Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **482,** 4528–4552 (2019)
Advance Access publication 2018 November 6

doi:10.1093/mnras/sty3007

The formation and hierarchical assembly of globular cluster populations

Kareem El-Badry [1,2★] Eliot Quataert,[1] Daniel R. Weisz [1] Nick Choksi [1] and Michael Boylan-Kolchin [3]

[1]*Department of Astronomy and Theoretical Astrophysics Center, University of California Berkeley, Berkeley, CA 94720, USA*
[2]*Max Planck Institute for Astronomy, D-69117 Heidelberg, Germany*
[3]*Department of Astronomy, The University of Texas at Austin, Austin, TX 78712, USA*

Accepted 2018 November 2. Received 2018 October 22; in original form 2018 May 9

ABSTRACT

We use a semi-analytic model for globular cluster (GC) formation built on dark matter merger trees to explore the relative role of formation physics and hierarchical assembly in determining the properties of GC populations. Many previous works have argued that the observed linear relation between total GC mass and halo mass points to a fundamental GC–dark matter connection or indicates that GCs formed at very high redshift before feedback processes introduced non-linearity in the baryon-to-dark matter mass relation. We demonstrate that at $M_{\rm vir}(z = 0) \gtrsim 10^{11.5}\,{\rm M}_{\odot}$, a constant ratio between halo mass and total GC mass is in fact an almost inevitable consequence of hierarchical assembly: by the central limit theorem, it is expected at $z = 0$ independent of the GC-to-halo mass relation at the time of GC formation. The GC-to-halo mass relation at $M_{\rm vir}(z = 0) < 10^{11.5}\,{\rm M}_{\odot}$ is more sensitive to the details of the GC formation process. In our fiducial model, GC formation occurs in galaxies when the gas surface density exceeds a critical value. This model naturally predicts bimodal GC colour distributions similar to those observed in nearby galaxies and reproduces the observed relation between GC system metallicity and halo mass. It predicts that the cosmic GC formation rate peaked at $z \sim 4$, too late for GCs to contribute significantly to the UV luminosity density during reionization.

Key words: globular clusters: general – galaxies: formation – galaxies: star clusters: general.

1 INTRODUCTION

Globular clusters (GCs) are relics of star formation under extreme conditions in the early Universe. Although it may soon become feasible to observe young GCs at high redshift as they form (Carlberg 2002; Katz & Ricotti 2013; Boylan-Kolchin 2017a; Renzini 2017; Vanzella et al. 2017; Zick, Weisz & Boylan-Kolchin 2018), at present, most of what we know about GCs comes from observations of the old GC populations of nearby galaxies.

Studies of GCs in the local Universe have highlighted striking differences between galaxies' GC and field star populations. While the galaxy stellar-to-halo mass relation is strongly non-linear, the total mass of GCs within a dark matter halo is a constant fraction of halo mass over almost five decades in mass (e.g. Blakeslee, Tonry & Metzger 1997; Harris, Harris & Alessi 2013; Durrell et al. 2014; Hudson, Harris & Harris 2014; Harris, Harris & Hudson 2015; Harris, Blakeslee & Harris 2017b). The GC populations of most individual halos exhibit bimodality in colour and/or metallicity (Zepf & Ashman 1993; Harris et al. 2006; Peng et al. 2006; Brodie et al. 2012), in contrast to the stars in the central galaxy or the stellar halo. And although GCs were once thought to be simple, uniform stellar populations formed in a single burst, detailed observations reveal evidence of multiple stellar populations and anomalous abundance patterns that remain poorly understood (e.g. Piotto et al. 2015; Bastian & Lardo 2017).

The observed constant GC-to-halo mass ratio and the old ages measured for Milky Way (MW) GCs have led many authors to suggest that most GCs, particularly those that are blue and metal-poor, formed at very early times, before feedback processes introduced non-linearity in the baryon-to-dark matter mass relation (Blakeslee et al. 1997; Kavelaars 1999; Diemand, Madau & Moore 2005; Moore et al. 2006; Bekki et al. 2008; Spitler et al. 2008; Spitler & Forbes 2009; Corbett Moran, Teyssier & Lake 2014; Hudson et al. 2014; Katz & Ricotti 2014; Harris et al. 2015; Trenti, Padoan & Jimenez 2015; Boylan-Kolchin 2017b). Such a formation scenario most directly implies a constant relation between GC mass and halo mass *at the time of GC formation*, but Boylan-Kolchin (2017b) showed that if a constant GC-to-halo mass ratio was set at high redshift, it would be preserved to $z = 0$ during hierarchical assembly.

Star formation is a local process. If GC formation did occur proportional to dark matter halo mass at high redshift, a successful GC

★ E-mail: kelbadry@berkeley.edu

formation theory must attempt to tie the mass of a dark matter halo at high redshift to local gas conditions conducive to GC formation. Doing so is challenging both because the properties of a DM halo do not uniquely determine the baryonic conditions in its central galaxy, even at high redshift (e.g. Wise et al. 2012; O'Shea et al. 2015), and because the local gas conditions required for the formation of massive bound clusters remain imperfectly understood (e.g McKee & Ostriker 2007; Krumholz 2014; Skinner & Ostriker 2015; Tsang & Milosavljevic 2017; Grudić et al. 2018b).

Some numerical studies have begun to resolve aspects of the GC formation process in a cosmological context (Kravtsov & Gnedin 2005; Boley et al. 2009; Trenti et al. 2015; Kimm et al. 2016; Kim, Kim & Ostriker 2016; Mandelker et al. 2017). Because GCs are much smaller than the scales typically resolved in cosmological zoom-in simulations, such works face strong trade-offs between resolution, simulation volume, and final redshift. Simulations reaching the sub-parsec scale resolution required to study details of the GC formation process have therefore to date focused on small volumes and have been terminated at high redshift, making comparison with observations difficult.

A complementary approach, which we take in this work, is to adopt simple prescriptions to predict the GC formation rate and/or the dynamical evolution of GCs in a halo as a function of galaxy-scale gas conditions or the properties of the dark matter halo. This approach, which has been fruitfully employed in a number of previous studies (Ashman & Zepf 1992; Côté, Marzke & West 1998; Beasley et al. 2002; Prieto & Gnedin 2008; Muratov & Gnedin 2010; Tonini 2013; Katz & Ricotti 2014; Li & Gnedin 2014; Kruijssen 2015; Choksi, Gnedin & Li 2018; Pfeffer et al. 2018), makes it possible to efficiently predict the observable GC populations of galaxies at $z = 0$ for a wide range of GC formation models. Such 'semi-analytic' models cannot predict the internal properties of GCs with high fidelity and are not guaranteed to capture all the physical processes relevant to GC formation. However, their simplicity aids their interpretability: because such models have only a few free parameters, they make it straightforward to gauge the sensitivity of observables to different aspects of the GC formation model.

This work models GC formation as the product of 'normal' star formation in the high-density discs of gas-rich galaxies. Motivated by simulations of molecular cloud collapse, we use the ansatz that massive bound clusters form preferentially when the gas surface density exceeds a critical threshold. We apply this ansatz to a semi-analytic gas model built on dark matter merger trees in order to predict the GC populations of halos at $z = 0$. We then explore how varying different aspects of the GC formation prescription changes the epoch at which GCs form, the GC-to-halo mass relation, and the $z = 0$ colour distributions of individual galaxies' GCs. In contrast to some previous works (e.g. Beasley et al. 2002; Tonini 2013; Amorisco 2018), our model does not explicitly assume separate formation modes for blue and red GCs; we simply predict the metallicity and colour of each GC based on the conditions in the galaxy in which it formed. Because the model is built on dark matter merger trees, it does not allow for straightforward predictions of GC formation divorced from dark matter (e.g. Zhao 2005; van Dokkum et al. 2018). However, we emphasize that GC formation in our model is most directly tied to baryonic conditions: the merger trees serve primarily to keep track of the gas conditions throughout the formation history of a GC population.

In agreement with previous work, we find that semi-analytic GC formation models can broadly reproduce many aspects of the observed GC population. However, one of our main results is that some observed GC scaling relations, particularly the constant GC-to-halo mass ratio, are primarily consequences of hierarchical assembly and are thus insensitive to details of the GC formation process.

The rest of this paper is organized as follows. In Section 2, we describe the assumptions and implementation of our semi-analytic model. We present the $z = 0$ GC populations predicted by our model and explore the model's sensitivity to several free parameters in Section 3. We summarize and discuss our findings in Section 4. We provide additional details of the underlying model in the appendices.

2 MODEL

The basic idea of our model is that massive bound clusters, including the progenitors of GCs, form primarily when the gas surface density exceeds a critical value. We motivate this ansatz in Section 2.1 and discuss the model's implementation in Section 2.2.

2.1 GC formation at high surface density

Massive bound clusters like the progenitors of GCs are not a typical outcome of star formation under normal ISM conditions at low redshift. Although a large fraction of stars in nearby galaxies form in clusters, most clusters become gravitationally unbound and disrupted within a few dynamical times (for a review, see Lada & Lada 2003). This cluster 'infant mortality' owes to the fact that the star formation efficiency of giant molecular clouds (GMCs) in galaxies with MW-like gas surface densities is low ($\sim$1 per cent), so initially bound clusters become unbound when stellar feedback expels most of a cluster's gas mass and shallows the gravitational potential (Tutukov 1978; Geyer & Burkert 2001; Bastian & Goodwin 2006; Baumgardt & Kroupa 2007). Clusters are more likely to remain bound if the star formation efficiency is high.

Massive young star clusters *are* observed in nearby galaxies with higher gas densities than the MW (e.g. Portegies Zwart, McMillan & Gieles 2010), and the fraction of stars formed in long-lived clusters is observed to be higher in high-density environments (Larsen & Richtler 2000; Keto, Ho & Lo 2005; Goddard, Bastian & Kennicutt 2010; Johnson et al. 2016). Theoretical star formation models suggest that environments of high density and pressure are conducive to the formation of proto-GC-like clusters (Elmegreen & Efremov 1997; Murray, Quataert & Thompson 2010; Kruijssen 2012, 2015), because (a) the free-fall time becomes shorter than the few-Myr massive stellar evolution time-scale, meaning that a large fraction of a gas cloud can turn into stars before the first supernovae explode (see Elmegreen 2017 and references therein), and (b) the self-gravity of a cloud increases more steeply with density than the energy injected by stellar feedback (Murray et al. 2010; Thompson & Krumholz 2016), so that at sufficiently high densities, the feedback energy budget is insufficient to prevent runaway star formation.

Grudić et al. (2018b, hereafter G18) recently argued that the formation of massive bound star clusters depends most directly on high gas *surface density*, as opposed to volume density, escape velocity, pressure, or other GMC properties. Using idealized cloud-collapse simulations of individual GMCs, they studied how the star formation efficiency, ε (i.e. the fraction of gas in a collapsing cloud that is converted to stars), and the cluster formation efficiency, Γ (the fraction of stars formed in bound clusters), scale with the structural parameters of a GMC. G18 found that at fixed cloud geometry, ε and Γ are primarily functions of the gas surface density, $\Sigma_{\rm GMC}$,

independent of the cloud mass and size.[1] In particular, G18 found ε to plateau, at a maximum value of order unity, for $\Sigma_{\rm GMC} \gg \Sigma_{\rm crit}$, and to fall off as $\epsilon \sim \Sigma_{\rm GMC}^{-\alpha}$, where $\alpha \sim 1$, at $\Sigma_{\rm GMC} \ll \Sigma_{\rm crit}$. A qualitatively similar scaling with $\Sigma_{\rm GMC}$ was found for Γ (Grudic et al., private communication). G18 found $\Sigma_{\rm crit} \approx 3000\ {\rm M_\odot\ pc^{-2}}$; several other works have predicted critical densities for GC formation in the range $\Sigma_{\rm crit} = 10^{3-4}\ {\rm M_\odot\ pc^{-2}}$ (Beasley et al. 2002; Elmegreen 2008; Fall, Krumholz & Matzner 2010; Kruijssen 2012; Kim et al. 2016; Raskutti, Ostriker & Skinner 2016; Li et al. 2017).

2.2 Fiducial model implementation

To predict the GC population of a dark matter halo at $z = 0$, we estimate the GC formation rate, based on estimates of the star formation rate (SFR) and gas surface density, throughout its assembly history. We propagate GCs formed at each point in the merger tree to $z = 0$, assuming that at each merger, the descendant halo inherits the GC populations of both its progenitors. The model is illustrated schematically in Fig. 1.

2.2.1 *Merger trees*

We generate merger trees based on extended Press Schechter theory (Bond et al. 1991), using the Monte Carlo algorithm described in Parkinson, Cole & Helly (2008).[2] Merger trees generated with this method have been shown to reproduce the statistical properties and mass accretion histories of merger trees extracted from *N*-body simulations with high fidelity (Jiang & van den Bosch 2014). Using cosmological parameters from Planck Collaboration XIII (2016), we generate merger trees for halos with $z = 0$ masses $M_{\rm vir} = 10^{10-14}\ {\rm M_\odot}$, with 0.1 dex spacing in $z = 0$ mass. Here $M_{\rm vir}$ is the halo mass within the evolving virial overdensity from Bryan & Norman (1998). We use a mass resolution of $m_{\rm res} = 10^7\ {\rm M_\odot}$ for $M_{\rm vir} \leq 10^{13}\ {\rm M_\odot}$ and $m_{\rm res} = 10^8\ {\rm M_\odot}$ for $M_{\rm vir} > 10^{13}\ {\rm M_\odot}$.

2.2.2 *Populating merger trees with GCs*

We express the GC formation rate as

$$\dot{M}_{\rm GCs} = \Gamma_{\rm GCs} \times {\rm SFR}, \tag{1}$$

where $\Gamma_{\rm GCs}$ is the fraction of stars forming in bound clusters that are sufficiently massive to survive until $z = 0$ (corresponding roughly to cluster birth masses $\gtrsim 10^5\ {\rm M_\odot}$; see Muratov & Gnedin 2010, and references therein). Motivated by the results of cloud-collapse simulations, $\Gamma_{\rm GCs}$ depends only on the mean surface density of GMCs. We estimate SFR and $\Sigma_{\rm GMC}$ throughout a merger tree as follows.

We assume an equilibrium model wherein the gas content of a galaxy is set by a balance between cosmological inflow and stellar feedback-driven outflow (e.g. Davé, Finlator & Oppenheimer 2012; Lilly et al. 2013; Rodríguez-Puebla et al. 2016a). In such models, the mass of a galaxy's cold gas reservoir is determined by

$$\dot{M}_{\rm gas,in} = \dot{M}_{\rm gas,out} + {\rm SFR}, \tag{2}$$

where $\dot{M}_{\rm gas,in}$ and $\dot{M}_{\rm gas,out}$ are inflow and outflow rates of cold gas, respectively. If stellar feedback expels gas from the galaxy in proportion to the mass of stars formed with mass loading factor $\eta = \dot{M}_{\rm gas,out}/{\rm SFR}$, the star formation rate can be written as

$${\rm SFR} = \dot{M}_{\rm gas,in}/\left(1+\eta\right). \tag{3}$$

The SFR at a given point in the merger tree thus depends on the cold gas accretion rate, $\dot{M}_{\rm gas,in}$, and the mass-loading factor, η.

At sufficiently high redshift, we expect $\dot{M}_{\rm gas,in} \approx f_{\rm b}\dot{M}_{\rm tot,in}$, where $f_{\rm b} = 0.165$ is the cosmic baryon fraction, and $\dot{M}_{\rm tot,in}$ is the total (dark matter plus baryon) accretion rate (e.g. Dekel et al. 2009). At later times, the fraction of all baryons that are in cold gas drops due to a combination of star formation and heating by the UV background, virial shocks, and stellar and AGN feedback. Following Davé et al. (2012), we approximate this suppression in the cold gas accretion rate as

$$\dot{M}_{\rm gas,in} = f_{\rm b}\dot{M}_{\rm tot,in} \times \zeta, \tag{4}$$

where $\zeta \leq 1$ is a function that represents the mass fraction of accreted material that is in cold gas relative to the cosmic baryon fraction. In practice, ζ is calculated as the product of several terms representing heating due to various sources; it varies with redshift and with the masses of the primary and accreted halo. Details on our adopted form of ζ, which is largely phenomenological, can be found in Appendix A. The basic effect of ζ is that at high z, $\dot{M}_{\rm gas,in} \approx f_{\rm b}\dot{M}_{\rm tot,in}$ over a wide range of masses, but at late times, the fraction of baryons in cold gas is suppressed at all masses, especially outside the range $10.5 \lesssim \log(M_{\rm vir}/\,{\rm M_\odot}) \lesssim 12$.

We calculate $\dot{M}_{\rm tot,in}$ directly from the dark matter merger trees. If a merger occurs in a given timestep (i.e. if a halo has more than one progenitor), we calculate $\dot{M}_{\rm tot,in}$ as the mass of the accreted satellite divided by the merger time-scale, which we describe in Section 2.2.3. We do not trace the accretion of halos below the resolution limit ('smooth' accretion).

The SFR also depends on the mass loading factor, η. This parameter is both observationally and theoretically poorly constrained (see Schroetter et al. 2015 and references therein). A generic prediction of many theoretical studies is that η is higher in lower-mass halos, since less energy is required to eject gas from a shallow potential well than from a deep one. We parametrize η as

$$\eta = \alpha_\eta \left(\frac{M_{\rm vir}}{10^{12}\ {\rm M_\odot}}\right)^{-\beta_\eta}, \tag{5}$$

where α_η and β_η are free parameters. As a fiducial estimate, we choose $\beta_\eta = 1/3$ and $\alpha_\eta = 1.0$, which matches the prediction for momentum-driven winds (Murray, Quataert & Thompson 2005; Davé, Oppenheimer & Finlator 2011) and is typical of the scalings predicted by simulations (see Schroetter et al. 2015, their fig. 10).

The parameters α_η and β_η are not independent, because η determines the SFR, and thus, the integrated stellar mass of a halo at $z = 0$. We treat β_η as the free parameter; for a given value of β_η, we choose α_η such that the total $z = 0$ stellar mass implied by the model matches the $z = 0$ total observed stellar-to-halo mass relation at the high-mass end (see Fig. 14). Increasing β_η increases the slope of the stellar-to-halo mass relation at low masses. Because a change in β_η causes $\eta(M_{\rm vir})$ to 'pivot' around $M_{\rm vir} = 10^{12}\ {\rm M_\odot}$, α_η (as required to produce the correct normalization of the stellar-

[1] G18 did not identify any unique relation between ε and other integrated cloud properties, suggesting that ε depends most directly on gas surface density. This is expected if the star formation efficiency is set by the balance of the force self-gravity, which scales as $F_{\rm gravity} \sim M^2/R^2$, where M and R are the mass and radius of a GMC, and that of stellar feedback, which scales as the stellar mass formed: $F_{\rm feedback} \sim M_{\rm star} \sim M$. For any fixed cloud geometry, the ratio of these quantities scales as $\Sigma_{\rm GMC} \sim M/R^2$.

