angular extent A, and redshift bin width Δz:

$$\phi_{cond}(m_{AB}, z, \delta_b, A, \Delta z)$$
$$= \langle \phi(m_{AB}, z) \rangle \left[1 + \varepsilon_{cv}(A, m_{AB}, z, \Delta z) \frac{\delta_b}{\sigma_{PB}} \right], \quad (20)$$

where $\langle \phi(m_{AB}, z) \rangle$ is the average UVLF and δ_b/σ_{PB} is the density of the region relative to a 1σ fluctuation. This conditional UVLF is similar in construction to that in Livermore et al. (2017), used to fit to lensed high-z galaxy survey data.

4.2 Parameter dependence of ε_{cv}

Our calculations so far have assumed our fiducial choices for the galaxy model (assuming the minimalist energy-regulated prescription) and mass function parameters. Here, we explore how sensitive our results are to variations in these assumptions.

First, we consider how cosmic variance depends on the star formation model. Fig. 5 shows the difference in ε_{cv} when using our fiducial energy-regulated feedback (solid lines) versus a redshift-independent version of feedback (dotted lines, see Section 3.1). There is little difference in the predictions for ε_{cv}. We find a similarly small difference when using momentum-regulated feedback and when using the dust correction from Vogelsberger et al. (2020) for $z < 8$ (the effects of these two are similar to the dotted lines; so are not plotted to reduce clutter). Also, our choice of K_{UV} (see equation 16) will not significantly affect our results, as ε_{cv} is not a particularly strong function of magnitude. These results suggest that cosmic variance is not strongly dependent on the details of star formation or dust.

Second, we explore if cosmic variance is strongly affected by large-scale galaxy environment, namely through differences in accretion.
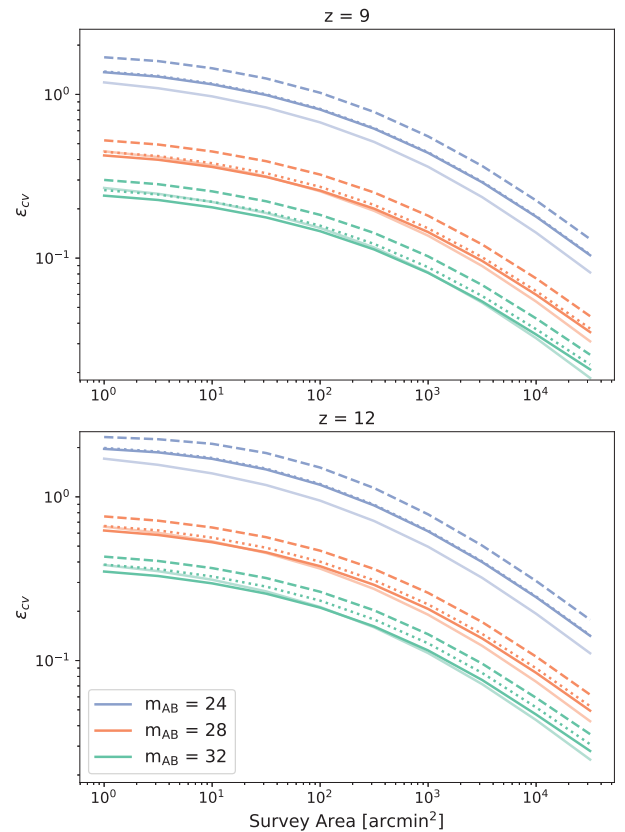
---

[9]Note that our fit assumes the survey subtends a square area on the sky.

Figure 5. The relative cosmic variance of the UVLF ε_{cv} as a function of survey area for various apparent magnitudes (solid lines, same as Fig. 4). The dotted lines show the effects of switching to a z-independent version of star formation. The faded solid lines show the linear bias method for estimating cosmic variance (equation 21). The dashed lines show the effects of using a different method for creating the CMF, specifically the 'Press–Schechter scaling' approach applied to the Trac et al. (2015) mass function (see equation 7). The three sets of lines correspond to magnitudes 32, 28, and 24 (bottom to top).

With the linear halo bias function from equation (5), we approximate

$$\varepsilon_{cv} \approx b_{Trac} \sigma_{PB}, \quad (21)$$

and show it in Fig. 5 (faded solid line).[10] While our full model allows for galaxies to have an environment-dependent accretion and thus luminosity, this linear method does not. However, it provides very similar results to the full method, though it slightly underpredicts cosmic variance at the bright end and overpredicts at the faint end due to the variance in accretion for those haloes (see Fig. 2). This result suggests that approximating the CMF via a simple bias factor is sufficient to capture the effects of cosmic variance (at least to linear order; large density excursions are discussed later in this section).

We conclude that the level of cosmic variance is not sensitive to the particulars of the galaxy formation model. Rather, cosmic variance is dominated by the underlying conditional halo mass function. In Section 2.1, we described an alternate method of creating a CMF: scaling the Trac et al. (2015) mass function by the conditional Press & Schechter (1974) mass function (equation 7). In Fig. 5, we show ε_{cv} when using that CMF (dashed lines). This change results in

---

[10]We connect the bias function b_{Trac} to galaxies using our model's average halo mass–UV luminosity relation.
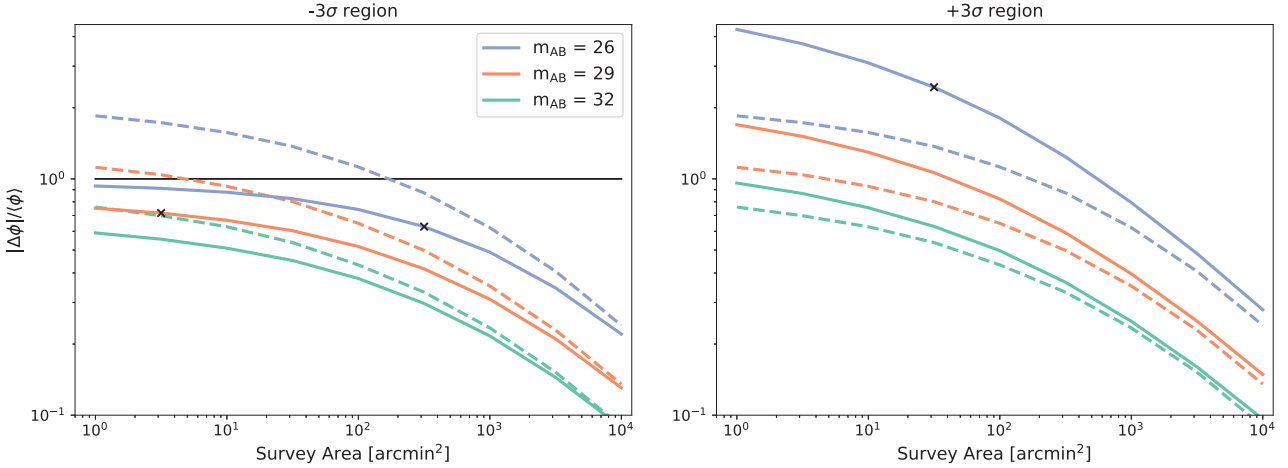
Figure 6. The relative excursion from the average UVLF $\langle\phi\rangle$ for a 3σ underdense region (left) and a 3σ overdense region (right) at redshift 9. We compare the full treatment (solid lines) and the linear approximation in equation (20) (dashed lines). In the left-hand panel, the linear approximation predicts there will be fewer galaxies than the full approach, and can even predict unphysical negative galaxy number densities (when above the horizontal black line). In the right-hand panel, again the linear approximation underpredicts the expected number of galaxies. The black 'x' marks the survey area where our model predicts there to be ~ 1 source in the corresponding magnitude bin. The solid and dashed lines show apparent magnitudes 32, 29, and 26 (bottom to top).
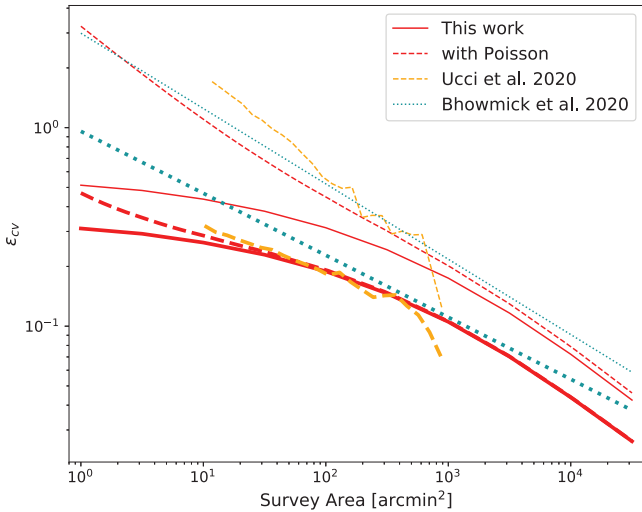
Figure 7. Comparison of cosmic variance predictions at $z = 9$ (with a redshift window of $\Delta z = 1$). The lower set of lines (thick) is at an apparent magnitude of $m_{AB} = 30$, while the upper set of lines (thin) is at $m_{AB} = 27$. The dashed red lines show our cosmic variance predictions with Poisson noise added, for comparison with Ucci et al. (2020).

~ 25 per cent more cosmic variance across the board, the largest effect of any model choice. Thus, in our model, the biggest uncertainty in ε_{cv} is in our understanding of the CMF.

While equation (20) provides a good approximation to the conditional UVLF for $\delta_b/\sigma_{PB} \lesssim 2$, the assumption of a Gaussian bias distribution breaks down at larger density excursions. Fig. 6 shows (at $z = 9$) the difference between using equation (20) (dashed lines) and our full treatment (solid lines) for the conditional UVLF (ϕ) for a 3σ density excursion. Equation (20) underestimates the number of galaxies in very underdense regions (even giving unphysical negative densities at the smallest survey areas), and it also underestimates the number of galaxies in very overdense regions. Equation (20) underpredicts the number of galaxies in both wings, even though it is more reliable near $\delta_b = 0$, because the true bias distribution (at fixed magnitude) is closer to a lognormal. However, where the

deviation from the Gaussian approximation is most pronounced, Poisson shot noise usually dominates the error. Thus, for most applications, equation (20) (ε_{cv}) adequately captures the behaviour of the conditional UVLF.

4.3 Comparison to other works

Here we compare our predictions of cosmic variance to those from two recent models in the literature: Bhowmick et al. (2020) and Ucci et al. (2020).

Bhowmick et al. (2020) provide public estimates for cosmic variance in a redshift range $z = 7$–14 and for apparent H-band magnitudes between $m_{AB} = 25$ and 30. They determine cosmic variance first by calculating the two-point correlation function of galaxies in their simulation box. They then fit the correlation function to a power law and integrate it in a pencil-beam volume (see their equation 2) to estimate the relative cosmic variance. They provide estimates of cosmic variance for all sources *brighter* than the listed magnitude, rather than for sources *at* the listed magnitude. This choice means their estimates of cosmic variance are higher than they would be at fixed magnitude, as cosmic variance increases for brighter sources. However, when we mimic this cumulative method, we find the effect is relatively small (cosmic variance $\lesssim 10$ per cent larger than fixed magnitude method).

Ucci et al. (2020) provide public estimates for cosmic variance in a redshift range $z = 6$–12 and for apparent magnitudes $m_{AB} = 24$–38 (at $z = 9$). Ucci et al. (2020) calculate cosmic variance as the relative standard deviation of galaxy number counts in many pencil-beam subvolumes of their simulation box. They have a slightly more limited survey area coverage, providing estimates between $A = 1$ and 1000 arcmin^2. We compare to the predictions from their 'photoionization' model. These predictions include Poisson variance, making them an estimate of the total variance rather than just cosmic variance.

Fig. 7 shows our predictions compared to those of Bhowmick et al. (2020) (blue dotted lines) and Ucci et al. (2020) (yellow dashed lines) at $z = 9$, with a redshift window of $\Delta z = 1$, at apparent magnitudes of $m_{AB} = 30$ (lower, thick curves) and $m_{AB} = 27$ (upper, thin curves).

As Ucci et al. (2020) report total variance, their predictions should be compared with the red dashed lines (our prediction plus Poisson

noise). Our predictions agree closely with those of Ucci et al. (2020) at $m_{AB} = 30$, and agree within $\sim$50 per cent at $m_{AB} = 27$ (though worsening towards low survey area). Our predictions diverge more significantly at a survey area of 100 arcmin^2, where Ucci et al. (2020) have the fewest independent volumes in their simulation. Also, note that differences in the underlying UVLF can strongly affect the strength of Poisson noise; when Poisson noise begins to dominate, our results should not be too closely compared with those of Ucci et al. (2020).

Our results (red solid lines) are systematically lower than those from Bhowmick et al. (2020) (dotted blue lines). However, our predictions remain within $\sim$25 per cent of each other except at smaller survey areas. At $m_{AB} = 30$, our predictions diverge at an area of $\sim$10 arcmin^2. At $m_{AB} = 27$, our predictions diverge for survey areas where Poisson noise begins to dominate.

Numerical simulations have the benefit of being able to capture the non-linear bias of haloes. This effect, along with differences in Poisson noise from differing mass functions, could help explain the discrepancy between our predictions and those of the simulations at small survey areas.

In comparison to estimates with numerical simulations, the principal benefit of our model is its flexibility. We can test our model with any mass function or star formation and feedback prescription. Simulations must also subtract their intrinsic Poisson noise (which is not known perfectly) to estimate the cosmic variance, while analytic models can easily separate the two effects. Finally, we cover a wider range of redshifts ($z = 5$–15) and magnitudes ($m_{AB} = 22$–38), and we can study larger volumes than simulations.

Our results agree quite well with those of Ucci et al. (2020), over the range to which we can compare, especially at faint luminosities. For bright sources, our estimates are slightly below theirs, but the discrepancy is comparable to the apparent uncertainty in the CMF ($\sim$25 per cent; see Section 4.2). We agree reasonably well with Bhowmick et al. (2020) on large scales as well.

5 IMPACT ON FUTURE SURVEYS

Cosmic variance will provide an unavoidable source of error for next generation telescopes, especially at the highest redshifts. It will dominate over Poisson noise for all but the brightest sources, and it is not easily avoided with deeper observations. Instead, it can only be minimized by probing larger volumes (at the cost of missing the more numerous faint sources) or by splitting up surveys into multiple independent pointings (at the cost of missing large-scale structure and making clustering measurements more difficult).