[2] We use an implementation of the algorithm provided by Yu Lu; it is available at https://github.com/ylu2010/mergertree.

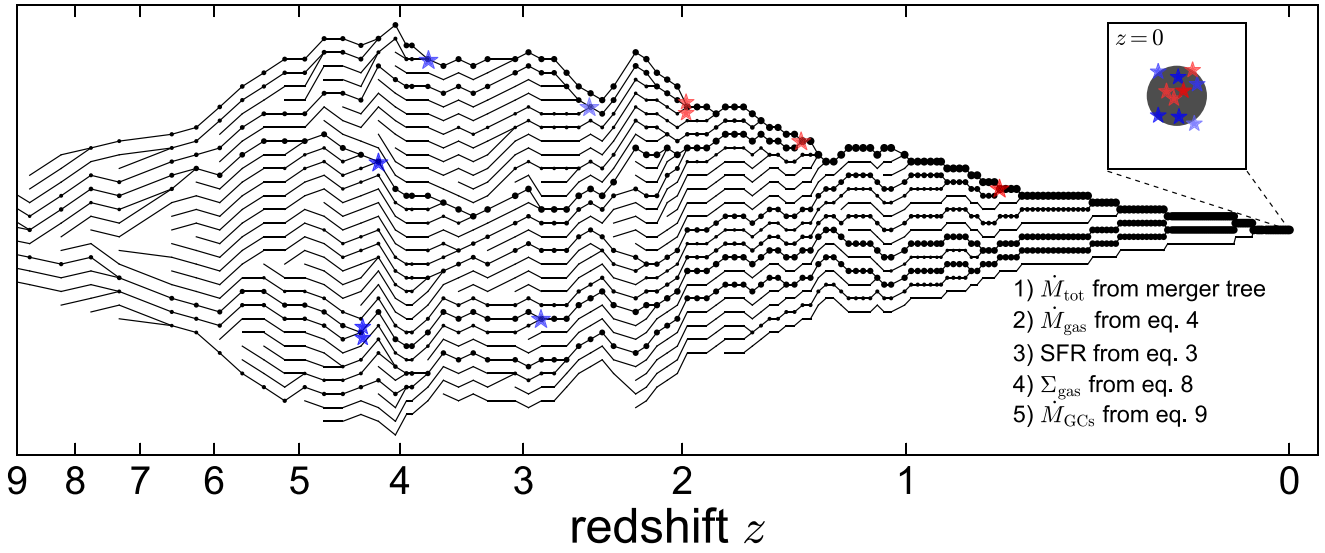

Figure 1. Schematic illustration of our model. Coloured stars represent GC formation events. At each node in the merger tree, we estimate the gas accretion rate from the total accretion rate (equation 4), the SFR from the gas accretion rate (equation 3), and the gas surface density from the SFR (equation 8). Star formation in massive bound clusters occurs when the gas surface density exceeds a critical value (equation 9). These clusters are propagated through the merger tree to $z = 0$, where they represent the observable GC population. We assume GCs form with the same metallicity as the gas in the galaxy in which they form, which is estimated from an analytic mass–metallicity relation (equation 13); as a result, most early-forming GCs have low metallicity (blue), while later-forming GCs have higher metallicity (red).

to-halo mass relation) is only weakly dependent on β_η, varying by ~ 50 per cent over $0 \lesssim \beta_\eta \lesssim 1/2$.

A smaller value of β_η causes a larger fraction of stars and GCs to form in low-mass halos. Given the fixed stellar mass–metallicity relation, we adopt (equation 13), changing β_η also changes the predicted metallicity distributions of GCs and field stars; this is discussed further in Appendix C.

To determine the gas surface density in a given halo, we use a KS-like relation that relates it to the star formation rate surface density. We first calculate the average star formation rate surface density as $\Sigma_{\rm SFR} \approx {\rm SFR}/(\pi R_{\rm d}^2)$, where $R_{\rm d}$ is the scale length of the gas disc. We estimate $R_{\rm d}$ using a model wherein the scale length of the disc is set by the specific angular momentum of the halo (e.g. Fall & Efstathiou 1980; Mo, Mao & White 1998),

$$R_{\rm d} = \frac{\lambda}{\sqrt{2}} R_{\rm vir} \approx 0.025 R_{\rm vir}, \tag{6}$$

where λ is the halo spin parameter that is fixed at a typical value of 0.035 (Bullock et al. 2001). Although the scaling of $R_{\rm d}$ with λ implied by equation (6) likely does not hold in detail (Desmond et al. 2017; Garrison-Kimmel et al. 2017; El-Badry et al. 2018b), the prediction of a constant scaling between disc size and $R_{\rm vir}$ has been found to hold within a factor of $\sim$2 over redshifts $0 < z < 8$ (Shibuya, Ouchi & Harikane 2015) and over nearly eight decades of stellar mass (Kravtsov 2013; Huang et al. 2017).[3]

Given $\Sigma_{\rm SFR}$, we estimate the corresponding gas surface density as described in Faucher-Giguère, Quataert & Hopkins (2013). In their model, gravity is balanced by feedback-driven turbulence such that discs self-regulate to a Toomre parameter $Q \sim 1$. This leads to a KS-like relation of the form

$$\Sigma_{\rm gas}^2 = \frac{(P_\star/m_\star)\,\mathcal{F}}{2\sqrt{2}\pi\, Q G} \Sigma_{\rm SFR}. \tag{7}$$

Here $(P_\star/m_\star)$ is the momentum ultimately injected into the ISM by supernovae per stellar mass formed (Cioffi, McKee & Bertschinger 1988; Ostriker & Shetty 2011), and $\mathcal{F}$ is a factor of order unity encapsulating various uncertainties in the model; Faucher-Giguère et al. (2013) found $\mathcal{F} = 2$ to provide a good match to observations. We set $Q = 1$, $(P_\star/m_\star) = 3000\,{\rm km\,s^{-1}}$, and $\mathcal{F} = 2$, yielding

$$\frac{\Sigma_{\rm gas}}{{\rm M_\odot\,pc^{-2}}} = 1.2 \times 10^3 \left(\frac{\Sigma_{\rm SFR}}{10\ {\rm M_\odot\,kpc^{-2}\,yr^{-1}}} \right)^{1/2}. \tag{8}$$

Here $\Sigma_{\rm gas}$ represents the disc-averaged surface density of cold gas. The surface densities of individual molecular clouds are expected to higher than the disc-averaged surface density. We adopt $\Sigma_{\rm GMC} = 5 \times \Sigma_{\rm gas}$, which is roughly the mean relation found in observations of nearby galaxies and in simulations over a wide range of gas densities (e.g. Bolatto et al. 2008; Hopkins, Quataert & Murray 2012). The factor of 5 is of course uncertain, but we do not leave it as a free parameter because varying it has exactly the same effect as varying $\Sigma_{\rm crit}$.

Given $\Sigma_{\rm gas}$, we can relate the SFR to the GC formation rate, $\dot{M}_{\rm GCs}$. Motivated by the theoretical prediction that the cluster formation efficiency plateaus at $\Sigma_{\rm GMC} \gg \Sigma_{\rm crit}$, we parametrize the GC formation efficiency as

$$\Gamma_{\rm GCs} \equiv \frac{\dot{M}_{\rm GCs}}{\rm SFR} = \frac{\alpha_\Gamma}{1 + (\Sigma_{\rm GMC}/\Sigma_{\rm crit})^{-\beta_\Gamma}}. \tag{9}$$

This parametrization causes the GC formation rate to approach $\alpha_\Gamma \times$ SFR when $\Sigma_{\rm GMCs} \gg \Sigma_{\rm crit}$ and to be suppressed at $\Sigma_{\rm GMCs} \ll \Sigma_{\rm crit}$. How strongly GC formation is suppressed at $\Sigma_{\rm GMC} \ll \Sigma_{\rm crit}$ is set by β_Γ (see Fig. B1). The case of $\beta_\Gamma = 0$ corresponds to *no* dependence on surface density; in this case, the GC formation rate will be a constant multiple of the total star formation rate implied by the model. Following G18, we set fiducial values of

[3] Kravtsov (2013) found the normalization constant 0.025 in equation (6) to be closer to 0.01 at low redshift for stellar discs, but noted that this is expected if the Mo et al. (1998) normalization held at the epoch of disc formation and halos subsequently grew by pseudo-evolution.

$\Sigma_{\rm crit} = 3000\ {\rm M}_\odot\ {\rm pc}^{-2}$ and $\beta_\Gamma = 1$. For particular values of $\Sigma_{\rm crit}$ and β_Γ, we set α_Γ such that the model reproduces the normalization of the observed GC-to-halo mass relation at the high-mass end. This implies $\alpha_\Gamma = 2.1 \times 10^{-3}$ for the fiducial model. In reality, a large fraction of star clusters born in high-density gas discs are disrupted shortly after their formation (e.g. Fall, Chandar & Whitmore 2005; Fall & Chandar 2012). This causes the effective formation efficiency of surviving clusters, represented by α_Γ in our model, to be small.

2.2.3 Merger time-scale

Each time a halo in the merger tree is accreted, our model requires an estimate of the accretion time-scale to determine the implied $\dot{M}_{\rm tot,\,in}$ (equation 4). One possibility is to use the output timestep of the merger trees. In this case, the value of $\Sigma_{\rm gas}$, and thus, the properties of the predicted GC population, will depend on the time resolution of the merger tree. We therefore instead define the merger time-scale to be roughly the dynamical time of the galaxy:

$$\tau_{\rm merger} = \frac{0.05 R_{\rm vir}}{V_{\rm vir}}, \tag{10}$$

where $0.05R_{\rm vir}$ approximates the size of the galaxy and $V = \sqrt{G M_{\rm vir}/R_{\rm vir}}$. Because the relative velocity of galaxies during a merger is of order $V_{\rm vir}$, and the merging galaxies travel a distance of order their size during a merger, this time-scale roughly represents how long the gas density is elevated during a merger. It depends only on redshift, varying from $\sim$10 Myr at $z = 5$ to $\sim$25 Myr at $z = 2$ to $\sim$100 Myr at $z = 0$.

When a halo of mass $M_{\rm vir,\,acc}$ is accreted, the resulting total mass accretion rate is

$$\dot{M}_{\rm tot,\,in} = \frac{M_{\rm vir,\,acc}}{\tau_{\rm merger}}. \tag{11}$$

The total GC mass formed in a GC formation event is then

$$\Delta M_{\rm GCs} = \dot{M}_{\rm GCs} \times \tau_{\rm merger}, \tag{12}$$

where $\dot{M}_{\rm GCs}$ is calculated from equation (1). The adopted $\tau_{\rm merger}$ has no effect on the total stellar mass formed, since the resulting change in the implied mass accretion rate (equation 11) is exactly balanced by the change in the time-scale over which stars and GCs form (equation 12). However, the merger time-scale does affect the total GC mass formed, because a decrease in $\tau_{\rm merger}$ implies an increase in the SFR, which implies an increase in $\Sigma_{\rm gas}$ and a higher fraction of stars formed in GCs. Increasing $\tau_{\rm merger}$ has the same effect as decreasing $\Sigma_{\rm crit}$.

2.2.4 GC masses

We draw the masses of individual clusters for each GC formation event from an m^{-2} power law with $m_{\rm min} = 10^5\ {\rm M}_\odot$, assuming that the majority of lower-mass clusters would be disrupted by $z = 0$ (e.g. Fall & Zhang 2001; Muratov & Gnedin 2010). Following Muratov & Gnedin (2010, their equations 11 and 12), we determine the mass of the most-massive cluster formed in a given event using 'optimal sampling' (Kroupa et al. 2013). Because we do not predict GC mass functions and do not implement mass-dependent GC disruption or evaporation in our fiducial model, these choices have little effect on our primary conclusions. We consider the effects of mass- and age-dependent GC disruption and evaporation in Appendix D; there, the GC mass spectrum does affect how much disruption occurs.

For a given minimum and maximum GC mass, we then compute $\overline{m}$, the mean mass of the GC mass function, and the predicted number of GCs formed, $\Delta N_{\rm GCs} = \Delta M_{\rm GCs}/\overline{m}$. The number of GCs formed in a single event must always be an integer. In order to ensure that our procedure for stochastically drawing GC masses on average forms the correct total $\Delta M_{\rm GCs}$ as predicted by equation (12), we use another random draw to determine the number of GCs formed. For example, if $\Delta N_{\rm GCs}$ is 2.7, we form 3 GCs with 70 per cent probability and 2 GCs with 30 per cent probability. If $\Delta N_{\rm GCs} < 0.5$, no GCs are formed. This limit prevents spurious GC formation in accretion events in which the mass of accreted cold gas is insufficient to form a GC.

2.2.5 Metallicities and colours

We assume that GCs inherit the gas-phase metallicity of the galaxy in which they formed, which we calculate using the mass–metallicity relation from Ma et al. (2016):[4]

$$[{\rm Fe/H}]_{\rm gas} = 0.35\left(\log\frac{M_{\rm star}}{{\rm M}_\odot} - 10\right) + 0.93\exp(-0.43z) - 1.25. \tag{13}$$

When calculating metallicities, we assign a stellar mass to each halo using the median stellar-to-halo mass relation from (Behroozi, Wechsler & Conroy 2013).[5] Following Tremonti et al. (2004), we assume an intrinsic Gaussian scatter in metallicity at fixed $M_{\rm star}$ of $\sigma_{\rm [Fe/H]} = 0.1$ dex. We calculate colours for model GCs using PARSEC isochrones (v1.2S; Bressan et al. 2012; Tang et al. 2014; Chen et al. 2014, 2015). We treat each GC as a simple stellar population with a Kroupa (2001) initial mass function. A model GC's colour at a given time thus depends only on its age and metallicty.

2.2.6 GC disruption

Our fiducial model does not include any GC disruption, tidal stripping, or mass loss due to two-body evaporation, besides the assumption that clusters with birth masses below $10^5\ {\rm M}_\odot$ will be disrupted by $z = 0$. Although disruption likely does have non-negligible effects on some observable properties of the GC population (Spitzer 1987; Gnedin, Lee & Ostriker 1999; Fall & Zhang 2001; McLaughlin & Fall 2008; Carlberg 2017), we do not believe that a model such as ours can capture disruption with much fidelity. The efficiency of disruption is highly dependent on the spatial distribution of GCs: GCs are subjected to strong tidal forces and can be rapidly destroyed as long as they reside in the discs within which they formed (Kruijssen et al. 2012). Tidal effects become much weaker once GCs migrate into the halo due to mergers (e.g. Kruijssen 2015) or feedback-driven fluctuations in the gravitational potential (e.g. El-Badry et al. 2016, 2018a). Because our model does not include information about the spatial distribution of GCs, it cannot account for these effects.

[4] Ma et al. provide a fitting function for the mass-weighted total metallicty, $\log(Z_{\rm gas}/Z_\odot)$. Following their convention, we then estimate $[{\rm Fe/H}]_{\rm gas} = \log(Z_{\rm gas}/Z_\odot) - 0.2$, where $[{\rm Fe/H}]_{\rm gas}$ represents the logarithmic iron abundance relative to the Solar value. The mass and redshift-evolution predicted by this relation are similar to those predicted in the model of Choksi et al. (2018). However, the normalization is lower at all masses and redshifts, by 0.2–0.3 dex on average. We note that Ma et al. assumed $Z_\odot = 0.02$; assuming $Z_\odot = 0.014$ would increase all [Fe/H] values by 0.15 dex.

[5] We do not use the stellar mass calculated directly from our model because it includes the mass of unmerged satellites while the mass–metallicity relation is for individual galaxies (see Section 4.2.1).

We consider the effects of a simple analytical model for GC disruption and stripping, which has also been employed in other recent semi-analytic works, in Appendix D. We find that the primary effect of disruption as implemented in this model is to change the normalization of $\Gamma_{\rm GCs}$; i.e. the α_Γ parameter in equation (9)) required to match the observed GC-to-halo mass relation.

2.3 Random GC formation model

We also construct a pathological random model for GC formation in which GC mass and halo mass are uncorrelated at the time of GC formation. We use this model to explore what aspects of the $z = 0$ GC population are expected purely due to hierarchical assembly.

In the random model, we select a random subset of all halos in the merger tree at $z > 2$ as GC formation sites. Each node in the merger tree at $z > 2$ is assigned the same probability of hosting a GC formation event. We then randomly assign each of these halos a GC mass to form, $\Delta M_{\rm GCs}$, which we draw from a log-uniform distribution over $10^{5-7}\,{\rm M}_\odot$. The absolute probability of hosting a GC formation event is set such that the normalization of the resulting $z = 0$ GC-to-halo mass relation matches the observed relation at the high-mass end. As in the fiducial model, masses of individual GCs for each formation event are sampled as described in Section 2.2.4. Both the halos in which GCs form and the GC mass formed in each halo are chosen without any consideration of halo mass, the mass accretion rate, or the implied SFR or gas density.

In the random model, the mean GC mass formed *per halo* is independent of halo mass. Because low-mass halos are more abundant than high-mass halos, the mean GC mass formed *per unit halo mass* decreases with $M_{\rm vir}$, roughly as $M_{\rm vir}^{-1}$. We experimented with an alternate random model in which the probability of hosting a GC formation event is instead proportional to halo mass, such that the mean GC mass formed *per unit halo mass* is independent of halo mass. All the results we present are unchanged under this alternate model.

3 RESULTS

3.1 GC system mass–halo mass relation

We first consider the $z = 0$ relation between halo mass and the total mass of all GCs in a halo ($M_{\rm GCs}$, which in our default model without GC disruption is simply the sum of the masses of all GCs formed in the merger tree). Fig. 2 shows the effect of varying β_Γ (left) and β_η (right). The black dashed line represents the observed relation, which is well-fit by a constant GC-to-halo mass ratio, $M_{\rm GCs} = 3.5 \times 10^{-5} M_{\rm vir}$, over $M_{\rm vir} \simeq 10^{11-15}\,{\rm M}_\odot$ (Harris 1996; Hudson et al. 2014; Harris et al. 2015). We plot the median total GC mass predicted by the model for four choices of β_Γ (left, while keeping $\beta_\eta = 1/3$ fixed) and four choices of β_η (right, while keeping $\beta_\Gamma = 1$ fixed). The median relation is calculated from 20 Monte Carlo merger trees at each 0.1 dex interval in $M_{\rm vir}$, which we find to be sufficient for all quantities to be converged.

The left-hand panel shows that the shape of the global GC-to-halo mass relation is not sensitive to β_Γ. The overall normalization of the relation *does* vary somewhat with β_Γ, but we always set the parameter α_Γ in equation (9) such that the normalization of the GC-to-halo mass matches the observed value at the high-mass end. With this constraint in place, the GC-to-halo mass ratio is essentially independent of β_Γ over all halo masses and is completely linear at high halo masses. As we discuss in Section 4.2, this owes to the self-similar assembly histories of dark matter halos. At low masses, the fiducial model with $\beta_\eta = 1/3$ produces lower total GC mass on average than predicted by a constant GC-to-halo mass relation. Halos with $M_{\rm vir} \lesssim 3 \times 10^{10}\,{\rm M}_\odot$ on average do not form any GCs.