In this section, we perform a case study of the effects of cosmic variance on two upcoming instruments, *JWST* and the *Roman Space Telescope*. We consider two potential high-redshift surveys: a *JWST* ultradeep (UD) survey (following Mason, Trenti & Treu 2015) along with a much wider-field *Roman Space Telescope* survey (similar to their planned supernova survey, and which we refer to as our SN survey). The UD survey has a detection limit of $m_{AB} \approx 32.0$ and survey area of $A = 40$ arcmin^2, while the SN survey has a detection limit of $m_{AB} \approx 28.3$ and a survey area of $A = 9$ deg^2. To begin, we assume that both are performed over a single contiguous area, requiring at minimum four and 30 separate pointings (neglecting overlap between the pointings).

Additionally, we present a method of fitting an average UVLF to data from the UD and SN surveys simultaneously. Our method makes use of our model effectively as a 'prior' on the cosmic variance in each survey field. The fitting process also accounts for the difference in shape between the local UVLF in each field and the average UVLF that we wish to fit.

5.1 Effects of cosmic variance on UD and SN surveys

We show the effects of cosmic variance on the UVLF of the UD and SN surveys in Fig. 8. The upper panels show the 1σ and 2σ fluctuations of the UVLF at $z = 6$ and 12. The lower panels show ε_{cv} for these surveys (lines) and Poisson shot noise (shaded bands).[11] The vertical lines denote the magnitude limit of the surveys.

A given survey has access to the UVLF over a limited magnitude range, bound on the faint side by the magnitude limit and on the bright side by Poisson noise. In between, the noise floor of cosmic variance determines the maximum accuracy one can achieve in measuring the average UVLF over the accessible magnitude range if using just the one survey.

Splitting up a survey into independent pointings can improve the measurement of the average UVLF. While each individual pointing has higher cosmic variance than a large mosaic, they may be combined, which results in a reduction by the square root of the number of fields. The effect of splitting the UD survey into four pointings is represented by the dashed lines in Fig. 8. Robertson (2010) found a similar improvement in cosmic variance when splitting surveys into independent volumes. See Section 5.4 for more details on the benefits/drawbacks of splitting up surveys.

When interpreting survey results, it is crucial to note that cosmic variance is correlated across all magnitudes. If a survey probes a 1σ underdense region, the expected number counts in *each* magnitude bin will be below the average by 1σ. In contrast, Poisson noise is uncorrelated between each magnitude bin, depending only on the expected number of sources in that bin.

5.2 Measuring the average UVLF

Here we introduce a method to account for cosmic variance in measuring the *average* UVLF of the Universe given data from multiple independent survey volumes. As an example of this method, we simulate mock UD and SN surveys of the UVLF and fit a model that extracts the average UVLF parameters; we then repeat this many times and compare those fits to the 'true' parameters predicted by our model.

In this section, we model the average UVLF as a modified Schechter function:

$$\phi(L)\,dL = \frac{\phi^*}{L^*} \left(\frac{L}{L^*}\right)^\alpha e^{-(L/L^*)^\Gamma}\,dL, \qquad (22)$$

where $\phi(L)\,dL$ is the number density of galaxies with luminosities in the range $(L, L + dL)$, ϕ^* is a normalization constant, L^* is the location of the exponential cut-off, α is the faint-end slope, and Γ is a parameter that governs the strength of the exponential cut-off. $\Gamma = 1$ corresponds to a normal Schechter function. Our models are fit best with $\Gamma = 0.5$, so we will use that value for this paper. We note that our use of $\Gamma = 0.5$ predicts a higher number of bright galaxies than a normal Schechter function. This effect is reminiscent of recent studies of very high-z surveys, which have found that the UVLF can be better fit by a double power law due to an excess of bright galaxies (Bowler et al. 2014, 2020). We do not use a double power law as our models are better fit by the modified Schechter function.

We explore four possible methods to measure the average UVLF of the Universe.

---

[11]Poisson shot noise is model dependent, so we represent it as a band that encompasses the predictions from the three different feedback prescriptions described in Section 3 and the variety of number counts predicted from cosmic variance itself.
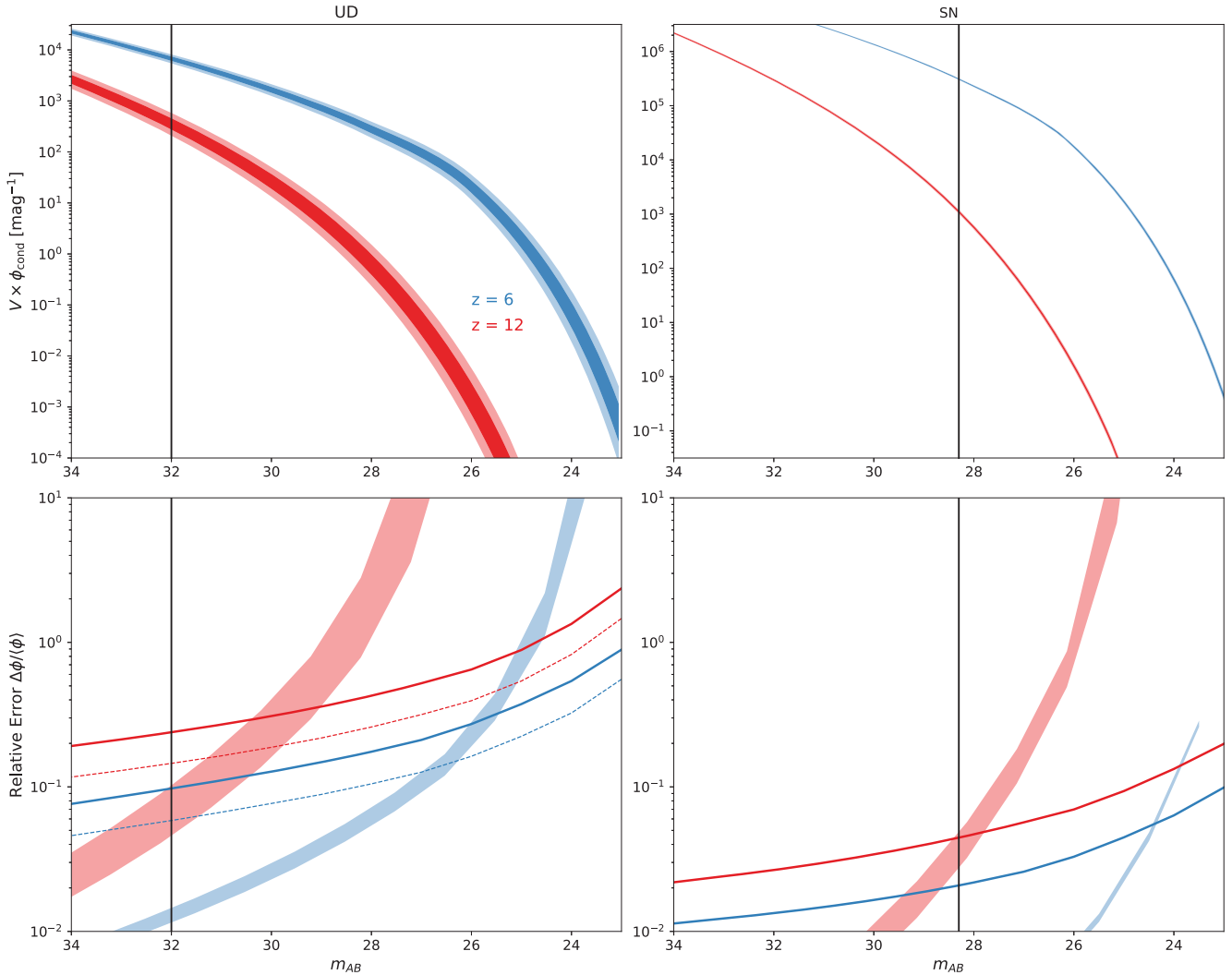
Figure 8. The effects of cosmic variance on the UVLFs of two high-redshift surveys (UD and SN). In the upper panels, the width of the curves represents the 1σ and 2σ (inner and outer shading) ranges of intrinsic UVLFs that could be found in that survey's volume. In the lower panels, the lines show ε_{cv} (upper set of lines are at $z = 12$, lower set at $z = 6$). If the UD survey is broken up into four independent pointings of *JWST*, each smaller subpointing has a higher variance than a large mosaic, but they may be combined for an overall reduction in measuring the average UVLF. This improvement is represented with the thin dashed lines. The vertical black lines are the magnitude limits of the surveys. The shaded band provides an estimate of Poisson shot noise. Cosmic variance acts as a noise floor for measuring the average UVLF, bounded on the faint end by the magnitude limit, and on the bright end by Poisson noise, except for the SN survey, which is entirely dominated by Poisson noise at high-redshift.