The right-hand panel of Fig. 2 shows that the shape of the GC-to-halo mass relation does depend somewhat on β_η, which determines how the mass loading factor varies with halo mass. A larger value of β_η leads to larger η, lower SFR, and thus, lower $\Sigma_{\rm GMC}$ at $M_{\rm vir} < 10^{12}\,{\rm M}_\odot$ (equation 5), so higher values of β_η lead to fewer GCs forming in low-mass halos. Interpreted at face value, Fig. 2 would appear to suggest that a value of β_η close to 0 provides the best match the observed constant GC-to-halo mass relation; however, we caution that there is significant scatter in the observed relation at low halo masses and some indication that observed GC systems on average also fall below the constant ratio at $M_{\rm vir} \lesssim 10^{12}\,{\rm M}_\odot$; see Choksi et al. (2018, their fig. 3). On the other hand, varying β_η has little effect on the shape of the GC-to-halo mass relation at higher masses, $M_{\rm vir} \gtrsim 10^{11.5}\,{\rm M}_\odot$. As we will now show, a constant GC-to-halo mass relation at high halo masses is a generic consequence of hierarchical assembly.

3.1.1 GC-to-halo mass relation with random GC formation

Fig. 3 compares the evolution of the GC-to-halo mass relation in the fiducial and random models. Each gray point represents a single halo at $z = 3$ (top), $z = 1$ (middle), and $z = 0$ (bottom). The left-hand panels show the fiducial model, with $\beta_\eta = 1/3$ and $\beta_\Gamma = 1$, for which the median relation was shown in Fig. 2. For the fiducial model, an approximately constant GC-to-halo mass ratio is already in place at high redshift, but the relation is offset to higher $M_{\rm GCs}$ at fixed $M_{\rm vir}$ relative to the $z = 0$ relation. Because the fraction of the total mass that is assembled at early times is larger for GCs than for halos, halos move rightward in the $M_{\rm GCs}$–$M_{\rm vir}$ plane as they evolve. By $z = 0$, the total GC-to-halo mass relation matches the observed constant ratio at high halo masses but falls somewhat below linearity at low halo masses.

In the right-hand panels, we show the relation predicted by the random model. There is no correlation between GC mass and halo mass at the time the GCs form,[6] but a linear relation emerges at later times as both GC system masses and halo masses grow through mergers. At $z = 0$, the random model – which does not attempt to model any of the physics of GC formation, except that GCs form at $z > 2$ – produces a tight, constant relation at $M_{\rm vir} \gtrsim 10^{11.5}\,{\rm M}_\odot$, with scatter comparable to that predicted by the fiducial model. At lower masses, the scatter grows larger, but the median GC mass does not drop off as it does for the fiducial model.

In the random model, mergers alone drive the GC-to-halo mass ratio towards a constant value. This is a manifestation of the central limit theorem. The total GC mass at $z = 0$ is essentially the result of adding together a long list of random numbers (the GC masses formed in each GC formation event). The same is true for the total halo mass at $z = 0$, which is the sum of all progenitor halo masses. At higher halo masses, the number of random numbers is larger, driving down the scatter in their sum. Irrespective of the GC formation model and the initial relation between GC mass and halo mass, a constant GC-to-halo mass ratio is expected at late times as long as GCs form relatively early, such that there are enough

[6]The substructure at high redshift and low $M_{\rm GCs}$ is an artefact of the GC mass sampling procedure for GC formation events in which only a single GC is formed (Section 2.2.4).

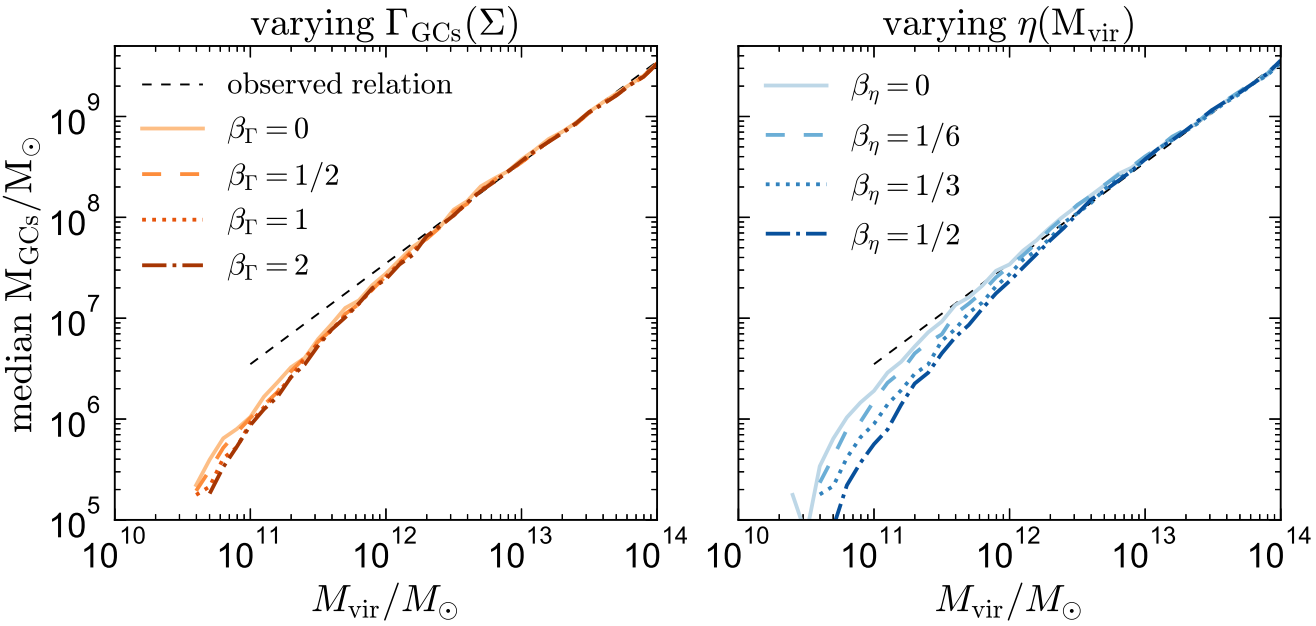

Figure 2. Median total GC-to-halo mass relation predicted by our model for different values of β_Γ (left) and β_η (right). The normalization of the relation is forced to match the observed value at the high-mass end. In the left (right)-hand panel we fix, $\beta_\eta = 1/3$ ($\beta_\Gamma = 1$). The black dashed line represents the observed linear relation, $M_{\rm GCs} \approx 3.5 \times 10^{-5} M_{\rm vir}$. *Left*: The relation is not sensitive to β_Γ, which parametrizes how steeply the cluster formation efficiency falls off at low $\Sigma_{\rm GMC}$ (equation 9). The model reproduces the observed linear behavior at $M_{\rm vir} \gtrsim 10^{12}\,{\rm M}_\odot$, but the prediction falls somewhat below linear at $M_{\rm vir} \lesssim 10^{11.5}\,{\rm M}_\odot$. *Right*: Increasing β_η leads to a decrease in $M_{\rm GCs}$ at low halo masses. At high halo masses, *all* models predict a linear GC-to-halo mass relation.

mergers after the majority of GCs form to drive the ratio towards a constant value.

The scatter in the GC-to-halo mass relation for the random model is large for halo masses at which the maximum $\Delta M_{\rm GCs}$ per formation event ($10^7\,{\rm M}_\odot$ in our implementation) exceeds the value of $M_{\rm GCs}$ implied by the observed relation. This corresponds to $M_{\rm vir} \lesssim 3 \times 10^{11}\,{\rm M}_\odot$ in Fig. 3. At higher masses, the $z = 0$ GC population is always the result of several GC formation events, driving the total GC population towards the mean relation. Increasing the maximum GC mass formed per formation event in the random model increases the scatter in the GC-to-halo mass relation at low masses.

In Fig. 4, we show the stellar mass-normalized GC frequency, $T_N = N_{\rm GCs}/(M_{\rm star}/10^9\,{\rm M}_\odot)$, predicted for the random and fiducial models. We compute the stellar mass of the host galaxy using the stellar-to-halo mass relation from Behroozi et al. (2013), including 0.22 dex scatter. We compare the predictions of both models to observations from the ACS Virgo Cluster Survey (Peng et al. 2008). Both the fiducial and random models match the shape of the observed relation; this is expected if the observed and model galaxies follow a similar stellar-to-halo mass relation. In low-mass galaxies ($M_{\rm star} \ll 10^9\,{\rm M}_\odot$), the random formation scenario predicts significantly higher values of T_N, as the fiducial model suppresses GC formation in low-mass halos. The observed scatter in T_N increases markedly at low masses, as is predicted if the constant GC-to-halo mass relation is primarily a consequence of the central limit theorem in hierarchical assembly.

Observational data are sparse in the mass range where the predictions of the random and fiducial models strongly differ but agree somewhat better with the random formation model. We caution, however, that because T_N depends on the *number* of GCs rather than on their total mass, neglecting GC disruption and evaporation decreases T_N predicted by the model at fixed $M_{\rm GCs}$ (see Appendix D). We discuss the GC populations of low-mass galaxies further in Section 4.2.2.

We have thus far neglected the possible effects of GC evaporation, stripping, and disruption on the GC-to-halo mass relation. In Appendix D, we show that although these processes are expected to change the normalization of the GC-to-halo mass relation at fixed α_Γ, they do not change the conclusion that a constant GC-to-halo mass ratio is expected at high masses purely due to the effects of mergers.

We also emphasize that the random model implemented here is not designed to produce a realistic GC population at $z = 0$, and it can be ruled out by comparing its predictions to observables beside the GC-to-halo mass relation. For example, because randomly selecting nodes in the merger tree as GC formation sites preferentially selects low-mass halos, the random model predicts a large majority of GCs to be metal-poor. Our contention is simply that mergers alone will drive the GC-to-halo mass relation towards a constant ratio for a large range of GC formation models. We return to this discussion in Section 4.2.

3.1.2 *GC-to-halo mass relation for red and blue GCs*

In Fig. 5, we show the $z = 0$ GC-to-halo mass relations predicted by our fiducial model for red and blue GCs separately. We divide red and blue clusters based on their $V - I$ colour (Section 2.2.5), with red GCs having $V - I > 1.0$. This corresponds roughly to the division between the two peaks in the GC colour distributions predicted by our model, and to a metallicity of [Fe/H] $= -1$ for typical GC ages (Section 3.3.2). Blue GCs dominate the population at low halo masses. GC colour is driven primarily by metallicity, and the metal-poor progenitors of low-mass halos form GCs that are metal-poor.

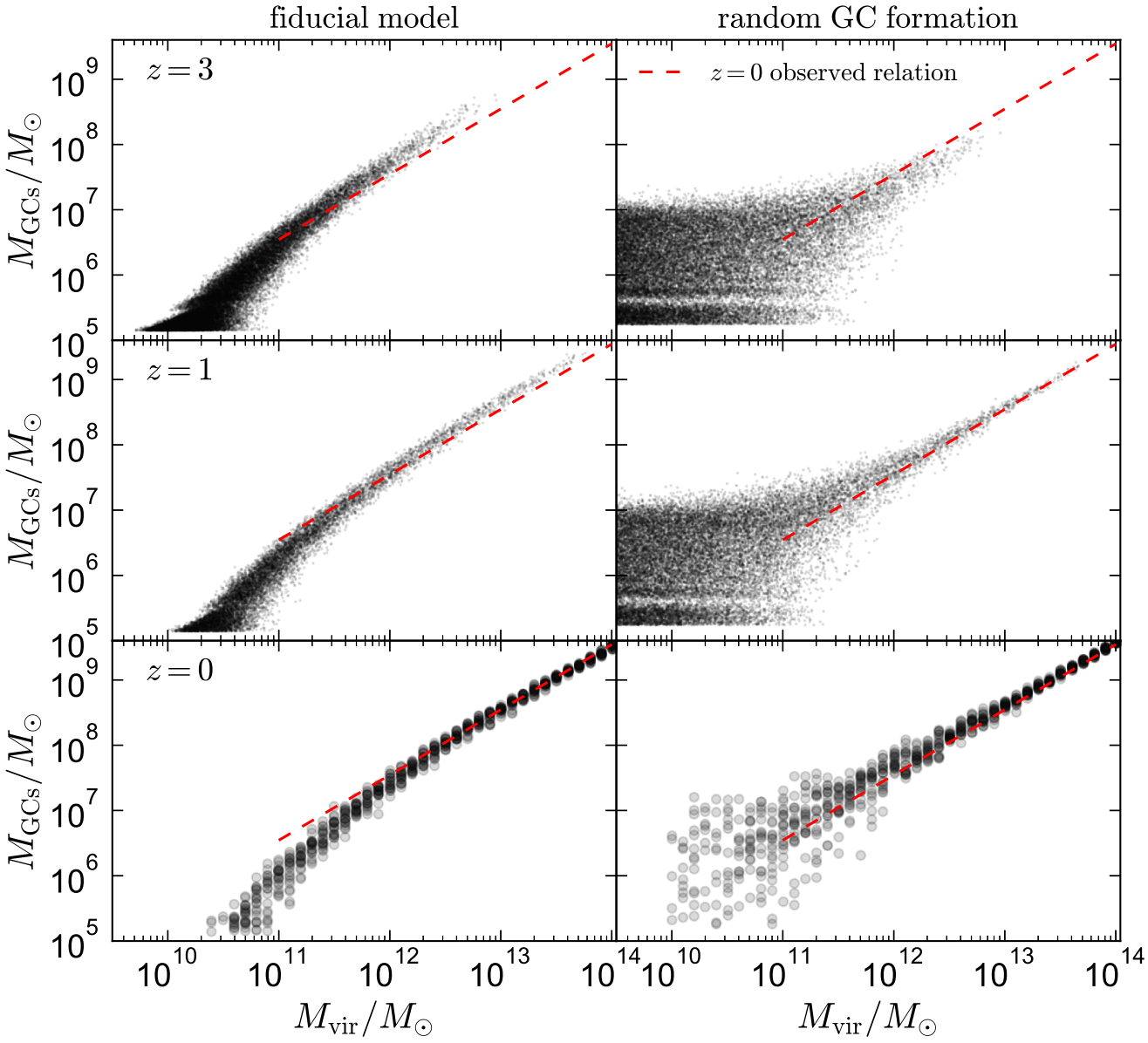

Figure 3. Total GC system mass versus halo mass relation. Each point represents a single halo at $z = 3$ (top), $z = 1$ (middle) and $z = 0$ (bottom); red line represents the observed $z = 0$ relation. *Left*: Fiducial model, in which GC formation occurs at high surface density (see Section 2.2). *Right*: Random model (Section 2.3), in which GCs form in an entirely random subset of all halos at $z > 2$. A tight, linear GC-to-halo mass relation is predicted by $z = 0$ at high halo masses purely as a result of the central limit theorem: high-mass halos form through mergers of low-mass halos, so the total GC-to-halo mass ratio tends to average out by $z = 0$. The linearity of the GC-to-halo mass relation at $M_{\rm vir} \gtrsim 10^{11.5}$ M$_\odot$ thus does not contain much information about GC formation beyond the fact that GCs are old.

$\beta_\Gamma = 1$ is fixed in both panels. The top panel shows predictions for the fiducial mass loading factor scaling of $\beta_\eta = 1/3$. Slightly more red GCs than blue GCs are predicted at the high-mass end, with equal GC mass in the two populations near $M_{\rm vir} = 10^{13}$ M$_\odot$. The bottom panel shows predictions for $\beta_\eta = 0$. In this case, η is lower in low-mass, metal-poor galaxies, making their SFR and $\Sigma_{\rm GMC}$ higher. This results in a higher fraction of blue GCs at all $z = 0$ halo masses. Although it is not shown in Fig. 5, we find that varying β_Γ also changes the relative numbers of red and blue GCs: at fixed $M_{\rm vir}$ and β_η, increasing β_Γ decreases the fraction of GCs that are red because a larger fraction of GCs form at high redshift.

Because the fraction of GCs that are red increases with halo mass, the GC-to-halo mass relation is steeper for red GCs than for blue GCs up to halo masses of a few $\times 10^{13}$ M$_\odot$. This is also found observationally: the best power-law fit to the observed blue GC-to-halo mass relation has a slope of ∼0.96, similar to the constant ratio (slope 1) observed for all GCs, while the slope for red GCs is steeper, at ∼1.21 (Harris et al. 2015). The red fraction predicted by our model flattens at the highest halo masses. Whether such flatting is also found for observed GC populations is unclear due to the small number of observed GC systems in high-mass halos; however, we note that there is substantial scatter in the observed $f_{\rm red}$ at fixed halo mass (e.g. Beasley et al. 2018).

The fraction of GCs that are red indeed decreases in low-mass halos in the local Universe (Brodie & Huchra 1991; Côté et al. 1998; Larsen et al. 2001). However, the sharpness of the transition

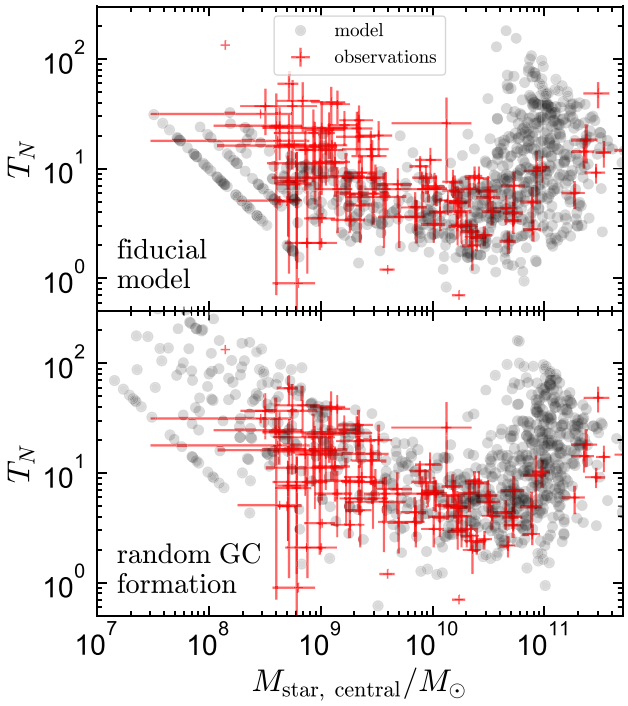

Figure 4. GC frequency, $T_N = N_{\rm GCs}/(M_{\rm star}/10^9\ {\rm M}_\odot)$, versus stellar mass of the host galaxy. Observational data are from the ACS Virgo Cluster Survey (Peng et al. 2008). Model values of $M_{\rm star,\ central}$ are computed assuming the stellar-to-halo mass relation from Behroozi et al. (2013). Predictions for the fiducial model and random GC formation scenario are very similar at $M_{\rm star,central} \gtrsim 10^9\ {\rm M}_\odot$ and are in good agreement with observations. GC formation in the fiducial model is suppressed in lower-mass galaxies, leading to lower T_N than in the random formation scenario.

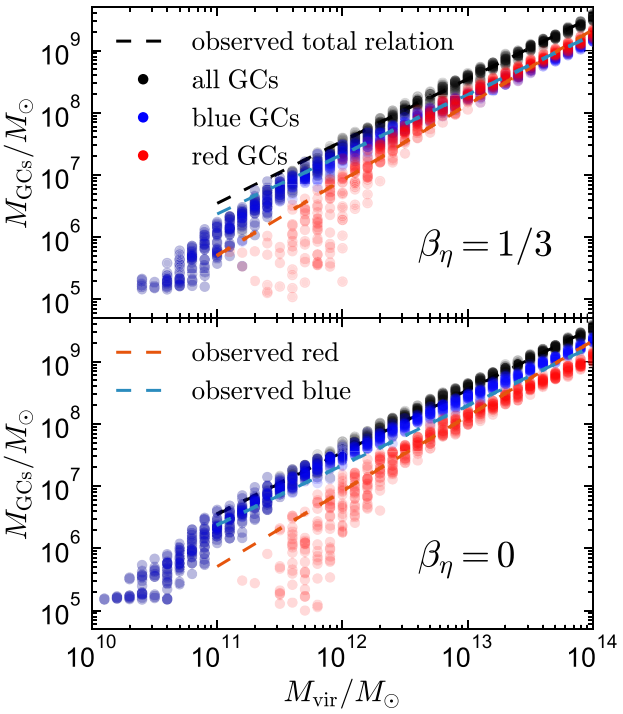

Figure 5. Total GC-to-halo mass relation for red GCs, blue GCs, and all GCs. Dashed line show median observed relations. Top panel shows our fiducial model with $\beta_\eta = 1/3$; bottom panel shows the extreme case in which the galactic wind mass loading factor η does not depend on halo mass. Because blue GCs form in lower-mass halos and at earlier times than red GCs, they fall on a tighter relation at lower masses, where they dominate the GC population. Red GCs make up an increasing fraction of the population at higher halo masses, so the GC-to-halo mass relation for red GCs is superlinear at intermediate halo masses.

from red to blue GCs predicted by our model is steeper than what is observed: some red GCs *are* observed in halos with masses $M_{\rm vir} < M^{11}\ {\rm M}_\odot$, where our model predicts all GCs to be blue. Harris et al. (2015) find a red fraction of ∼30 per cent at $M_{\rm vir} = 10^{12}\ {\rm M}_\odot$ and ∼20 per cent at $M_{\rm vir} = 10^{11}\ {\rm M}_\odot$. The observed red fraction does eventually reach ∼0, but only at $M_{\rm vir} \lesssim 10^{10}\ {\rm M}_\odot$ (Georgiev et al. 2010).

The mean metallicity predicted by our fiducial model agrees well with observations (see Section 3.3), so the dearth of red GCs primarily reflects the fact that the scatter in GC metallicity at fixed mass predicted by our model is lower than is observed. Our model assumes an intrinsic scatter in the galaxy mass–metallicity relation of only 0.1 dex. This value is consistent with what is observed for galaxies in the local Universe (Tremonti et al. 2004), but the relation is uncertain at low masses and high redshifts, where its scatter may also be larger. The observed scatter in the *stellar* mass–metallicity relation at low masses is of order 0.2 dex in the Local Group (Kirby et al. 2013). Our model also assigns the same metallicity to all GCs formed in a given GC formation event, implicitly treating the ISM as homogeneous. The ISM in real galaxies can exhibit spatial abundance fluctuations over a wide range of scales (e.g. Sanders et al. 2012; Krumholz & Ting 2018), and it is also possible that GCs self-enrich during formation (Bailin & Harris 2009). Increased scatter in the mass–metallicity relation due to such effects could also plausibly account for the red GCs observed in low-mass halos.

3.2 Cosmic GC formation rate

To calculate the cosmic mean GC formation rate at a given redshift, we calculate the mean GC formation rate per halo as a function of halo mass and redshift and then weight by the halo mass function:

$$\frac{{\rm d}\dot{M}_{\rm GCs}}{{\rm d}M_{\rm vir}} = \frac{{\rm d}\dot{M}_{\rm GCs}}{{\rm d}n}\frac{{\rm d}n}{{\rm d}M_{\rm vir}} \tag{14}$$

Here ${\rm d}\dot{M}_{\rm GCs}/{\rm d}n = \langle \Delta M_{\rm GCs}/\Delta t \rangle\,(M_{\rm vir}, z)$ is the mean GC formation rate per halo[7] for halos of a particular mass and redshift, and ${\rm d}n/{\rm d}M_{\rm vir}$ is the halo mass function. We use the halo mass function measured from the Bolshoi–Planck and MultiDark–Planck simulations by Rodríguez-Puebla et al. (2016b, their equation 23). Fig. 6 shows the resulting cosmically averaged distribution of GC formation sites at different redshifts. The peak of the distribution is set by the competing effects of a higher average GC formation rate in more massive halos and a larger absolute number of low mass halos; it moves to higher masses at late times as the Schechter mass increases and accretion of cold gas is suppressed in low-mass halos.

If the GC formation rate scaled linearly with halo mass, the distributions in Fig. 6 would be flat below the Schechter mass because each decade in $M_{\rm vir}$ contributes the same total mass for a ${\rm d}n/{\rm d}M_{\rm vir} \sim M_{\rm vir}^{-2}$ halo mass function. The fact that this is *not* the

[7]This quantity is averaged over all halos, including those not forming any GCs. $\Delta M_{\rm GCs}$ represents the GC mass formed in a particular timestep, and Δt, the length of the timestep.

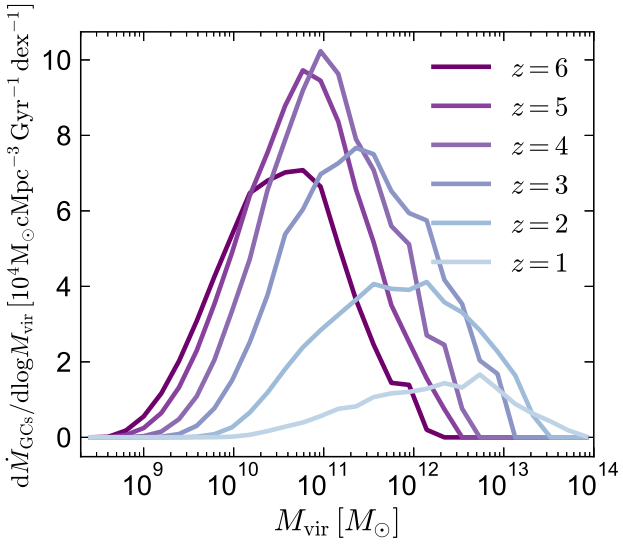

Figure 6. Cosmically averaged distribution of halo masses hosting GC formation at different redshifts, for our fiducial model with $\beta_\eta = 1/3$ and $\beta_\Gamma = 1$. GCs form in progressively more massive halos at later times.

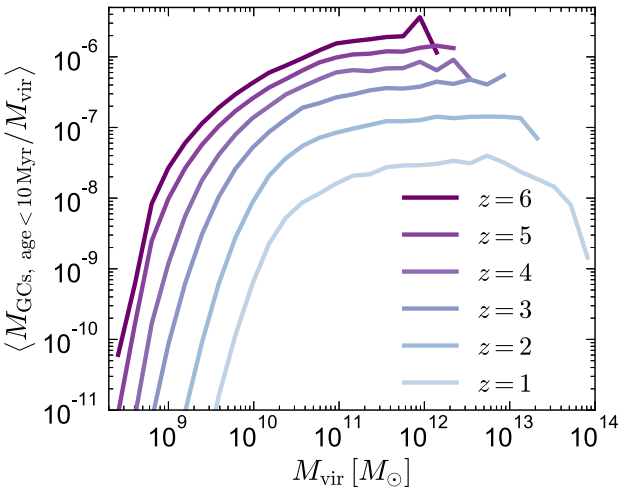

Figure 7. Mean GC-to-halo mass ratio predicted by our fiducial model for young GCs only (age < 10 Myr). GC formation is suppressed at low halo masses, and at high halo masses at late times. The ratio is highest at high z, but the cosmic abundance of massive halos is lower at high z.

case is primarily a consequence of the cold gas-to-dark matter relation adopted in our model (Appendix A and Fig. A1), which imprints a mass scale at which GC formation is most efficient. This can be seen explicitly in Fig. 7, which shows the GC-to-halo mass ratio for young GCs only. This ratio is nearly constant at intermediate halo masses[8] but drops off sharply at low halo masses, and at later times, at high halo masses. Thus, although our fiducial model predicts the integrated GC mass in a halo at a given redshift to scale linearly with halo mass (Fig. 3), the specific GC formation *rate* at any redshift varies with halo mass.

[8]This occurs because the gas accretion rate scales nearly linearly with halo mass (Dekel et al. 2009).

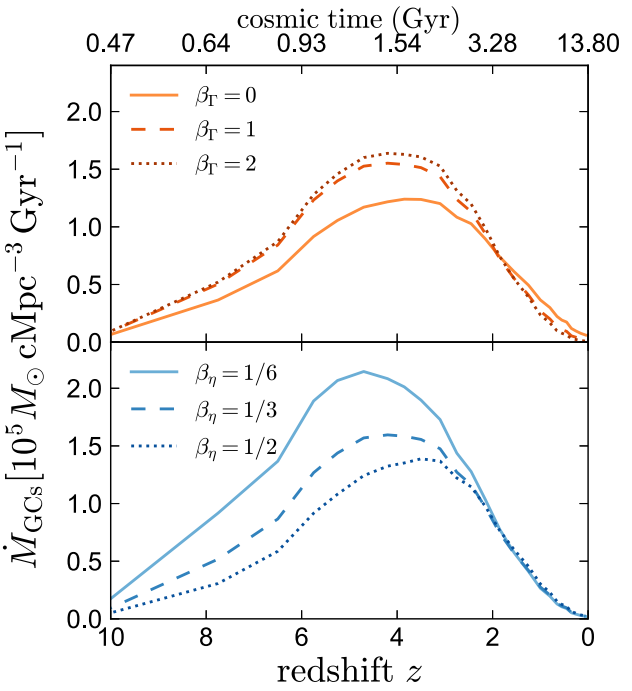

Figure 8. Cosmic GC formation rate predicted by our model; i.e. the result of integrating the distributions in Fig. 6 over all halo masses. Top panel varies β_Γ while holding $\beta_\eta = 1/3$ fixed; bottom panel holds $\beta_\Gamma = 1$ fixed and varies β_η. The GC formation rate peaks at $3 \lesssim z \lesssim 5$. As a result, GCs in these models contribute only a few per cent of the UV luminosity during reionization.

The total cosmic GC formation rate can be computed by integrating over all halo masses:

$$\dot{M}_{\rm GCs}(z) = \int \frac{{\rm d}\dot{M}_{\rm GCs}}{{\rm d}M_{\rm vir}}(M_{\rm vir}, z)\,{\rm d}M_{\rm vir}. \quad (15)$$

The resulting GC formation rate per comoving volume is shown for three values of β_Γ and β_η in Fig. 8. For typical model choices, the GC formation rate peaks at $z \sim 3$–5. This peak is set primarily by the balance between lower $\Sigma_{\rm GMC}$ at low redshift and a dearth of massive halos at high redshift. The peak moves towards higher z for higher β_Γ or lower β_η, both of which cause a larger fraction of GCs to form in low-mass halos at early times.

Figs 6 and 8 imply that although the GC formation rate peaked at $z \sim 3$–5, some GCs should continue to form at late times in gas-rich galaxies with high SFRs over a wide range of halo masses. These GCs can be associated with massive star clusters observed forming in the nearby Universe (e.g. Portegies Zwart et al. 2010). The high GC formation rate predicted at $z \gtrsim 2$ is also in agreement with previous works (e.g. Shapiro, Genzel & Förster Schreiber 2010) that have associated GC formation with the bright star-forming clumps observed in galaxies at $z \sim 2$ (Förster Schreiber et al. 2009, 2011; Adamo et al. 2013), or with compact bright sources seen in lensed fields at higher redshifts (Bouwens et al. 2017; Vanzella et al. 2017).

We note that although the GC formation *rate* predicted by our model peaks at $z \sim 3$–5, the most common GC formation redshift is somewhat younger, typically corresponding to $z_{\rm form} \sim 2.5$ (see Fig. 9). There is more time for GCs to form at low redshifts, so the lower formation rate predicted by our model at late times still contributes significantly to the total GC population. Integrated over all formation redshifts, our fiducial model predicts a mean cosmic

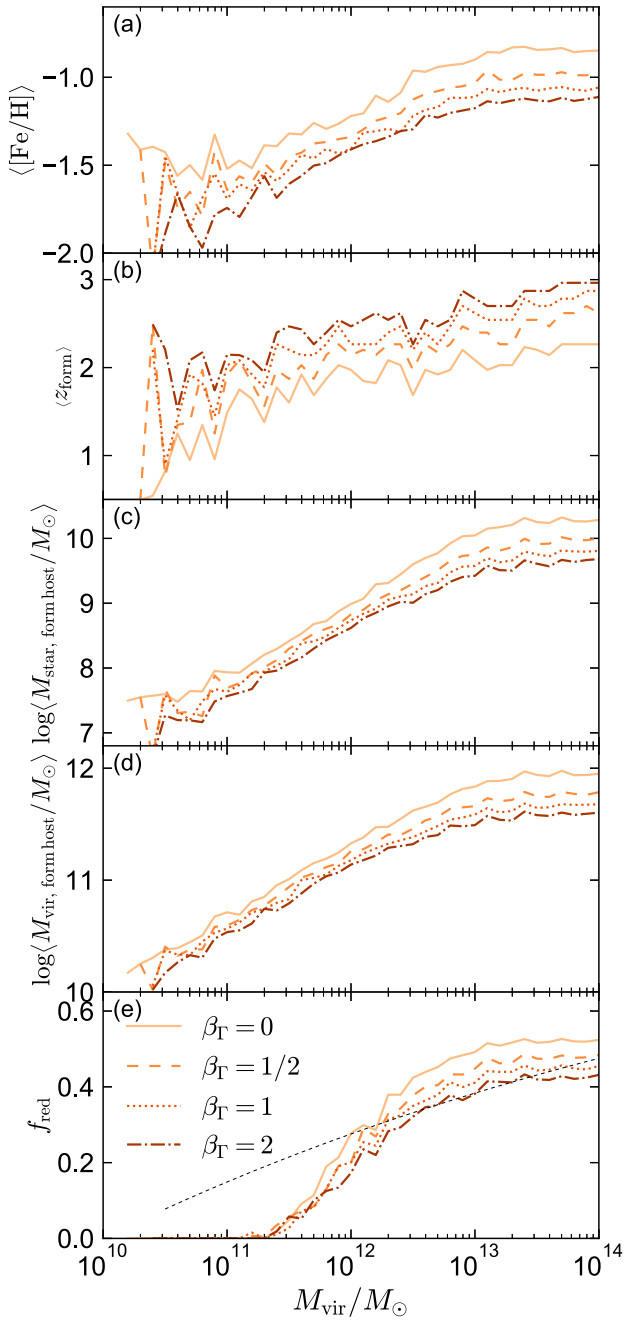

Figure 9. Median GC properties as a function of halo mass at $z = 0$. (a): GC metallicity. (b): GC formation redshift. (c): stellar mass of the host galaxy in which the GC formed, at the time of the GC's formation. (d): virial mass of the host halo in which the GC formed, at the time of the GC's formation. (e): fraction of GCs that are red (V − I > 1.0 mag); dashed black line shows a fit to observations.

GC mass density of $\phi_{\rm GCs} = 5 \times 10^5$ M$_\odot$ Mpc^{-3} at $z = 0$, as is required to match the observed GC-to-halo mass relation.

We calculate the contribution of GCs to the cosmic UV luminosity density using approximations for the time evolution of the UV luminosity at 1500 Angstroms of simple stellar populations from Boylan-Kolchin (2017b, their equations 17–18). At $z = 6$, our fiducial model predicts $\rho_{\rm UV} = 1.95 \times 10^{24}$ erg s^{-1} Hz^{-1} cMpc^{-3}. This is roughly 1 per cent of the total UV luminosity density found by Bouwens et al. (2015) at the same redshift when integrating the luminosity function down to $M_{\rm UV} = -13$. The predicted UV luminosity density due to GCs is less than 2 per cent of the total cosmic value over $4 < z < 9$, so in the fiducial model, GCs form too late to contribute substantially to reionization.

3.3 GC populations

3.3.1 Trends with halo mass

We now examine how properties of the GC population predicted by our model scale with halo mass. Fig. 9 shows the median GC metallicity, formation redshift, and birth galaxy and halo mass, as well as the fraction of GCs that are red. We fix $\beta_\eta = 1/3$ in all panels and show predictions for four different values of β_Γ, corresponding to the cluster formation efficiencies shown in Fig. B1. At fixed halo mass, increasing β_Γ causes the GC population to form in lower-mass halos and become older, bluer, and more metal poor. A higher value of β_Γ limits GC formation to galaxies with higher $\Sigma_{\rm GMC}$, and $\Sigma_{\rm GMC}$ is on average higher at high z. We find that decreasing β_η has qualitatively similar effects to increasing β_Γ.

The median metallicity of GC systems (panel a) and stellar and halo masses of GC formation sites (panels c and d), as well as the fraction of GCs that are red (panel e) all increase monotonically with halo mass at $M_{\rm vir} \lesssim 10^{13}$ M$_\odot$ and then flatten off at high halo masses. The primary reason for this flattening is that our model suppresses cold gas accretion at high halo masses and late times (see Appendix A). Thus, few GCs form in halos with $M_{\rm vir} \gtrsim 10^{13}$ M$_\odot$. The $z = 0$ GC populations of these halos consist primarily of GCs that formed in lower-mass halos that subsequently merged. Thus, halos with $M_{\rm vir} \gtrsim 10^{13}$ M$_\odot$ all have similar GC population demographics, reflecting the average demographics of the lower-mass progenitors in which most of the GCs formed. The uniformity of GC systems predicted at high halo masses is a consequence of the fact that most GCs form relatively early, before high-mass halos assembled. At the highest $z = 0$ halo masses, most GCs formed in halos with $M_{\rm vir} \simeq 10^{11-12}$ M$_\odot$ and $M_{\rm star} \simeq 10^{9-10.5}$ M$_\odot$ at the time of GC formation; this is the mass regime in which cold gas accretion is most efficient. Such halos form earlier on average in overdense regions that collapse into cluster-mass halos by $z = 0$ than in underdense regions; this causes the median GC formation redshift to increase weakly with halo mass (panel b).