(i) *NoCV*. We assume cosmic variance does not exist. Every region of the Universe has the exact same underlying UVLF, so Poisson noise is the only source of error.

(ii) *Naive*. Cosmic variance exists, but we fit the average UVLF without attempting to account for it.

(iii) *Standard*. We fit for the average UVLF using a common method to account for cosmic variance.

(iv) *Full*. Our fiducial method. We fit for the average UVLF parameters using the conditional UVLF developed in this paper[12] (equation 20).

The *NoCV* method assumes (unrealistically!) that cosmic variance does not exist. We simulate galaxy counts for the UD and SN surveys by drawing from the average UVLF that our model predicts, adding

Poisson noise, and then fitting equation (22) to the combined mock data.[13] The solid curves in Figs 9 and 11 show the probability density functions (PDFs) of the best fits of the UVLF to 2000 sets of simulated data with *no* cosmic variance (for $z = 9$ and 12, solid lines). Unsurprisingly, this method recovers the 'true' values (black crosses) of the average UVLF parameters well, as the SN probes the bright end and the UD the faint end, with some overlap between. Of course, cosmic variance *does* exist; this method is only to be used as a comparison to our more realistic scenarios.

For the other three methods we use our model to simulate data for each survey, including cosmic variance. We first draw from

---

[12]A similar method is implemented in Livermore et al. (2017); they consider cosmic variance in lensed surveys, and construct a conditional luminosity function using cosmic variance estimates from Robertson et al. (2014).

[13]We assume in this paper that the UD and SN surveys are perfect, in that they detect every galaxy and are able to accurately place each source in a magnitude bin of width $\Delta m_{AB} = 0.5$ and a redshift bin of $\Delta z = 1$. These are clearly not all accurate assumptions, especially the first one, but this treatment may be taken as a best possible scenario.
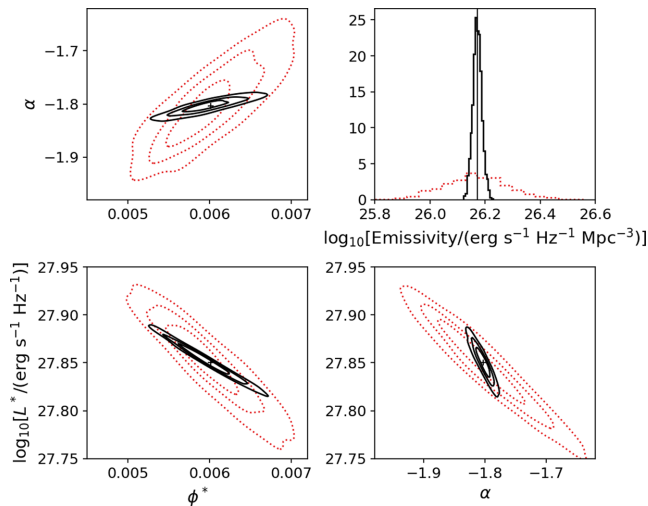
Figure 9. Recovered luminosity function parameters and uncertainties in the *NoCV* (solid contours) and *Naive* (red dotted contours) methods. The contours represent the distribution of best-fitting average UVLF parameters (see equation 22) for 2000 simulated pairs of UD and SN surveys at $z = 9$. Cosmic variance adds a large amount of uncertainty to the determination of the 'true' parameters (black crosses) if not treated properly in the fitting technique. The top right-hand panel shows the emissivity of each fit (integrated down to m_{min}), with the 'true' average emissivity shown as the vertical line. The contours in this and all other figures are equally spaced between zero and the peak value of each normalized distribution.

the distribution of possible density environments for the UD survey $p(\delta_b|R_e, z)$ (see Appendix A) and then use equation (20) to generate the UVLF for that survey.[14] We then calculate the expected number of galaxies in each magnitude bin and apply Poisson shot noise. We repeat these steps for the SN survey. We then repeat this process 2000 times to generate many possible pairs of surveys.

In the *Naive* method, we simply joint fit equation (22) to the 2000 UD+SN mock data pairs with no attempt to correct for cosmic variance. The red dotted lines in Figs 9–12 show the resulting best-fitting PDF for the average UVLF. The recovered parameter range is far wider than the *NoCV* method because the *Naive* method completely ignores the effects of cosmic variance; the measured luminosity functions in the two surveys are *not* the same so cannot easily be reconciled by a single fit.

The *Standard* method, originally developed by Sandage et al. (1979) and used by e.g. Efstathiou et al. (1988) and Bouwens et al. (2015), fits a universal *shape* to the UVLF, ignoring the field-to-field normalization. Then, the normalization is fixed at the end to reproduce the correct total number of galaxies across all surveys. Using this method, we fit to the mock data with cosmic variance. The blue dashed lines in Figs 10 and 12 show the PDFs of the best-fitting parameters for the average UVLF. This method recovers the average UVLF parameters much more accurately than the *Naive* method.

While the *Standard* method is relatively robust to cosmic variance, it does not take into account any changes in the shape of the UVLF due to environment. Additionally, it does not incorporate any information about expected levels of cosmic variance, and it can produce biased results, as seen in this example by its systematic underprediction of the values of ϕ^* and α (most noticeably in Fig. 10, upper left-hand panel).

[14]For $\langle\phi(m_{AB})\rangle$ in equation (20), we use the average UVLF predicted by our model, fit by equation (22) to obtain the 'true' parameters.
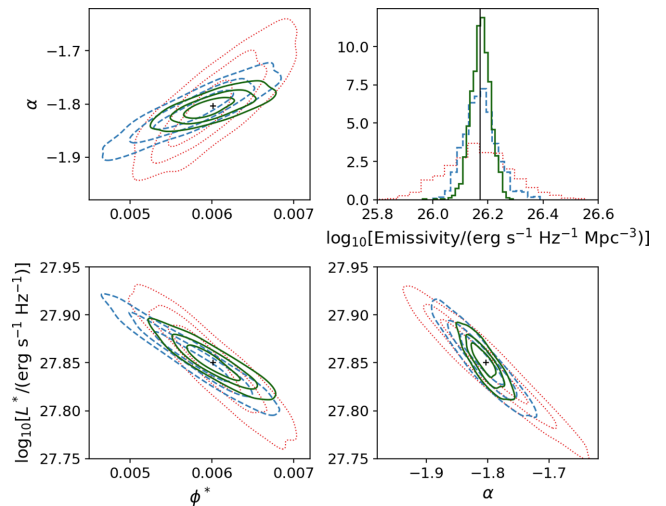

Figure 10. Recovered luminosity function parameters and uncertainties in the *Full* (green solid contours), *Standard* (blue dashed contours), and *Naive* (red dotted contours) methods. The contours represent the distribution of best-fitting average UVLF parameters (see equation 22) for 2000 simulated pairs of UD and SN surveys at $z = 9$. The *Full* and *Standard* methods are significant improvements over the *Naive* method, though the *Full* method does the best job recovering the 'true' parameters (black crosses). The *Standard* method is also slightly biased towards recovering a high L^*, low ϕ^*, and steeper α in this case.
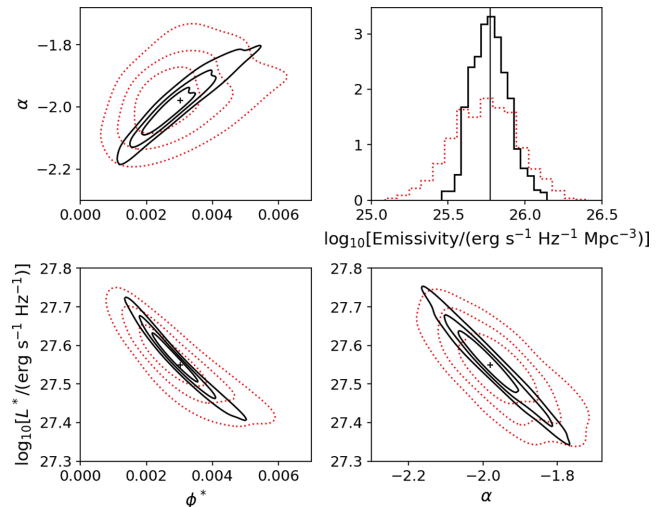

Figure 11. The same as Fig. 9, but at $z = 12$.

Finally, in the *Full* method, we fit equation (20) (with the modified Schechter function in equation 22 as $\langle\phi(m_{AB})\rangle$) simultaneously to each of the pairs of mock surveys, allowing for different values of δ_b for each survey. The green solid lines in Figs 10 and 12 show the PDFs of best-fitting parameters for the average UVLF. Unsurprisingly (because we are fitting with the same function used to generate the mock data), the 'true' parameters are recovered better than with the *Standard* method.

The upper right-hand panels of Figs 9–12 show the total emissivity of the Universe as inferred from the parameters of the best fit (integrating down to m_{min}), compared to the 'true' average emissivity that our model predicts (vertical line). The *Full* method does a slightly better job at recovering the average emissivity of the Universe
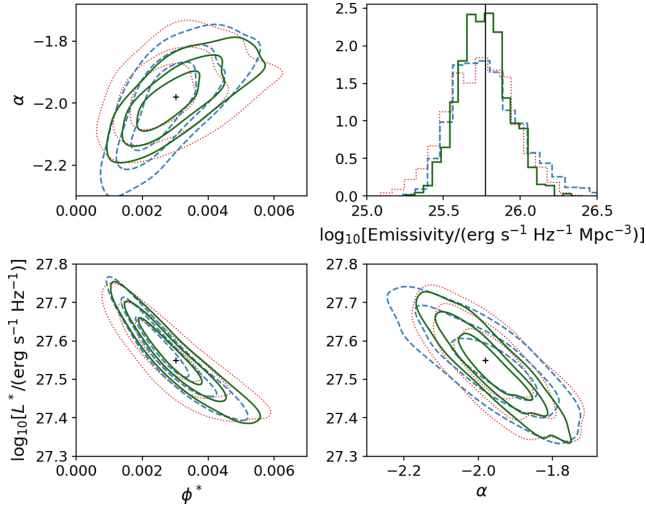
Figure 12. The same as Fig. 10, but at $z = 12$. The difference between the *Full*, *Standard*, and *Naive* models is less pronounced at $z = 12$, though the *Full* method still performs best, and with the least amount of bias.

compared to the *Standard* method, and both do much better than the *Naive* method.

While it is certainly to be expected that the *Full* method performs better than the *Standard* method in our calculations (given that we use our model to generate the mock data *and* to fit to the data), the *Full* method still has benefits. First, it provides estimates of the dark matter overdensity δ_b for each survey field, while the *Standard* method by design throws out field-to-field variance information. Thus, the *Full* method can be used to test our understanding of cosmic variance, because it effectively has a prior on the level of cosmic variance allowed. It penalizes very high field-to-field variance, unlike the *Standard* method that effectively uses a flat prior on the amount of cosmic variance that is allowed during fitting. If real data were fit with the *Full* and *Standard* methods, and the *Standard* method provided a better fit, that would indicate that our understanding of cosmic variance is flawed. We could use our model to investigate *where* and *why* our understanding of cosmic variance breaks down in terms of our physically motivated inputs.

One could investigate the time evolution of UVLF parameters after determining the best-fitting values at a variety of redshifts. However, this experiment would need to be done with care, as Figs 9–12 show that the UVLF parameters are highly correlated. Thus, their time evolution must be fit jointly and with a good estimation of their covariance. That covariance can depend strongly on the treatment of cosmic variance.

5.3 The benefits of multiple surveys

Next, we consider the importance of measuring the UVLF with multiple complementary surveys. Fig. 13 shows the range of best-fitting parameters for the $z = 9$ UVLF when fitting to SN survey data alone (dotted contours), the UD survey alone (dashed contours), and with both simultaneously fit (solid contours, identical to those in Fig. 10). The SN survey alone provides good constraints on ϕ^* and L^*, but the faint-end slope α is constrained better by the joint fit than either survey alone.

At $z = 12$, shown in Fig. 14, the combination of these two surveys is even more crucial, as neither survey can provide good constraints on any parameter by itself.
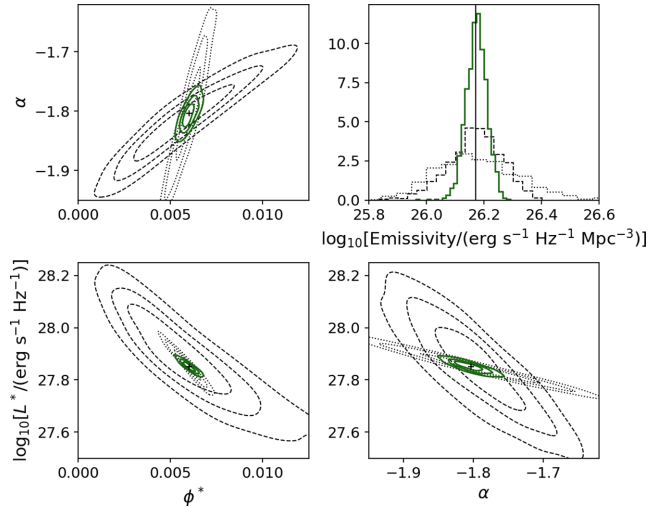