Our model predicts no red GCs in halos with $M_{\rm vir} \lesssim 10^{11}$ M$_\odot$ (panel e). The dashed black line shows a log-linear fit to the $f_{\rm red}$ values for nearby galaxies compiled in Harris et al. (2015), highlighting the discrepancy between the observed non-zero occurrence rate of blue GCs in low-mass halos and the predictions of the model. We note that although the observed red fraction is higher than what is predicted by our model at low halo masses, the mean metallicity predicted by our model at the low-mass end is in good agreement with observed values, which find $-2 \lesssim \langle$[Fe/H]$\rangle \lesssim -1.5$ in the lowest-mass galaxies hosting GCs (e.g. Georgiev et al. 2010; de Boer & Fraser 2016; Choksi et al. 2018). Thus, the tension between our model's predictions and observations relates to the large scatter in the colours and metallicities of observed GC systems in low-mass halos.

3.3.2 GC population bimodality

The colour distributions of the GC populations of many nearby galaxies are bimodal, so GCs are often divided into 'red' and 'blue' subpopulations (Ashman & Zepf 1992; Zepf & Ashman 1993; Harris et al. 2006; Peng et al. 2006; Brodie et al. 2012). Colour bi-

modality has been interpreted as indicative of bimodality in GC metallicity (e.g. Brodie et al. 2012) and possibly also GC age (e.g. Woodley et al. 2010; Dotter, Sarajedini & Anderson 2011; Leaman, VandenBerg & Mendel 2013; but see Strader et al. 2005). We now investigate what ranges of model parameters lead to bimodal colour distributions in our model.

To quantify bimodality, we introduce a 'bimodality statistic', $\mathcal{B}$. Given an array of values x_i, we define $x_{\mathrm{upper},i}$ and $x_{\mathrm{lower},i}$ as the upper and lower halves of the sorted array. We then compute

$$\mathcal{B} \equiv \frac{\mathrm{med}\left(x_{\mathrm{upper},i}\right) - \mathrm{med}\left(x_{\mathrm{lower},i}\right)}{\mathrm{std}\left(x_{\mathrm{upper},i}\right) + \mathrm{std}\left(x_{\mathrm{lower},i}\right)}. \quad (16)$$

$\mathcal{B}$ measures the separation of the 'upper' and 'lower' sub-populations relative to their internal dispersion. We find that a clearly bimodal distribution similar to the observed GC colour distributions of many giant ellipticals has $\mathcal{B} \sim 1.8$, a marginally bimodal distribution without clear separation between the two peaks has $\mathcal{B} \sim 1.5$ and a Gaussian has $\mathcal{B} = 1.1$. Because the separation between the upper and lower sub-populations always occurs at the median value of the sample (not at a fixed colour cut), a high value of $\mathcal{B}$ can occur only when the GC population contains a comparable number of red and blue GCs. To make $\mathcal{B}$ values more stable to stochastic fluctuations, we only compute $\mathcal{B}$ for clusters with $0.65 < V{-}I < 1.35$. This includes the vast majority of GCs formed in our model but excludes GCs younger than ∼2 Gyr.

We explore the range of model parameters β_Γ and β_η that produce a bimodal GC colour distribution in Fig. 10. For each point in β_η–β_Γ parameter space, we predict the GC population for 20 merger tree realizations with $M_{\mathrm{vir}} = 10^{13}\,\mathrm{M}_\odot$ (roughly the mass where the observed colour bimodality is most pronounced; e.g. Harris et al. 2017a) and then compute the median $\mathcal{B}$. We assume photometric uncertainties of 0.02 mag. The colour scale in the left-hand panel shows the median $\mathcal{B}$ for 20 merger tree realizations; in the right-hand panels, we show representative colour distributions corresponding to several points in β_η–β_Γ parameters space that are marked in the left-hand panel. Point (c) corresponds to our fiducial model with $\beta_\eta = 1/3$ and $\beta_\Gamma = 1$.

Consistent with expectations from Figs 5 and 9, the bimodality of the population depends on both β_η and β_Γ. A high value of β_η or a low value of β_Γ suppress GC formation at early times in low-mass halos, leading to a unimodal, red GC population. Conversely, low β_η or high β_Γ leads to a unimodal, blue GC population formed in low-mass halos at early times. However, models across a wide swath of β_η–β_Γ parameter space produce roughly equal numbers of red and blue GCs, and the right-hand panels show that the GC populations predicted by these models generally have two distinct peaks.

Fig. 10 thus shows that requiring the $z = 0$ GC population to exhibit colour bimodality does not strongly constrain the GC formation process, at least in the absence of priors imposed from other observables. This is in some sense unsurprising, since a large number of other semi-analytic GC formation models (Ashman & Zepf 1992; Côté et al. 1998; Beasley et al. 2002; Muratov & Gnedin 2010; Tonini 2013; Li & Gnedin 2014; Kruijssen 2015; Choksi et al. 2018; Pfeffer et al. 2018) have predicted bimodal GC colour distributions while employing a wide range of GC formation prescriptions, mass–metallicity relations, and assumptions regarding the origin of the red and blue GCs. However, some models that predict bimodal GC populations can be ruled out on other grounds. For example, models with $\beta_\eta \sim 0$ imply unrealistically metal-poor metallicity distributions for field stars (see Appendix C). Models with $\beta_\Gamma \sim 0$ can be excluded because they produce GC populations with the same age and metallicity distribution as field stars.

Fig. 11 shows how the colour distributions predicted by our model vary with halo mass. The right-hand panel shows colour distributions for our fiducial model parameters (point (c) in Fig. 10) for halos of different masses. At $M_{\mathrm{vir}} = 10^{11}\,\mathrm{M}_\odot$, all GCs are blue. This is simply a consequence of our adopted stellar-to-halo mass and stellar mass–metallicity relations: a halo of mass $M_{\mathrm{vir}} = 10^{11}\,\mathrm{M}_\odot$ hosts a galaxy with $z = 0$ gas-phase metallicity $[\mathrm{Fe/H}]_{\mathrm{gas}} = -0.8$, which is barely metal-rich enough to form a red GC (see Fig. 13). Its lower-mass progenitors at higher redshift had even lower metallicities, so without GC self-enrichment or additional scatter in the mass–metallicity relation, there is no possibility of forming a red GC.

At higher halo masses, red GCs make up an increasing fraction of the total population. A red mode is barely apparent in the GC population predicted for MW-mass halos but is already pronounced at $M_{\mathrm{vir}} = 10^{13}\,\mathrm{M}_\odot$. The GC populations predicted by our model do not change significantly at $M_{\mathrm{vir}} \gg 10^{13}\,\mathrm{M}_\odot$, because the GCs in these systems almost all formed in lower-mass halos (see Section 3.3). The left-hand panel of Fig. 11 shows that this behavior is qualitatively similar across all sets of model parameters β_η and β_Γ: the strength of the colour bimodality increases with mass up to $M_{\mathrm{vir}} = 10^{13}\,\mathrm{M}_\odot$ and then flattens off.

Although the GC colour distributions of most observed massive halos can be well-fit by a sum of two Gaussians (Peng et al. 2006; Harris et al. 2016), the colour distributions of some massive systems appear more complex (Strader et al. 2011; Harris et al. 2017a) and have been interpreted as exhibiting either unimodality or trimodality. Likely due to the simplicity of our model and limited sources of scatter, the colour distributions we predict at high halo masses are fairly uniform and almost all have two peaks.

3.3.3 *Origin of bimodality*

Fig. 12 illustrates the origin of the metallicity bimodality for the GC population of a typical massive elliptical galaxy. The left-hand panel shows when and where the GCs in the halo at $z = 0$ formed. More than half of the GCs formed in the first 2.5 Gyr of cosmic history ($z_{\mathrm{form}} > 2.6$). These early-forming GCs form primarily in lower-mass galaxies with typical stellar masses of $M_{\mathrm{star}} < 10^{10}\,\mathrm{M}_\odot$ at $z \sim 2$. On the other hand, most GCs with $z_{\mathrm{form}} < 2$ formed in galaxies with $M_{\mathrm{star}} > 10^{10}\,\mathrm{M}_\odot$ and have $[\mathrm{Fe/H}] > -1$. Consistent with the GC age estimates of some works (e.g. Woodley et al. 2010; Dotter et al. 2011; VandenBerg et al. 2013), our model predicts the metal-rich GCs to be younger than the metal-poor GCs by ∼2 Gyr on average. The distribution of GC formation times is not bimodal but has a long tail towards late formation times. The distribution of M_{star} of the host galaxy at the time of formation is marginally bimodal. The distribution of GC metallicities is more strongly bimodal, because at fixed M_{star}, later-forming GCs have higher metallicity. This scenario is consistent with the conclusions of Li & Gnedin (2014), who identified the redshift evolution of the galaxy mass–metallicity relation as an important factor in producing bimodal GC metallicity distributions.

Fig. 13 shows how the distribution of GC ages and metallicities predicted by our model translates to a colour distribution at $z = 0$. GC colour is a stronger function of metallicity than of age for GCs older than ∼3 Gyr. Most red GCs ($V - I > 1$) have $[\mathrm{Fe/H}] \gtrsim -1$. For young GCs, colour is more strongly dependent on age than on

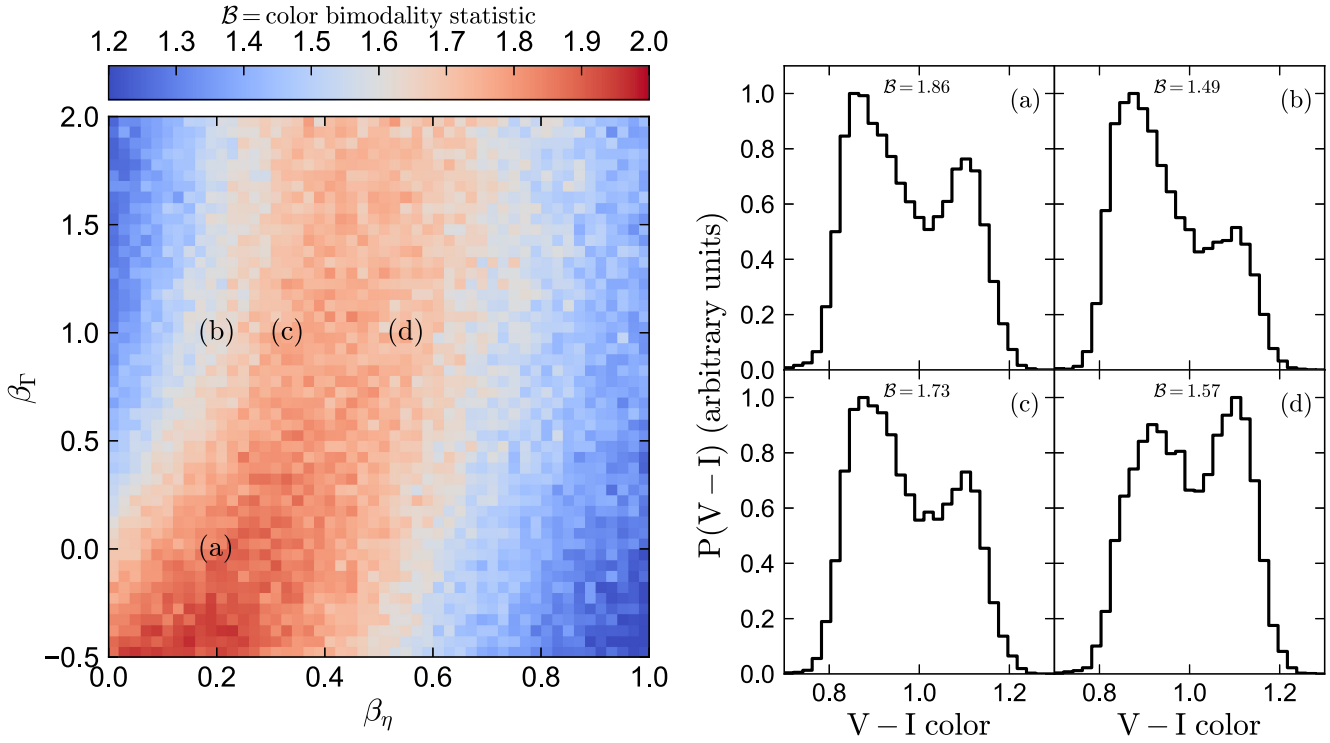

Figure 10. *Left*: GC $V-I$ colour bimodality statistic (equation 16) for halos with $M_{\rm vir} = 10^{13}\,{\rm M}_\odot$. We show bimodality values predicted for a wide range of model parameters. β_Γ parametrizes how the cluster formation efficiency varies with $\Sigma_{\rm GMC}$ (equation 9). β_η parametrizes how the mass-loading factor, which relates the SFR and gas accretion rate, varies with $M_{\rm vir}$ (equation 5). *Right*: Example GC colour distributions for different combinations of β_η and β_Γ. Our model does not produce a bimodal GC colour distribution for $\beta_\eta \gtrsim 2/3$, because then too few GCs are produced in low-mass, metal-poor galaxies; however, a wide range of GC formation efficiencies (i.e. different choices of β_Γ) predicts bimodal GC colour distributions at $z = 0$.

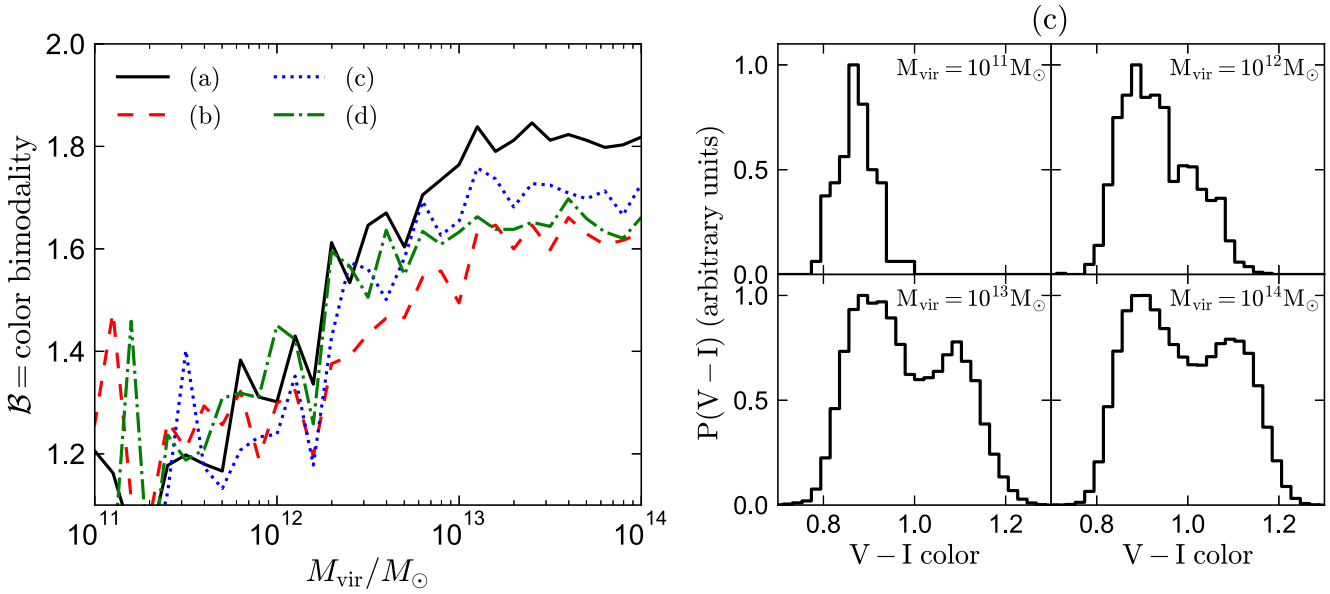

Figure 11. *Left*: GC colour bimodality (equation 16) versus host halo mass. Lines show the median of an ensemble of different merger tree realizations and correspond to the same choice of model parameters as in Fig. 10. Strong colour bimodality implies $\mathcal{B} \gtrsim 1.6$. *Right*: GC colour distribution for a range of halo masses, all for our fiducial parameters of $\beta_\eta = 1/3$ and $\beta_\Gamma = 1$; i.e. line (c).

metallicity, but young GCs constitute a negligible fraction of the total GC population in most cases.

For our fiducial model parameters, massive ellipticals are predicted to have bimodal distributions of both colour and metallicity. However, this is not generically true: for some choices of β_Γ and β_η, the model predicts single-peaked [Fe/H] distributions while still predicting double-peaked colour distributions; see Appendix E. This can occur because GCs with a range of colours and ages can fall on a line of constant colour, such that a unimodal [Fe/H] distribution transforms a bimodal colour distribution. Because the GC popula-

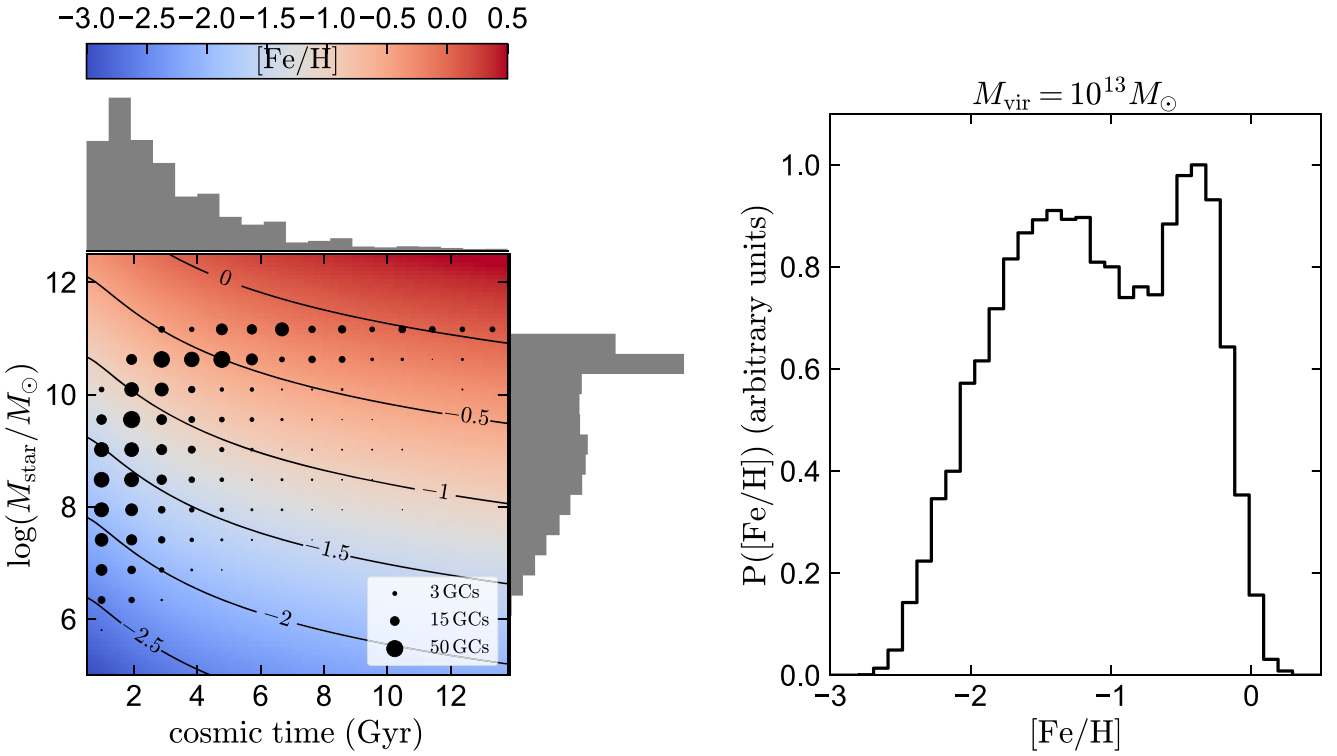

Figure 12. *Left*: Colour scale shows the gas-phase mass–metallicity relation from Ma et al. (2016), which is built into our model. Black circles show the total number of GCs formed in each grid cell; i.e. in galaxies within each $M_{\rm star}$ interval at a given time interval, for a merger tree with $M_{\rm vir} = 10^{13}$ M$_\odot$ at $z = 0$. The area of each circle scales with the number of GCs formed. GCs form with the metallicity of the galaxy in which they form. *Right*: Metallicity distribution for all GCs formed by $z = 0$. Metallicity bimodality is driven primarily by bimodality in $M_{\rm star}$ of the galaxy from which the GC formed.