Figure 13. The range of parameters obtained with the SN survey alone (dotted contours), the UD survey alone (dashed contours), and with both simultaneously fit (green solid contours, same as those in Fig. 10). The upper right-hand panel shows the distribution of emissivities calculated from these distributions.
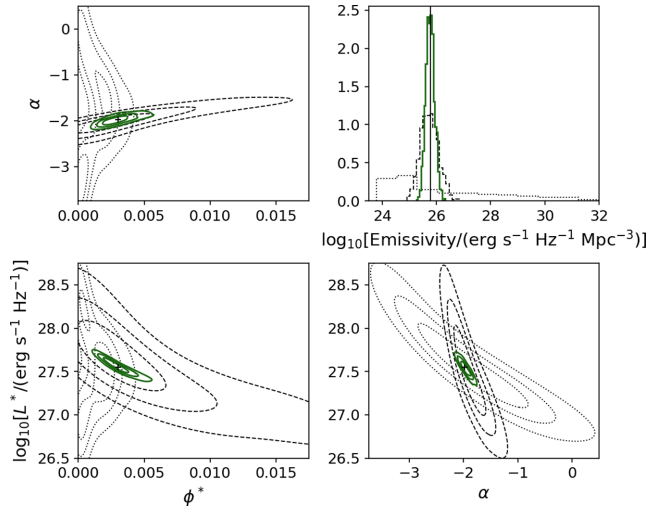

Figure 14. Same as Fig. 13, but at $z = 12$. Each individual survey has much lower constraining power alone.

We also investigated the effects of splitting up the UD survey into four independent pointings and rerunning the *Full* and *Standard* methods. This method gives a significantly better determination of the average number density of very faint sources. However, it only results in a slightly better determination of the average UVLF parameters, because the faint-end slope α is not very sensitive to cosmic variance, and the SN survey dominates the constraints of ϕ^* and L^*.

We see that tiered surveys, including both wide and deep strategies, will be essential for providing an accurate census of the high-z galaxy population.

5.4 Time allocation and survey design strategies

One important use for our results is to identify survey design strategies that result in the best constraints on the average UVLF parameters (Figs 10 and 12). One could use simulations of our model to optimize the design given constraints on telescope time, survey

depth, and area, but here we provide a strategy for a good initial guess.

Given a single magnitude bin and an error requirement ϵ_{req} in measuring the average number counts in that bin, there is a minimum survey area below which cosmic variance will exceed the error requirement. For example, say we wish to design a survey that measures the average UVLF at $z = 9$ (and $\Delta z = 1$) at apparent magnitudes of 30 and 26 with contributions from cosmic variance at those magnitudes below 15 per cent and 10 per cent, respectively. Reading off Fig. 4 or using our PYTHON package GALCV, we find that these would require ∼300 arcmin^2 and ∼2.8 deg^2, respectively.

Alternatively, if we were willing to split each survey into independent pointings, we could satisfy the same error requirements with four ∼3 arcmin^2 surveys down to $m_{AB} = 30$ and nine ∼0.11 deg^2 surveys down to $m_{AB} = 26$. This observing plan requires ∼25× less telescope time for the deep survey and ∼3× less telescope time for the wide-field survey. Splitting up surveys is an especially efficient way to mitigate cosmic variance for narrow surveys because the curves in Fig. 4 are flattest at small survey area, so there is little penalty for moving to even narrower independent pointings. However, we do note that our model does not include non-linear clustering that may become more important in such narrow survey fields.

Unfortunately, splitting up a survey into smaller and smaller subpointings is not without its drawbacks. Large mosaics can be used to measure clustering of galaxies; splitting up a survey leaves many spatial scales inaccessible, and clustering is typically more difficult to measure in the radial direction. Also, multiple small fields of view can miss interesting large-scale structures such as protoclusters. Splitting surveys also increases observing overhead and survey design complexity. An efficient compromise would be a tiered approach: the majority of a survey's area is in one contiguous location, while a smaller fraction is split into a few independent pointings to calibrate for cosmic variance.

6 CONCLUSIONS

Cosmic variance will be an unavoidable source of error for next generation telescopes when measuring average properties of the Universe, especially at higher redshifts. Cosmic variance will dominate over Poisson noise for all but the brightest sources. This study integrates cosmic variance into the galaxy model developed in Furlanetto et al. (2017). We first consider how star formation rates vary with environment in the model. Next, we construct a conditional UVLF and provide its linear approximation for a wide variety of survey parameters with the parameter ε_{cv} via equation (20). We then study what parts of our model are most important in determining ε_{cv}. Finally, we propose a method for using these estimates as a prior on cosmic variance to improve fitting luminosity functions to high-z data.

In our model, the choice of star formation and feedback prescriptions has little effect on the relative strength of cosmic variance, and haloes of fixed mass are similar in all environments. Therefore, the main driver of cosmic variance in the UVLF is cosmic variance in the underlying dark matter halo population. The halo mass function is also the main driver in the uncertainty in our model; a more accurate conditional mass function would allow for a better prediction of cosmic variance.

A simple dark matter halo bias function along with an average halo mass–UV luminosity relation can adequately describe the relative effects of cosmic variance in the UVLF, except for density excursions exceeding ∼2σ. In those regions, cosmic variance becomes non-Gaussian, and a full treatment is required.

We provide linear approximations of cosmic variance via ε_{cv} in terms of apparent (rest-UV) AB magnitude, survey area, and redshift. This approximation may be easily applied to any average UVLF via equation (20). We provide a public PYTHON package GALCV for easy access to our results. This package provides values of ε_{cv} over a wide range of redshifts, magnitudes, survey areas, and redshift bin widths. It also includes two options for the conditional mass function used, which can be used as an estimate of the model uncertainty in the value of ε_{cv} (see Data Availability section for more details). We compare our results with cosmic variance predictions from simulations (Bhowmick et al. 2020; Ucci et al. 2020) and find good general agreement except at the smallest survey volumes (where Poisson noise begins to dominate and non-linear halo bias could be significant), or at volumes that are sizable fractions of their simulations' box size.

We also present a method for using our model as a prior on cosmic variance when fitting a UVLF to galaxy survey data. This method can inform our understanding of cosmic variance while also improving the quality of and reducing the bias in fitting the UVLF. It allows us to quantify the gains from splitting surveys into independent pointings and combining independent observations. In particular, we have shown that the combination of a shallow wide survey and a deep narrow survey is essential for fully constraining the UVLF. We also show that splitting up a survey can be an effective way to reduce the effects of cosmic variance.

Our model treats galaxy formation in a very simple manner. The primary simplification is in modelling only the *average* galaxy population in a given environment. We also ignore the effects of dust, mergers, scatter in the halo mass–UV luminosity relation, the evolution of the IMF, and the spatial distribution of star formation within a dark matter halo. Fortunately, these shortcomings pertain to (1) the details of star formation, which we have shown hardly affect the *relative* cosmic variance results ε_{cv}; and (2) individual galaxies, which are likely averaged out (to an extent) when considering cosmic variance in an ensemble of galaxies.

An understanding of cosmic variance is essential for quantifying the uncertainty in future surveys with observatories like *JWST* and the *Roman Space Telescope*. We hope that our flexible model, and the method we have introduced to incorporate cosmic variance explicitly into fitting multiple fields, can offer better constraints not just on the galaxy luminosity function but also on cosmic variance itself.

DATA AVAILABILITY

We provide a public PYTHON package GALCV that can be used to calculate ε_{cv} for a wide range of survey parameters (redshift, apparent AB magnitude, survey area, and redshift bin width). This package may be installed in a PYTHON environment via 'pip install galcv' and then imported via 'import galcv'. The code for this package and instruction for its installation and use may be found online at https://github.com/adamtrapp/galcv. Requests for bug fixes and suggestions for additions are welcome.