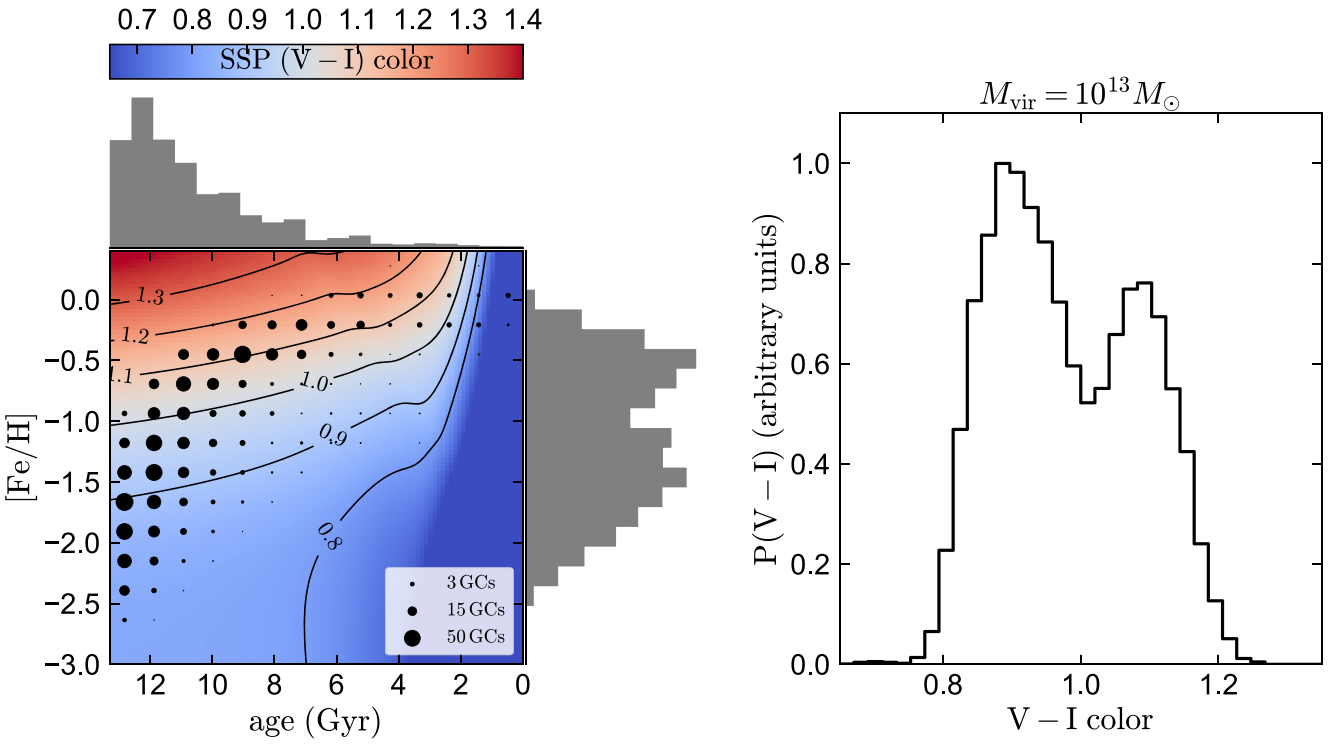

Figure 13. *Left*: IMF-integrated colour-metallicity-age relation from Padova isochrones adopted by our model. Black circles show the total number of GCs formed in each age-metallicity grid cell for a merger tree with $M_{\rm vir} = 10^{13}$ M$_\odot$ at $z = 0$. The area of each circle scales with the number of GCs formed. *Right*: Distribution of GC colours at $z = 0$. The colour bimodality is driven primarily by the metallicitity bimodality.

tions of most giant ellipticals do not have spectroscopic metallicity measurements, some previous works (e.g. Richtler 2006; Yoon, Yi & Lee 2006) have proposed that effects similar to this are responsible for the observed bimodal colour distributions. In the few cases where spectroscopic metallicity measurements are available (e.g. Brodie et al. 2012), the [Fe/H] distributions do also appear to be bimodal.

4 SUMMARY AND DISCUSSION

4.1 Summary

We have used a semi-analytic model for GC formation to explore the sensitivity of the observable properties and scaling relations of low-redshift GC populations to details of the GC formation process. Our model uses dark matter merger trees to predict the GC populations of halos at $z = 0$, treating GC formation as an extension of normal star formation that occurs at high surface densities. Our primary results are as follows.

(i) *GC system mass–halo mass relation*: At $z = 0$, all the models we consider produce a constant GC-to-halo mass relation at high halo masses, independent of the details of the GC formation model (Fig. 2). In fact, a tight GC-to-halo mass relation at $M_{\rm vir} \gtrsim 10^{11.5}\,{\rm M}_{\odot}$ is predicted even when we adopt a pathological random model for GC formation in which the GC formation probability is not tied to any properties of the host halo (Fig. 3). This remains true when we add an approximate treatment of GC disruption and mass loss to the model (Appendix D and Fig. D2). The GC specific frequency predicted by both the fiducial and random models is U-shaped, reflecting the non-linearity in the stellar-to-halo mass relation (Fig. 4).

A constant GC-to-halo mass ratio is predicted for a wide range of models as a result of the central limit theorem. Large halos are formed through mergers of smaller halos, and both the halo masses and GC system masses are summed during mergers. After many mergers, the ratio of total GC mass to halo mass tends to average out, irrespective of the GC-to-halo mass relation when GCs formed. This holds true as long as GCs form relatively early ($z \gtrsim 2$), such that enough mergers occur after the bulk of the GC population forms to drive the population towards the mean relation. GC age constraints from stellar models suggest that most GCs are indeed ancient. We therefore conclude that the observed constant GC-to-halo mass relation does not necessarily imply any fundamental GC-dark matter connection.

At low halo masses ($M_{\rm vir} \lesssim 10^{11.5}\,{\rm M}_{\odot}$ in our fiducial model), mergers alone are insufficient to produce a tight, linear GC-to-halo mass relation. In this regime, the GC-to-halo mass relation predicted by our fiducial model falls below linear, and small number statistics drive up the scatter in the $z = 0$ GC-to-halo mass relation in the absence of a correlation between GC and halo mass at the time of GC formation (Fig. 3). The GC populations of low-mass halos thus retain the most information about the physical conditions under which GCs formed.

(ii) *Cosmic GC formation rate*: Our model predicts the cosmically averaged GC formation rate to peak at $z \simeq 3$–5 (Fig. 8). Our fiducial model predicts that GCs contributed 1–2 per cent of the total UV luminosity density during reionization. Most GCs form in halos with $M_{\rm vir} \simeq 10^{10-12}\,{\rm M}_{\odot}$, with the typical halo mass hosting GC formation increasing over cosmic time (Fig. 6). Although the integrated GC-to-halo mass relation predicted by the model is constant at $z = 0$, the GC formation rate at a particular redshift falls off at low and high halo masses (Fig. 7), largely due to our input model for the gas accretion rate (Appendix A). Because there is more time for GCs to form at lower redshifts, the median GC formation redshift is $z \sim 2.5$ (Fig. 9).

(iii) *GC colour/metallicity bimodality*: Our model predicts the metallicity (Fig. 12) and colour (Fig. 13) distributions of GCs in massive galaxies to be bimodal at $z = 0$ down to MW-mass halos (Fig. 11). The fraction of GCs predicted to be red increases with halo mass, with all GCs in halos with $M_{\rm vir} \lesssim 10^{11}\,{\rm M}_{\odot}$ predicted to be blue (Fig. 5). Bimodal GC colour distributions are predicted for a wide range of model parameters (Fig. 10). Red, metal-rich GCs are on average younger by $\sim$2 Gyr than blue, metal-poor GCs. The metallicity bimodality predicted by our model arises primarily due to bimodality in the masses of the galaxies in which GCs form and is strengthened by the redshift evolution of the mass–metallicity relation (Fig. 12). The median formation redshifts of red and blue GCs are $z_{\rm form} \sim 1.9$ and $z_{\rm form} \sim 3.9$, respectively.

4.2 Discussion: The GC–dark matter connection

Many previous works have proposed causal models for the origin of the constant observed GC-to-halo mass ratio. Peebles & Dicke (1968) first suggested that GCs formed immediately following recombination with a characteristic scale set by the cosmological Jeans mass at $z \sim 1000$. Peebles (1984) revised this model in the context of the CDM paradigm, suggesting that GCs formed in the centers of dark matter minihalos at $z \sim 50$–100 (see also Fall & Rees 1985; Rosenblatt, Faber & Blumenthal 1988). Considerations of the inefficiency of cooling in primordial gas have pushed the preferred epoch of GC formation in similar, more recent models to $z \sim 10$–12, still in dark matter halos at the highest density peaks (Mashchenko & Sills 2005; Moore et al. 2006; Bekki et al. 2008; Boley et al. 2009; Spitler & Forbes 2009; Corbett Moran et al. 2014). These and other works (e.g. Santos 2003; Bekki 2005) have suggested that GC formation was truncated by reionization at $z \gtrsim 6$, at least for metal-poor GCs.

Such models are appealing because they explain the uniformity of GCs found in very different environments and because if reionization truncated GC formation at roughly the same time throughout the Universe, they predict the $z = 0$ GC system mass to scale with halo mass. However, absolute GC age constraints from stellar models have systematic uncertainties of $\pm$1–2 Gyr (e.g. Chaboyer et al. 2017) and thus cannot distinguish between scenarios in which GCs form at $z \gtrsim 2$ and those in which they form prior to reionization. We also note that the apparent lack of dark matter halos around observed GCs (Moore 1996; Baumgardt et al. 2009; Conroy, Loeb & Spergel 2011; Ibata et al. 2013) poses a challenge for dark matter minihalo GC formation models.

Irrespective of whether GCs formed in individual dark matter halos or in galactic discs, a number of recent works (e.g. Harris et al. 2013; Hudson et al. 2014; Harris et al. 2015, 2017b) have argued that (a) a constant GC-to-halo mass ratio at the time of GC formation implies that GC formation was largely unaffected by feedback from UV radiation, stellar winds, supernovae, and AGN, and (b) the constant GC-to-halo mass ratio observed at $z = 0$ implies a constant GC-to-halo mass ratio at the time of formation.

An alternative interpretation of the GC-to-halo mass relation was proposed by Kruijssen (2015). In this model, the total GC mass at $z = 0$ is determined primarily by the fraction of GCs that survive a 'rapid destruction' phase in the discs of high-redshift galaxies. This fraction depends on the stellar mass of the host galaxy at the time of GC formation. Largely by coincidence, the mass-scaling of the

stellar-to-halo mass relation and the surviving GC-to-stellar mass relation nearly cancel in this model, such that the GC-to-halo mass ratio after the rapid destruction phase is nearly constant. Kruijssen (2015) then argues that *once a constant GC-to-halo mass relation is established*, it is likely to be preserved and/or strengthened by hierarchical mergers, as was shown explicitly by Boylan-Kolchin (2017b).

In contrast to the previous work, we find that the existence of a constant GC-to-halo mass ratio at $z = 0$ does not imply the existence of such a relation at high redshift: it is predicted by all the models we consider, including the pathological case in which GC formation occurs at random (Fig. 3). If GCs are relatively old and the $z = 0$ GC population is viewed as the composite population of GCs formed in progenitor halos and assembled through mergers, no coupling between GCs and dark matter halos is needed to explain the observed relation, at least at high halo masses.[9]

The fact that a constant GC-to-halo mass relation is expected due to mergers alone is perhaps most obvious when one considers the GC populations of galaxy clusters. Because galaxy clusters have long dynamical friction time-scales, their GC populations are – unlike those of MW-mass halos – often dominated by GCs bound to satellites, not to the central galaxy. When only the GCs associated with the central galaxy are accounted for, massive clusters are found to have lower GC system masses than predicted for a constant GC-to-halo mass ratio (Spitler & Forbes 2009). On the other hand, massive clusters fall on the observed constant ratio when $M_{\rm GCs}$ also includes both GCs associated directly with the individual member galaxies and intracluster GCs that are not bound to any individual galaxy (see e.g. Spitler & Forbes 2009; Peng et al. 2011; Durrell et al. 2014; Harris et al. 2015). Given that the individual member galaxies in clusters are known to fall on a constant GC-to-halo mass relation and clusters are composed of individual member galaxies (some already tidally destroyed), it follows that the GC population of a whole cluster will have the same GC-to-halo mass ratio as the constituent galaxies.

The mechanism that enforces a constant GC-to-halo mass ratio in our model does not apply uniquely to GCs: it is expected to create a constant ratio at late times between halo mass and any property that is set at relatively early times and is passed on through mergers. In fact, a non-causal, merger-driven scenario is widely recognized as a plausible explanation for the observed constant black hole-to-bulge mass ratio (Peng 2007; Hirschmann et al. 2010; Jahnke & Macciò 2011): because mergers are expected to cause both the bulges and central black holes of merging galaxies to combine, they drive galaxies towards a constant bulge-to-black hole mass ratio. Provided that GCs are not preferentially destroyed during mergers, this scenario is probably *more* applicable for GCs than for black holes: while most GCs are unambiguously old, massive black holes grow both by mergers and by accretion of gas at late times (e.g. Kulier et al. 2015).

Indeed, it has been shown observationally that the ratio between total GC mass and black hole mass is also constant, independent of galaxy or halo mass (Burkert & Tremaine 2010; Harris, Poole & Harris 2014). This fact is not generally interpreted as indicative of a causal connection between GCs and black holes, as both GC and black hole mass are known independently to scale with halo mass (e.g. Gnedin, Ostriker & Tremaine 2014; Kruijssen 2015). Unsurprisingly, such a correlation is also naturally predicted purely due to hierarchical assembly, as noted by Jahnke & Macciò (2011).

[9]Our results of course do not *rule out* the possibility that GC formation is directly linked to properties of dark matter halos. They do imply, however, that the $z = 0$ relation cannot strongly distinguish between different GC formation scenarios.

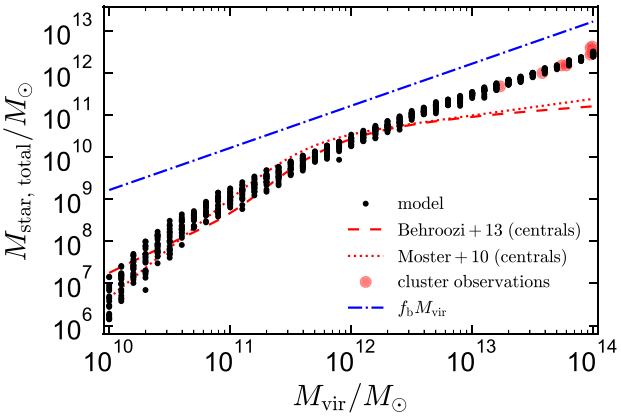

Figure 14. Black points show total stellar mass–halo mass relation implied by our model at $z = 0$; here $M_{\rm star,\,total}$ represents the sum of all stellar mass within the halo (including satellites and the ICL, which dominate at high halo mass). Red curves show parametrizations of the median stellar mass–halo mass relation for distinct halos (excluding satellites); red points show observations of clusters at $z \sim 0$ (including satellites and the ICL). The relation implied by our simplified model agrees with observational constraints to within a factor of a few. At high halo masses, an almost linear *total* stellar-to-halo mass relation is predicted for the same reasons a linear GC-to-halo mass is predicted: most of the stars in clusters are accumulated via mergers.

Finally, we note that some other properties of observed GC scaling relations hint that they arise in large part due to hierarchical assembly. The observed GC-to-halo mass relation is tighter for blue GCs than for red GCs (Peng et al. 2006; Harris et al. 2015). Since blue GCs likely formed in lower-mass halos and at earlier times than red GCs, they are expected to have gone through more mergers by $z = 0$ than red GCs, providing more opportunity to linearize their GC-to-halo mass relation. Perhaps relatedly, the fraction of GCs that are red is at fixed halo mass higher for late-type galaxies than for early-type galaxies (Harris et al. 2015). Since early-type galaxies on average have gone through more mergers, one might expect their GC populations to more closely reflect those of the lower-mass galaxies from which they formed.

Our fiducial model ignores the effects of GC disruption and mass loss. We show in Appendix D that applying an analytic recipe for GC disruption and mass loss does not substantially change the prediction of a constant GC-to-halo mass relation at high halo masses. *If* the disruption efficiency varied strongly with halo mass and disruption occurred primarily at late times, after most mergers, then one might expect disruption to change the shape of the $z = 0$ GC-to-halo mass relation. However, we think such a scenario unlikely because the strongest GC disruption is expected to occur in the tidal fields of the gas disc from which GCs form, *before* mergers liberate GCs from the galaxies in which they formed and deposit them in the halo (e.g. Kruijssen et al. 2012; Kruijssen 2015).

4.2.1 Other scaling relations resulting from hierarchical assembly

Although they are not always interpreted as arising due to mergers, other known scaling relations with halo mass may have a similar origin to the GC-to-halo mass relation. The total number of surviving ancient stars in a halo is predicted to scale almost linearly with

halo mass (see Griffen et al. 2018; their fig. 10). The same is true for the total stellar mass in groups and clusters (see Yang et al. 2007 and their fig. 5). To further illustrate this point, we show in Fig. 14 the total $z = 0$ stellar mass implied by our model as a function of halo mass; i.e. the result of integrating the SFR from equation (3) over all nodes in the merger tree. Black points correspond to individual merger tree realizations, and the two red lines show stellar-to-halo mass relations *for individual galaxies* calculated from abundance matching.

At the high-mass end, the total stellar mass predicted by our model greatly exceeds the stellar mass predicted by the Moster et al. (2010) and Behroozi et al. (2013) relations for central galaxies. This is because the total stellar mass represents not only the stellar mass of the main galaxy, but also the stellar mass of satellite galaxies that have not yet merged with the central galaxy and the mass of stars contributing to the intracluster light (ICL). Because star formation is inefficient in high-mass galaxies and the dynamical friction time-scale is long in cluster-mass halos, these components are in fact the dominant contributors to the total stellar mass in massive galaxy clusters, exceeding the mass of the central galaxy by factors of 5–10 (Lin, Mohr & Stanford 2004; Yang et al. 2007; Leauthaud et al. 2012; Kravtsov, Vikhlinin & Meshcheryakov 2018). To make a fair comparison with our model, we also plot as red hexagons the total stellar mass within a number of intermediate-mass galaxy clusters at low redshift;[10] these are in good agreement with the predictions of our model.

At the high-mass end, the total stellar-to-halo mass relation in clusters is log-linear for the same reason we predict the GC-to-halo mass relation to be linear. Because a larger fraction of stars than GCs form at late times and star formation is suppressed within massive halos at late times (see Appendix A), the logarithmic slope of the total stellar-to-halo mass relation predicted at high masses is somewhat less than one: we find $M_{\rm star,\,tot} \sim M_{\rm vir}^{0.9}$. Fitting the data from observed clusters, we find a very similar value. At still-higher masses, $M_{\rm vir} = 10^{14-15}\,{\rm M}_\odot$, the best-fitting exponent is ∼0.7 (Vale & Ostriker 2006; Becker 2015; Kravtsov et al. 2018).

4.2.2 *The GC-to-halo mass relation in low-mass halos*

A tight, constant GC-to-halo mass ratio cannot be explained purely as a consequence of hierarchical assembly at low halo masses, largely because low-mass halos experience fewer mergers (e.g. Fitts et al. 2018). The precise scale below which mergers fail to enforce a constant GC-to-halo mass ratio depends on the typical total mass formed per GC formation event, the typical GC formation redshift, and the halo mass limit below which GCs do not form; for both our random and fiducial models, it is $M_{\rm vir} \lesssim 10^{11.5}\,{\rm M}_\odot$. Below this mass scale, the GC-to-halo mass ratio drops systematically in the fiducial model, and the scatter in the random model increases.

In the fiducial model, most halos with masses below $M_{\rm vir}(z = 0) \sim 4 \times 10^{10}\,{\rm M}_\odot$ do not form any GCs. The gas accretion rates on the progenitors of halos below this mass are sufficiently low, and η is sufficiently high, that the implied $\Delta M_{\rm GCs}$ is less than the minimum GC mass we adopt for halos to survive to $z = 0$. Some galaxies in halos below this mass *are* observed to host GCs (de Boer & Fraser 2016; Georgiev et al. 2010), and there are some indications that the observed GC-to-halo mass ratio remains constant down to a few

[10]We compile observations from Leauthaud et al. (2012), Gonzalez et al. (2013), and Kravtsov et al. (2018). The points from Leauthaud et al. (2012) are median values computed for several objects in two mass bins.

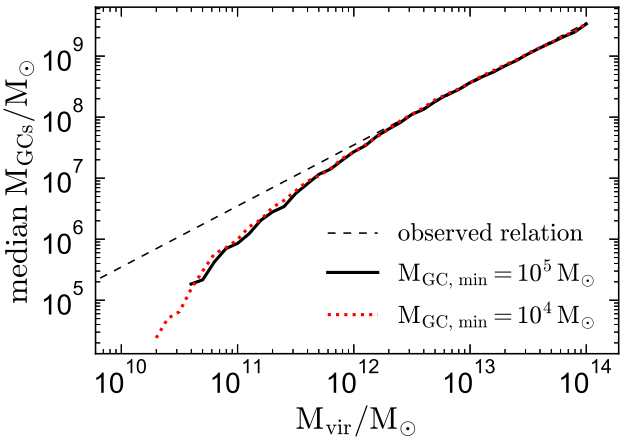

Figure 15. Median predicted GC-to-halo mass relation for two choices of the minimum GC mass. Our fiducial model assumes $M_{\rm GC,\,min} = 10^5\,{\rm M}_\odot$. Decreasing this value to $10^4\,{\rm M}_\odot$ has minimal effects on the GC-to-halo mass relation predicted at high halo masses, but it allows more GCs to form in lower-mass halos, where the fiducial model predicts that most halos will not host any GCs.

$\times\,10^{10}\,{\rm M}_\odot$ (Hudson et al. 2014; Zaritsky, Crnojević & Sand 2016; Harris et al. 2017b). This may indicate that GC formation is more efficient at early times than is predicted by our model, or that there is a causal origin of the GC-to-halo relation at this mass scale.

However, substantial uncertainties remain in the observed GC-to-halo mass relation at low masses. At least some dwarf galaxies may deviate strongly from a linear relation (e.g. Amorisco et al. 2018; Lim et al. 2018; van Dokkum et al. 2018), and the GC-to-halo mass ratio at $M_{\rm vir} \lesssim 10^{11}\,{\rm M}_\odot$ appears to be systematically higher than average for dwarf ellipticals and lower than average for dwarf spheroidals (Spitler & Forbes 2009; Georgiev et al. 2010). Measurements of the GC-to-halo mass ratio at low halo masses are also complicated by the fact that most of the lowest-mass observed GC systems are hosted by satellite galaxies, whose halos may have undergone significant tidal stripping.

Recently, Forbes et al. (2018) found that the observed mean GC-to-halo mass ratio remains constant down to at least $M_{\rm vir} = 10^9\,{\rm M}_\odot$, though the scatter increases substantially at low halo masses. This result is inconsistent with our fiducial model, which predicts a decrease in the GC-to-halo mass ratio below $M_{\rm vir} = 10^{11.5}\,{\rm M}_\odot$. The increased scatter found at low masses is consistent with what is expected if the constant GC-to-halo mass ratio at higher masses arises from the central limit theorem in hierarchical assembly, though we note that observational measurements of halo mass are also more uncertain at low masses.

A potential concern is that the decrease in $M_{\rm GCs}$ predicted by our model at low masses is an artefact of the minimum GC mass we adopt in sampling GC masses. We test this possibility in Fig. 15, which compares the GC-to-halo mass relations predicted for two choices of the minimum GC mass. A lower minimum GC mass allows some GCs to form when the expectation value of the GC mass formed in an accretion event is $\Delta M_{\rm GCs} \ll 10^5\,{\rm M}_\odot$. This leads to a lower minimum halo mass for hosting a GC, but does not significantly change the GC-to-halo mass relation at $M_{\rm vir} > 5 \times 10^{10}\,{\rm M}_\odot$.

The fiducial model's prediction of a drop in the GC-to-halo mass relation at low mases is thus not primarily a consequence of the procedure for sampling GC masses. We have also verified that it is not sensitive to merger tree resolution: it results from the lower

gas surface densities and higher mass loading factors predicted for low-mass galaxies in our model. Substantial uncertainties remain in observational measurements of halo mass at the low-mass end. If the observed GC-to-halo mass relation is confirmed to remain constant at low masses, it will imply that our fiducial model assumptions break down at low halo masses. In the context of our fiducial model, a constant GC-to-halo mass ratio at low masses could result from a higher GC formation efficiency, $\Gamma_{\rm GCs}$, in low-mass halos, or a higher GC survival probability in low-mass galaxies.

4.3 Low GC formation efficiency

In order for our model to match the normalization of both the GC-to-halo mass relation and the stellar-to-halo mass relation, the coefficient α_Γ (equation 9), which determines the fraction of star formation that occurs in GC progenitors at asymptotically high surface density, must be quite small, with $\alpha_\Gamma \approx 2.1 \times 10^{-3}$ in the fiducial model. Such a low value of $\Gamma_{\rm GCs}$ is somewhat unexpected, since idealized simulations predict the total cluster formation efficiency, Γ (which represents the fraction of star formation that occurs in any bound clusters, not only GC progenitors), to asymptote to a value of order unity at high surface density (G18). We consider some possible explanations for this discrepancy below.

4.3.1 GC disruption

Our default model does not include GC disruption or stripping. Disruption and mass loss may significantly reduce the total GC mass at $z = 0$ relative to the total GC mass formed, increasing the value of α_Γ required to match observations. However, the fraction of the total GC mass formed that survives until $z = 0$ is highly uncertain, as the efficiency of GC disruption depends sensitively both on the redshift at which GCs form (e.g. Katz & Ricotti 2014; Carlberg 2017) and on the spatial distribution of GCs within their host halos (Chernoff & Weinberg 1990; Gieles & Baumgardt 2008; Kruijssen & Mieske 2009; Kruijssen et al. 2012; Kruijssen 2015; Pfeffer et al. 2018).

The simplified model for disruption and evaporation that we test in Appendix D leads to a significant but not overwhelming reduction in the total GC mass, requiring an increase in α_Γ by a factor of $\sim$2.6. Some other models, particularly those attempting to explain observations of anomalous abundances in GCs as the result of enrichment from a population of stars that was preferentially stripped at late times, invoke much higher disruption efficiencies, typically requiring a reduction in bound GC mass between formation and $z = 0$ by factors of 10–100 (D'Ercole et al. 2008; Conroy & Spergel 2011; Conroy 2012). Combined with the other factors discussed below, such efficient disruption could bring the value of α_Γ required to match the observed GC-to-halo mass relation into agreement with the expectation from idealized simulations. We note, however, that the trends predicted by self-enrichment models between the fraction of GC stars with anomalous abundances and other cluster properties are generally not observed (Bastian & Lardo 2015). If GCs *were* much more massive when they formed than they are today, then young GCs likely contribute significantly to the faint end of the high-z luminosity function (Bouwens et al. 2017; Boylan-Kolchin 2017a). Future observations with *JWST* will be able to test such a scenario. Searches for disrupted GC stars in the MW bulge and stellar halo (e.g. Martell et al. 2016; Schiavon et al. 2017) can also place limits on the efficiency of GC disruption.

4.3.2 Contributions from lower-mass clusters

Because low-mass clusters ($m \lesssim 10^5$ M$_\odot$ at birth) are expected to be strongly affected by two-body evaporation over a Hubble time (Spitzer 1987; Fall & Zhang 2001; Prieto & Gnedin 2008; Muratov & Gnedin 2010; Gieles, Heggie & Zhao 2011), we do not consider them as possible progenitors of $z = 0$ GCs. However, the total cluster formation efficiency Γ *does* include bound lower-mass clusters, so we generically expect Γ to exceed $\Gamma_{\rm GCs}$. For a dn/d$m \sim m^{-2}$ initial cluster mass function, each decade of cluster mass contributes equal total mass, so the total mass formed in all clusters is expected to exceed the mass formed in clusters sufficiently massive to survive until $z = 0$ by a factor of a few. The shape of the initial cluster mass function is imperfectly understood (e.g. Vesperini & Zepf 2003; Parmentier & Gilmore 2005; Elmegreen 2010). If GCs that survive until $z = 0$ represent the objects near the high-mass cutoff of a Schechter-type mass function, the mass fraction in since-disrupted lower-mass clusters could be much higher.

4.3.3 Additional requirements for bound cluster formation

In our model, the fraction of star formation that occurs in bound clusters depends only on the local gas surface density, $\Sigma_{\rm GMC}$. The low value of α_Γ that we find needed to match observations with this model may indicate that other conditions must be met for the formation of GC progenitors, such that our current model based on surface density alone overestimates the fraction of star formation that occurs in massive bound clusters at fixed surface density. For example, the fraction of stars forming in bound clusters likely also depends to some degree on factors such as cloud geometry, the cloud virial parameter, the local tidal field, the amount of shear from neighboring clouds, the Toomre Q parameter, and the gas metallicity (Krumholz & McKee 2005; Howard, Pudritz & Harris 2016; Kruijssen 2012).

It is also possible that our model underestimates the true critical density for GC formation, $\Sigma_{\rm crit} = 3000$ M$_\odot$ pc^{-2}, or equivalently, that our merger tree gas calculations overestimate $\Sigma_{\rm GMC}$. Increasing the value of $\Sigma_{\rm crit}$ does increase the value of α_Γ needed to match the observed GC-to-halo mass ratio for any positive β_Γ (see Appendix B). However, for the fiducial value of $\beta_\Gamma = 1$, it is necessary to increase $\Sigma_{\rm crit}$ by two orders of magnitude in order to produce an order-unity value of α_Γ. Because idealized cloud-collapse calculations find values of $\Sigma_{\rm crit}$ similar to our fiducial value (or lower; e.g. Raskutti et al. 2016), we regard a much-higher value of $\Sigma_{\rm crit}$ as unlikely.

4.4 Limitations of the model

4.4.1 Treatment of mergers

Our model treats GC formation as the high-surface density extension of normal star formation. Because the gas surface density in our model is directly linked to the mass accretion rate calculated from merger trees, a large fraction of GCs form following major mergers (See Fig. E1). This prediction is consistent with observations of massive clusters forming in nearby major mergers (e.g. Wilson et al. 2006; Bastian et al. 2009), and with simulations that find massive bound clusters to form during major mergers (e.g. Bournaud, Duc & Emsellem 2008). Major mergers are explicitly modeled as the dominant site of GC formation in several semi-analytic models (Ashman & Zepf 1992; Muratov & Gnedin 2010; Li & Gnedin 2014; Choksi et al. 2018).

In our model, the increased star and GC formation following major mergers is simply a result of the increased inflow rate. However, major mergers are also known to produce large-scale torques that can compress gas and drive it towards the galactic center (e.g. Mihos & Hernquist 1996; Hopkins et al. 2013), potentially amplifying the subsequent burst of star formation. Our model is designed to approximate this phenomenologically, not to model the details of star or GC formation with high fidelity.

A related limitation is that equilibrium models may be less applicable at high redshift, when galaxies may be in a 'gas accumulation phase' and cannot process gas as fast as they receive it (e.g. Davé et al. 2012; Krumholz & Dekel 2012). Following the arguments of Davé et al. (2012, their equation 15), we find that more than 95 per cent of GCs form after the redshift $z_{\rm eq}$ when galaxies reach equilibrium on average. However, galaxies can depart from equilibrium at later times during major mergers.[11] We find that roughly half of the GCs in our model form during such periods. Exploring the effects of a gas accumulation phase on GC formation is a possible avenue for future work.

4.4.2 *Merger time-scale and dynamical friction*

In our model, star and GC formation events corresponding to a particular merger event occur immediately following the merger in the merger tree, on a time-scale set by the galaxy dynamical time (Section 2.2.3). In reality, accreted objects will orbit within their host halos as satellites before spiraling to the center of the main halo (e.g. Lacey & Cole 1993; Cole et al. 2000; Boylan-Kolchin, Ma & Quataert 2008). The dynamical friction time-scale scales with the dynamical time of the main halo (which is shorter at high redshift) and with the mass ratio of the accreted halo to the primary halo. The majority of GCs in our model form at early times and in major mergers. In these cases, the dynamical friction time-scale is short, minimizing the error caused by treating the initial inspiral as instantaneous. A small fraction of GCs in our model do form at late times in halos in which the dynamical friction time-scale is of order the Hubble time; in these cases, the model may overpredict the true SFR and GC formation rate. In a model similar to the one introduced in this work, Li & Gnedin (2014) tested the effects of including an analytic model for dynamical friction. They found that including dynamical friction typically reduced the number of GCs at $z = 0$ by 5–10 per cent.

It *is* appropriate to account for the GCs of potentially unmerged satellites (which only contribute a large fraction of the total GC mass in cluster-mass halos) when calculating $M_{\rm GCs}$. Observational studies that find a constant GC-to-halo mass ratio at high halo masses also include satellite GC populations, either by direct counting or by extrapolating the GC number density profile to large radii.

ACKNOWLEDGEMENTS

We are grateful to Yu Lu for making public his implementation of the Parkinson & Cole merger tree algorithm. We thank the anonymous referee for constructive comments that improved the paper. We also thank Mike Grudic, Charlie Conroy, Chris McKee, Joss Bland-Hawthorn, Aldo Rodriguez-Puebla, and Rohan Naidu for helpful discussions. Ideas for this project were developed in part at the Near/Far Globular Cluster Workshop in Napa in 2017 December. KE and DRW are grateful for the hospitality provided by the MPIA in Heidelberg during the writing of this paper. KE acknowledges support from an NSF Graduate Research Fellowship. EQ and KE are supported by a Simons Investigator Award from the Simons Foundation and by NSF grant AST-1715070. DRW is supported by fellowships provided by the Alfred P. Sloan Foundation and the Alexander von Humboldt Foundation. MBK acknowledges support from NSF grant AST-1517226 and CAREER grant AST-1752913 and from NASA grants NNX17AG29G and HST-AR-13888, HST-AR-13896, HST-AR-14282, HST-AR-14554, HST-AR-15006, HST-GO-12914, and HST-GO-14191 from STScI. The analysis in this paper relied on the PYTHON packages `NumPy` (Van Der Walt, Colbert & Varoquaux 2011), `Matplotlib` (Hunter 2007), and `AstroPy` (Astropy Collaboration 2013).

[11] We estimate that galaxies will depart from equilibrium and temporarily accumulate gas when the SFR implied by the equilibrium model exceeds the gas consumption rate, i.e. when $\dot{M}_{\rm gas,\,in}/(1+\eta) > 0.02 M_{\rm gas}/t_{\rm dyn}$, where the gas consumption rate is based on estimates from the local Universe. During such periods, the SFR is likely somewhat lower than predicted by the equilibrium model.