REFERENCES

Akeson R. et al., 2019, preprint (arXiv:1902.05569)
Atek H., Richard J., Kneib J.-P., Schaerer D., 2018, MNRAS, 479, 5184
Barkana R., Loeb A., 2004, ApJ, 609, 474
Behroozi P., Wechsler R. H., Hearin A. P., Conroy C., 2019, MNRAS, 488, 3143
Bhowmick A. K., Somerville R. S., Di Matteo T., Wilkins S., Feng Y., Tenneti A., 2020, MNRAS, 496, 754
Bond J. R., Cole S., Efstathiou G., Kaiser N., 1991, ApJ, 379, 440
Bouwens R. J. et al., 2015, ApJ, 803, 34
Bouwens R. J. et al., 2016, ApJ, 830, 67
Bowler R. A. A., Jarvis M. J., Dunlop J. S., McLure R. J., McLeod D. J., Adams N. J., Milvang-Jensen B., McCracken H. J., 2020, MNRAS, 493, 2059
Bowler R. A. A. et al., 2014, MNRAS, 440, 2810
Dekel A., Mandelker N., 2014, MNRAS, 444, 2071
Dore O. et al., 2019, BAAS, 51, 341
Efstathiou G., Ellis R. S., Peterson B. A., 1988, MNRAS, 232, 431
Eisenstein D. J., Hu W., 1998, ApJ, 496, 605
Fakhouri O., Ma C.-P., Boylan-Kolchin M., 2010, MNRAS, 406, 2267
Faucher-Giguère C.-A., Kereš D., Ma C.-P., 2011, MNRAS, 417, 2982
Faucher-Giguère C.-A., Quataert E., Hopkins P. F., 2013, MNRAS, 433, 1970
Finkelstein S. L. et al., 2015, ApJ, 810, 71
Furlanetto S. R., Mirocha J., Mebane R. H., Sun G., 2017, MNRAS, 472, 1576
Gao L., White S. D. M., 2007, MNRAS, 377, L5
Goerdt T., Ceverino D., Dekel A., Teyssier R., 2015, MNRAS, 454, 637
Hayward C. C., Hopkins P. F., 2017, MNRAS, 465, 1682
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Kitzbichler M. G., White S. D. M., 2007, MNRAS, 376, 2
Krumholz M. R., Burkhart B., Forbes J. C., Crocker R. M., 2018, MNRAS, 477, 2716
Lacey C., Cole S., 1993, MNRAS, 262, 627
Livermore R. C., Finkelstein S. L., Lotz J. M., 2017, ApJ, 835, 113
Loeb A., Furlanetto S. R., 2013, The First Galaxies in the Universe. Princeton Univ. Press, Princeton, NJ
Lovell C. C., Vijayan A. P., Thomas P. A., Wilkins S. M., Barnes D. J., Irodotou D., Roper W., 2020, priprint (arXiv:2004.07283)
Madau P., Dickinson M., 2014, ARA&A, 52, 415
Mason C. A., Trenti M., Treu T., 2015, ApJ, 813, 21
McBride J., Fakhouri O., Ma C.-P., 2009, MNRAS, 398, 1858
McKinney W., 2010, in van der Walt S., Millman J., eds, Proceedings of the 9th Python in Science Conference (SciPy 2010). SciPy, p. 51
Mesinger A., Furlanetto S., Cen R., 2011, MNRAS, 411, 955
Mirocha J., Furlanetto S. R., Sun G., 2017, MNRAS, 464, 1365
Mirocha J., Mason C., Stark D. P., 2020, MNRAS, 498, 2645
Mo H. J., White S. D. M., 1996, MNRAS, 282, 347
Moster B. P., Somerville R. S., Newman J. A., Rix H.-W., 2011, ApJ, 731, 113
Muñoz J. A., Trac H., Loeb A., 2010, MNRAS, 405, 2001
Newman J. A., Davis M., 2002, ApJ, 564, 567
Oesch P. A., Bouwens R. J., Illingworth G. D., Labbé I., Stefanon M., 2018, ApJ, 855, 105
Patiri S. G., Betancort-Rijo J. E., Prada F., Klypin A., Gottlöber S., 2006, MNRAS, 369, 335

Perez F., Granger B. E., 2007, Comput. Sci. Eng., 9, 21
Planck Collaboration XIII, 2016, A&A, 594, A13
Press W. H., Schechter P., 1974, ApJ, 187, 425
Robertson B. E., 2010, ApJ, 713, 1266
Robertson B. E., Ellis R. S., Dunlop J. S., McLure R. J., Stark D. P., McLeod D., 2014, ApJ, 796, L27
Rubiño-Martín J. A., Betancort-Rijo J., Patiri S. G., 2008, MNRAS, 386, 2181
Sandage A., Tammann G. A., Yahil A., 1979, ApJ, 232, 352
Sheth R. K., 1998, MNRAS, 300, 1057
Somerville R. S., Lee K., Ferguson H. C., Gardner J. P., Moustakas L. A., Giavalisco M., 2004, ApJ, 600, L171
Spergel D. et al., 2015, preprint (arXiv:1503.03757)
Stark D. P., Loeb A., Ellis R. S., 2007, ApJ, 668, 627
Sun G., Furlanetto S. R., 2016, MNRAS, 460, 417
Tinker J., Kravtsov A. V., Klypin A., Abazajian K., Warren M., Yepes G., Gottlöber S., Holz D. E., 2008, ApJ, 688, 709
Trac H., Cen R., Mansfield P., 2015, ApJ, 813, 54
Tramonte D., Rubiño-Martín J. A., Betancort-Rijo J., Dalla Vecchia C., 2017, MNRAS, 467, 3424
Trenti M., Stiavelli M., 2008, ApJ, 676, 767
Ucci G. et al., 2020, preprint (arXiv:2004.1109)
Vale A., Ostriker J. P., 2004, MNRAS, 353, 189
van den Bosch F. C., Jiang F., Hearin A., Campbell D., Watson D., Padmanabhan N., 2014, MNRAS, 445, 1713
van der Walt S., Colbert S. C., Varoquaux G., 2011, Comput. Sci. Eng., 13, 22
Virtanen P. et al., 2020, Nat. Methods, 17, 261
Vogelsberger M. et al., 2020, MNRAS, 492, 5167

APPENDIX A: THE EULERIAN VOLUME CORRECTION

In this appendix, we construct $p(\delta_b|R_e, z)\,d\delta_b$, the fraction of volume in the Universe with linear density between $(\delta_b, \delta_b + d\delta_b)$ when averaged over the Eulerian scale R_e. This distribution is in contrast to $p(\delta_b|R, z)\,d\delta_b$, the fraction of *mass* in the Universe with linear density between $(\delta_b, \delta_b + d\delta_b)$ when averaging over the Lagrangian scale R.

As described in Section 2.2, $p(\delta_b|R, z)$ is by definition equal to a zero-mean Gaussian with variance $\sigma^2(M, z)$, where M is the mass of a region of radius R and average density. Unfortunately, $p(\delta_b|R, z)$ considers a fixed mass scale R,[15] which corresponds to a density-dependent range of different volumes.

Let us choose one fixed Eulerian scale R_e. We consider that scale's corresponding Lagrangian radii,

$$R^3 = R_e^3(1 + \delta_r), \tag{A1}$$

where δ_r is the true, non-linear density of the region. Following Mo & White (1996), the *real* density of a region may be related to the linear density via the following approximation (assuming spherical collapse):

$$\delta_b = -1.35(1 + \delta_r)^{-2/3} + 0.78785(1 + \delta_r)^{-0.58661}$$
$$- 1.12431(1 + \delta_r)^{-1/2} + 1.68647. \tag{A2}$$

Inserting this value into equation (A1), we now have $R(\delta_b|R_e)$, a relation between linear density and Lagrangian radius at fixed Eulerian radius. We convert R to σ via $\sigma(M = 4\pi/3\ \bar{\rho}R^3)$ and convert δ_b to δ_0 via the growth function $\delta_0 = \delta_b/F_g(z)$. That process

---

[15] R is a mass (Lagrangian) scale because it is defined as the radius of a region of mass M *if* that region were at average density. In reality, regions of mass M can have different physical volumes depending on their densities, as we will show in Appendix A.
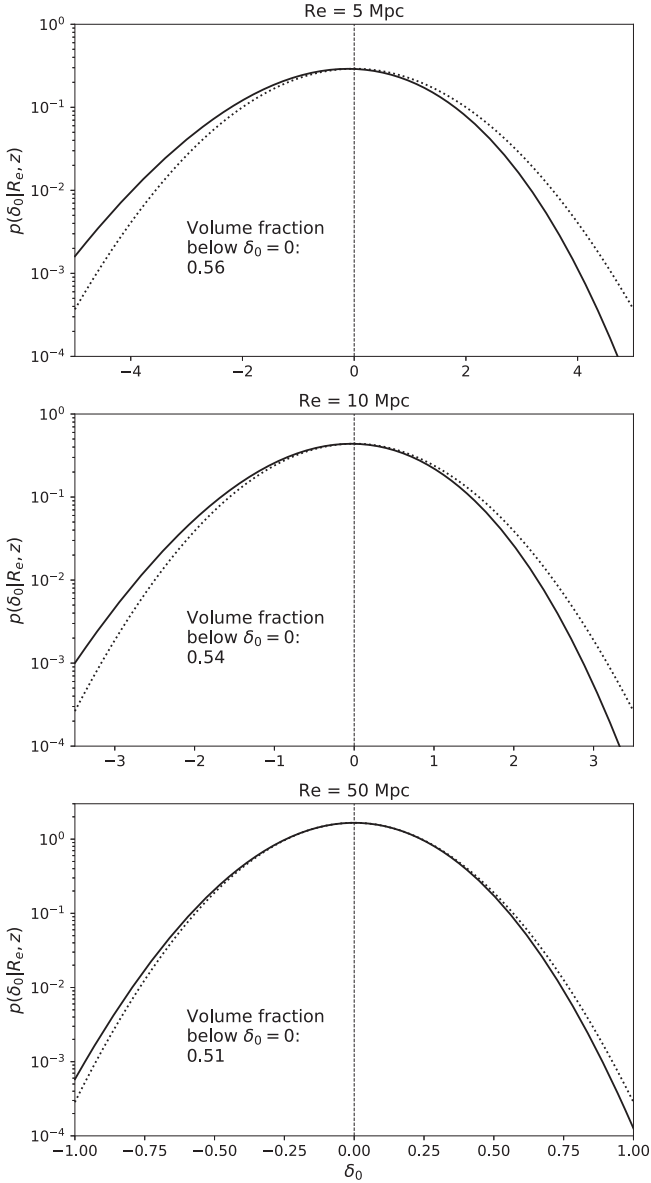
Figure A1. Effects of the Eulerian correction at $z = 9$. The solid lines show $p(\delta_0|R_e, z)$, the distribution of linear densities at fixed scale (densities are extrapolated to $z = 0$ via the growth function). The dotted lines show the Lagrangian distribution of densities $p(\delta_0|R, z)$: Gaussian distributions with standard deviation $\sigma(R_e, z)$. The volume of the Universe that is below average density at each scale is indicated in each panel.

provides $\sigma^2(\delta_0|R_e, z)$, a locus in (σ^2, δ_0) space of constant Eulerian radius R_e.

With $\sigma^2(\delta_0|R_e, z)$, we can use the excursion set formalism to solve for $f_R(\sigma^2|R_e, z)$, the distribution of mass in the Universe that is associated with a region with σ^2 (and thus corresponding R and δ_0) at fixed R_e. The excursion set formalism describes a random walk in dark matter density δ_0 as one averages over first a very large volume (small σ), and then successively smaller volumes (larger σ) centred at a single point in space (Bond et al. 1991; Lacey & Cole 1993). The distribution of random walks that first cross the barrier $\sigma^2(\delta_0|R_e, z)$ defines $f_R(\sigma^2|R_e, z)$.

$f_R(\sigma^2|R_e, z)$ has no analytic solution for an arbitrary barrier shape, so we approximate $\sigma^2(\delta_0|R_e, z)$ as a straight line,[16]

$$B(\sigma^2|R_e, z) = B_0 + B_1\sigma^2, \tag{A3}$$

where B is the density δ_0, and B_0 and B_1 are fit parameters (corresponding to the *y*-intercept and slope, respectively).

Fortunately, the first-crossing distribution f_R for a linear barrier in (σ^2, δ_0) space has been solved analytically by Sheth (1998):

$$f_R(\sigma^2, |R_e, z)\,d\sigma^2$$
$$= \frac{B(0|R_e, z)}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{B^2(\sigma^2|R_e, z)}{2\sigma^2}\right)\frac{d\sigma^2}{\sigma^2}. \tag{A4}$$

This is an Inverse Gaussian distribution.

We convert $f_R(\sigma^2|R_e, z)$, a mass fraction distribution in σ^2, to $p(\delta_b|R_e, z)$, a volume fraction distribution in δ_b, following equation (16) of Sheth (1998):

$$p(\delta_b|R_e, z)\,d\delta_b = \frac{1}{(1 + \delta_r)}f_R(\sigma^2|R_e, z)\,d\sigma^2. \tag{A5}$$

In principle, dividing by the non-linear function $(1 + \delta_r)$ can result in a $p(\delta_b|R_e, z)$ that is not normalized. In practice, $p(\delta_b|R_e, z)$ remains normalized within 1 per cent for all cases we consider.

For the range of redshifts and scales considered in this paper, $p(\delta_b|R_e, z)$ is near to a Gaussian with standard deviation $\sigma(M = 4\pi/3\ \bar\rho R_e^3)$. However, the distribution is skewed towards negative densities, resulting in a boost in the negative wing and suppression in the positive wing, an effect that is most significant for volumes with radii less than $\sim$10 Mpc (see Fig. A1).

At $z = 9$, we find that for regions with scales $R_e = 5$, 10, and 50 Mpc, the fraction of volume in the Universe that is below average cosmic density is 56, 54, and 51 per cent, respectively. These fractions increase slightly at lower redshifts as underdense regions continue to expand relative to overdense regions. This result indicates that surveys will be slightly more likely to probe underdense regions. Using different methods, Muñoz et al. (2010) also found that surveys are more likely to probe an underdense region because of those regions' more rapid cosmic expansion.

---

[16]We approximate $\sigma^2(\delta_0|R_e, z)$ as a line by fitting it to the barrier near where most trajectories cross the barrier: $\delta_0 = 0$.

This paper has been typeset from a TₑX/LᴬTₑX file prepared by the author.