REFERENCES

Adamo A., Östlin G., Bastian N., Zackrisson E., Livermore R. C., Guaita L., 2013, ApJ, 766, 105
Amorisco N. C., 2018, preprint (arXiv:1802.00812)
Amorisco N. C., Monachesi A., Agnello A., White S. D. M., 2018, MNRAS, 475, 4235
Ashman K. M., Zepf S. E., 1992, ApJ, 384, 50
Astropy Collaboration, 2013, A&A, 558, A33
Bailin J., Harris W. E., 2009, ApJ, 695, 1082
Bastian N., Goodwin S. P., 2006, MNRAS, 369, L9
Bastian N., Lardo C., 2015, MNRAS, 453, 357
Bastian N., Lardo C., 2017, preprint (arXiv:1712.01286)
Bastian N., Trancho G., Konstantopoulos I. S., Miller B. W., 2009, ApJ, 701, 607
Baumgardt H., Kroupa P., 2007, MNRAS, 380, 1589
Baumgardt H., Côté P., Hilker M., Rejkuba M., Mieske S., Djorgovski S. G., Stetson P., 2009, MNRAS, 396, 2051
Beasley M. A., Baugh C. M., Forbes D. A., Sharples R. M., Frenk C. S., 2002, MNRAS, 333, 383
Beasley M. A., Trujillo I., Leaman R., Montes M., 2018, Nature, 555, 483
Becker M. R., 2015, preprint (arXiv:1507.03605)
Behroozi P. S., Wechsler R. H., Conroy C., 2013, ApJ, 770, 57
Bekki K., 2005, ApJ, 626, L93
Bekki K., Yahagi H., Nagashima M., Forbes D. A., 2008, MNRAS, 387, 1131
Blakeslee J. P., Tonry J. L., Metzger M. R., 1997, AJ, 114, 482
Bolatto A. D., Leroy A. K., Rosolowsky E., Walter F., Blitz L., 2008, ApJ, 686, 948
Boley A. C., Lake G., Read J., Teyssier R., 2009, ApJ, 706, L192
Bond J. R., Cole S., Efstathiou G., Kaiser N., 1991, ApJ, 379, 440
Bournaud F., Duc P.-A., Emsellem E., 2008, MNRAS, 389, L8
Bouwens R. J., Illingworth G. D., Oesch P. A., Caruana J., Holwerda B., Smit R., Wilkins S., 2015, ApJ, 811, 140
Bouwens R. J., van Dokkum P. G., Illingworth G. D., Oesch P. A., Maseda M., Ribeiro B., Stefanon M., Lam D., 2017, preprint (arXiv:1711.02090)
Boylan-Kolchin M., 2017a, preprint (arXiv:1711.00009)
Boylan-Kolchin M., 2017b, MNRAS, 472, 3120
Boylan-Kolchin M., Ma C.-P., Quataert E., 2008, MNRAS, 383, 93
Bressan A., Marigo P., Girardi L., Salasnich B., Dal Cero C., Rubele S., Nanni A., 2012, MNRAS, 427, 127
Brodie J. P., Huchra J. P., 1991, ApJ, 379, 157
Brodie J. P., Usher C., Conroy C., Strader J., Arnold J. A., Forbes D. A., Romanowsky A. J., 2012, ApJ, 759, L33
Bryan G. L., Norman M. L., 1998, ApJ, 495, 80
Bullock J. S., Dekel A., Kolatt T. S., Kravtsov A. V., Klypin A. A., Porciani C., Primack J. R., 2001, ApJ, 555, 240

Burkert A., Tremaine S., 2010, ApJ, 720, 516
Carlberg R. G., 2002, ApJ, 573, 60
Carlberg R. G., 2017, preprint (arXiv:1706.01938)
Chaboyer B. et al., 2017, ApJ, 835, 152
Chen Y., Girardi L., Bressan A., Marigo P., Barbieri M., Kong X., 2014, MNRAS, 444, 2525
Chen Y., Bressan A., Girardi L., Marigo P., Kong X., Lanza A., 2015, MNRAS, 452, 1068
Chernoff D. F., Weinberg M. D., 1990, ApJ, 351, 121
Choksi N., Gnedin O., Li H., 2018, preprint (arXiv:1801.03515)
Cioffi D. F., McKee C. F., Bertschinger E., 1988, ApJ, 334, 252
Cole S., Lacey C. G., Baugh C. M., Frenk C. S., 2000, MNRAS, 319, 168
Conroy C., 2012, ApJ, 758, 21
Conroy C., Spergel D. N., 2011, ApJ, 726, 36
Conroy C., Loeb A., Spergel D. N., 2011, ApJ, 741, 72
Corbett Moran C., Teyssier R., Lake G., 2014, MNRAS, 442, 2826
Côté P., Marzke R. O., West M. J., 1998, ApJ, 501, 554
D'Ercole A., Vesperini E., D'Antona F., McMillan S. L. W., Recchi S., 2008, MNRAS, 391, 825
Davé R., Oppenheimer B. D., Finlator K., 2011, MNRAS, 415, 11
Davé R., Finlator K., Oppenheimer B. D., 2012, MNRAS, 421, 98
de Boer T. J. L., Fraser M., 2016, A&A, 590, A35
Dekel A. et al., 2009, Nature, 457, 451
Desmond H., Mao Y.-Y., Wechsler R. H., Crain R. A., Schaye J., 2017, MNRAS, 471, L11
Diemand J., Madau P., Moore B., 2005, MNRAS, 364, 367
Dotter A., Sarajedini A., Anderson J., 2011, ApJ, 738, 74
Durrell P. R. et al., 2014, ApJ, 794, 103
El-Badry K., Wetzel A., Geha M., Hopkins P. F., Kereš D., Chan T. K., Faucher-Giguère C.-A., 2016, ApJ, 820, 131
El-Badry K. et al., 2018a, preprint (arXiv:1804.00659)
El-Badry K. et al., 2018b, MNRAS, 473, 1930
Elmegreen B. G., 2008, ApJ, 672, 1006
Elmegreen B. G., 2010, ApJ, 712, L184
Elmegreen B. G., 2017, ApJ, 836, 80
Elmegreen B. G., Efremov Y. N., 1997, ApJ, 480, 235
Fall S. M., Chandar R., 2012, ApJ, 752, 96
Fall S. M., Efstathiou G., 1980, MNRAS, 193, 189
Fall S. M., Rees M. J., 1985, ApJ, 298, 18
Fall S. M., Zhang Q., 2001, ApJ, 561, 751
Fall S. M., Chandar R., Whitmore B. C., 2005, ApJ, 631, L133
Fall S. M., Krumholz M. R., Matzner C. D., 2010, ApJ, 710, L142
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
Fitts A. et al., 2018, preprint (arXiv:1801.06187)
Forbes D. A., Read J. I., Gieles M., Collins M. L. M., 2018, MNRAS
Förster Schreiber N. M. et al., 2009, ApJ, 706, 1364
Förster Schreiber N. M. et al., 2011, ApJ, 739, 45
Garrison-Kimmel S. et al., 2017, preprint (arXiv:1712.03966)
Georgiev I. Y., Puzia T. H., Goudfrooij P., Hilker M., 2010, MNRAS, 406, 1967
Geyer M. P., Burkert A., 2001, MNRAS, 323, 988
Gieles M., Baumgardt H., 2008, MNRAS, 389, L28
Gieles M., Heggie D. C., Zhao H., 2011, MNRAS, 413, 2509
Gnedin O. Y., Lee H. M., Ostriker J. P., 1999, ApJ, 522, 935
Gnedin O. Y., Ostriker J. P., Tremaine S., 2014, ApJ, 785, 71
Goddard Q. E., Bastian N., Kennicutt R. C., 2010, MNRAS, 405, 857
Gonzalez A. H., Sivanandam S., Zabludoff A. I., Zaritsky D., 2013, ApJ, 778, 14
Griffen B. F., Dooley G. A., Ji A. P., O'Shea B. W., Gómez F. A., Frebel A., 2018, MNRAS, 474, 443
Grudić M. Y., Hopkins P. F., Quataert E., Murray N., 2018a, preprint (arXiv:1804.04137)
Grudić M. Y., Hopkins P. F., Faucher-Giguère C.-A., Quataert E., Murray N., Kereš D., 2018b, MNRAS, 475, 3511
Harris W. E., 1996, AJ, 112, 1487
Harris W. E., Whitmore B. C., Karakla D., Okoń W., Baum W. A., Hanes D. A., Kavelaars J. J., 2006, ApJ, 636, 90
Harris W. E., Harris G. L. H., Alessi M., 2013, ApJ, 772, 82
Harris W. E., Harris G. L., Hudson M. J., 2015, ApJ, 806, 36
Harris W. E., Blakeslee J. P., Whitmore B. C., Gnedin O. Y., Geisler D., Rothberg B., 2016, ApJ, 817, 58
Harris W. E., Ciccone S. M., Eadie G. M., Gnedin O. Y., Geisler D., Rothberg B., Bailin J., 2017a, ApJ, 835, 101
Harris W. E., Blakeslee J. P., Harris G. L. H., 2017b, ApJ, 836, 67
Harris G. L. H., Poole G. B., Harris W. E., 2014, MNRAS, 438, 2117
Hirschmann M., Khochfar S., Burkert A., Naab T., Genel S., Somerville R. S., 2010, MNRAS, 407, 1016
Hopkins P. F., Quataert E., Murray N., 2012, MNRAS, 421, 3488
Hopkins P. F., Cox T. J., Hernquist L., Narayanan D., Hayward C. C., Murray N., 2013, MNRAS, 430, 1901
Howard C. S., Pudritz R. E., Harris W. E., 2016, MNRAS, 461, 2953
Huang K.-H. et al., 2017, ApJ, 838, 6
Hudson M. J., Harris G. L., Harris W. E., 2014, ApJ, 787, L5
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Ibata R., Nipoti C., Sollima A., Bellazzini M., Chapman S. C., Dalessandro E., 2013, MNRAS, 428, 3648
Jahnke K., Macciò A. V., 2011, ApJ, 734, 92
Jiang F., van den Bosch F. C., 2014, MNRAS, 440, 193
Johnson L. C. et al., 2016, ApJ, 827, 33
Katz H., Ricotti M., 2013, MNRAS, 432, 3250
Katz H., Ricotti M., 2014, MNRAS, 444, 2377
Kavelaars J. J., 1999, in Merritt D. R., Valluri M., Sellwood J. A., eds, ASP Conf. Ser. Vol. 182, Galaxy Dynamics - A Rutgers Symposium. Astron. Soc. Pac., San Francisco
Keto E., Ho L. C., Lo K.-Y., 2005, ApJ, 635, 1062
Kim J.-G., Kim W.-T., Ostriker E. C., 2016, ApJ, 819, 137
Kimm T., Cen R., Rosdahl J., Yi S. K., 2016, ApJ, 823, 52
Kirby E. N., Cohen J. G., Guhathakurta P., Cheng L., Bullock J. S., Gallazzi A., 2013, ApJ, 779, 102
Kravtsov A. V., 2013, ApJ, 764, L31
Kravtsov A. V., Gnedin O. Y., 2005, ApJ, 623, 650
Kravtsov A. V., Vikhlinin A. A., Meshcheryakov A. V., 2018, Astron. Lett., 44, 8
Kroupa P., 2001, MNRAS, 322, 231
Kroupa P., Weidner C., Pflamm-Altenburg J., Thies I., Dabringhausen J., Marks M., Maschberger T., 2013, The Stellar and Sub-Stellar Initial Mass Function of Simple and Composite Populations. p. 115
Kruijssen J. M. D., 2012, MNRAS, 426, 3008
Kruijssen J. M. D., 2015, MNRAS, 454, 1658
Kruijssen J. M. D., Mieske S., 2009, A&A, 500, 785
Kruijssen J. M. D., Maschberger T., Moeckel N., Clarke C. J., Bastian N., Bonnell I. A., 2012, MNRAS, 419, 841
Krumholz M. R., 2014, Phys. Rep., 539, 49
Krumholz M. R., Dekel A., 2012, ApJ, 753, 16
Krumholz M. R., McKee C. F., 2005, ApJ, 630, 250
Krumholz M. R., Ting Y.-S., 2018, MNRAS, 475, 2236
Kulier A., Ostriker J. P., Natarajan P., Lackner C. N., Cen R., 2015, ApJ, 799, 178
Lacey C., Cole S., 1993, MNRAS, 262, 627
Lada C. J., Lada E. A., 2003, ARA&A, 41, 57
Larsen S. S., Richtler T., 2000, A&A, 354, 836
Larsen S. S., Brodie J. P., Huchra J. P., Forbes D. A., Grillmair C. J., 2001, AJ, 121, 2974
Leaman R., VandenBerg D. A., Mendel J. T., 2013, MNRAS, 436, 122
Leauthaud A. et al., 2012, ApJ, 746, 95
Li H., Gnedin O. Y., 2014, ApJ, 796, 10
Li H., Gnedin O. Y., Gnedin N. Y., Meng X., Semenov V. A., Kravtsov A. V., 2017, ApJ, 834, 69
Lilly S. J., Carollo C. M., Pipino A., Renzini A., Peng Y., 2013, ApJ, 772, 119
Lim S., Peng E. W., Côté P., Sales L. V., den Brok M., Blakeslee J. P., Guhathakurta P., 2018, ApJ, 862, 82
Lin Y.-T., Mohr J. J., Stanford S. A., 2004, ApJ, 610, 745

Ma X., Hopkins P. F., Faucher-Giguère C.-A., Zolman N., Muratov A. L., Kereš D., Quataert E., 2016, MNRAS, 456, 2140
Mandelker N., Dekel A., Ceverino D., DeGraf C., Guo Y., Primack J., 2017, MNRAS, 464, 635
Martell S. L. et al., 2016, ApJ, 825, 146
Mashchenko S., Sills A., 2005, ApJ, 619, 243
McKee C. F., Ostriker E. C., 2007, ARA&A, 45, 565
McLaughlin D. E., Fall S. M., 2008, ApJ, 679, 1272
Mihos J. C., Hernquist L., 1996, ApJ, 464, 641
Moore B., 1996, ApJ, 461, L13
Moore B., Diemand J., Madau P., Zemp M., Stadel J., 2006, MNRAS, 368, 563
Moster B. P., Somerville R. S., Maulbetsch C., van den Bosch F. C., Macciò A. V., Naab T., Oser L., 2010, ApJ, 710, 903
Mo H. J., Mao S., White S. D. M., 1998, MNRAS, 295, 319
Muratov A. L., Gnedin O. Y., 2010, ApJ, 718, 1266
Murray N., Quataert E., Thompson T. A., 2005, ApJ, 618, 569
Murray N., Quataert E., Thompson T. A., 2010, ApJ, 709, 191
O'Shea B. W., Wise J. H., Xu H., Norman M. L., 2015, ApJ, 807, L12
Okamoto T., Gao L., Theuns T., 2008, MNRAS, 390, 920
Ostriker E. C., Shetty R., 2011, ApJ, 731, 41
Parkinson H., Cole S., Helly J., 2008, MNRAS, 383, 557
Parmentier G., Gilmore G., 2005, MNRAS, 363, 326
Peebles P. J. E., 1984, ApJ, 277, 470
Peebles P. J. E., Dicke R. H., 1968, ApJ, 154, 891
Peng E. W. et al., 2006, ApJ, 639, 95
Peng E. W. et al., 2008, ApJ, 681, 197
Peng E. W. et al., 2011, ApJ, 730, 23
Peng C. Y., 2007, ApJ, 671, 1098
Pfeffer J., Kruijssen J. M. D., Crain R. A., Bastian N., 2018, MNRAS, 475, 4309
Piotto G. et al., 2015, AJ, 149, 91
Planck CollaborationXIII, 2016, A&A, 594, A13
Portegies Zwart S. F., McMillan S. L. W., Gieles M., 2010, ARA&A, 48, 431
Prieto J. L., Gnedin O. Y., 2008, ApJ, 689, 919
Raskutti S., Ostriker E. C., Skinner M. A., 2016, ApJ, 829, 130
Renzini A., 2017, MNRAS, 469, L63
Richtler T., 2006, Bull. Astron. Soc. India, 34, 83
Rodríguez-Puebla A., Primack J. R., Behroozi P., Faber S. M., 2016a, MNRAS, 455, 2592
Rodríguez-Puebla A., Behroozi P., Primack J., Klypin A., Lee C., Hellinger D., 2016b, MNRAS, 462, 893
Rosenblatt E. I., Faber S. M., Blumenthal G. R., 1988, ApJ, 330, 191
Sanders N. E. et al., 2012, ApJ, 758, 132
Santos M. R., 2003, in Kissler-Patig M., ed., Extragalactic Globular Cluster Systems. p. 348
Schiavon R. P. et al., 2017, MNRAS, 465, 501
Schroetter I., Bouché N., Péroux C., Murphy M. T., Contini T., Finley H., 2015, ApJ, 804, 83
Shapiro K. L., Genzel R., Förster Schreiber N. M., 2010, MNRAS, 403, L36
Shibuya T., Ouchi M., Harikane Y., 2015, ApJS, 219, 15
Skinner M. A., Ostriker E. C., 2015, ApJ, 809, 187
Spitler L. R., Forbes D. A., 2009, MNRAS, 392, L1
Spitler L. R., Forbes D. A., Strader J., Brodie J. P., Gallagher J. S., 2008, MNRAS, 385, 361
Spitzer L., 1987, Dynamical Evolution of Globular Clusters.
Strader J., Brodie J. P., Cenarro A. J., Beasley M. A., Forbes D. A., 2005, AJ, 130, 1315
Strader J., Brodie J. P., Spitler L., Beasley M. A., 2006, AJ, 132, 2333
Strader J. et al., 2011, ApJS, 197, 33
Tang J., Bressan A., Rosenfield P., Slemer A., Marigo P., Girardi L., Bianchi L., 2014, MNRAS, 445, 4287
Thompson T. A., Krumholz M. R., 2016, MNRAS, 455, 334
Tonini C., 2013, ApJ, 762, 39
Tremonti C. A. et al., 2004, ApJ, 613, 898
Trenti M., Padoan P., Jimenez R., 2015, ApJ, 808, L35
Tsang B. T.-H., Milosavljevic M., 2017, preprint (arXiv:1709.07539)
Tutukov A. V., 1978, A&A, 70, 57
Vale A., Ostriker J. P., 2006, MNRAS, 371, 1173
Van Der Walt S., Colbert S. C., Varoquaux G., 2011, preprint (arXiv:1102.1523)
van Dokkum P. et al., 2018, ApJ, 856, L30
VandenBerg D. A., Brogaard K., Leaman R., Casagrande L., 2013, ApJ, 775, 134
Vanzella E. et al., 2017, MNRAS, 467, 4304
Vesperini E., Zepf S. E., 2003, ApJ, 587, L97
Wilson C. D., Harris W. E., Longden R., Scoville N. Z., 2006, ApJ, 641, 763
Wise J. H., Turk M. J., Norman M. L., Abel T., 2012, ApJ, 745, 50
Woodley K. A., Harris W. E., Puzia T. H., Gómez M., Harris G. L. H., Geisler D., 2010, ApJ, 708, 1335
Yang X., Mo H. J., van den Bosch F. C., Pasquali A., Li C., Barden M., 2007, ApJ, 671, 153
Yoon S.-J., Yi S. K., Lee Y.-W., 2006, Science, 311, 1129
Zaritsky D., Crnojević D., Sand D. J., 2016, ApJ, 826, L9
Zepf S. E., Ashman K. M., 1993, MNRAS, 264, 611
Zhao H., 2005, preprint (arXiv:e-print)
Zick T. O., Weisz D. R., Boylan-Kolchin M., 2018, preprint (arXiv:1802.06801)

APPENDIX A: COLD GAS ACCRETION RATE FROM TOTAL ACCRETION RATE

As discussed in Section 2.2, the equilibrium model that we use to estimate the SFR and gas surface density throughout a merger tree depends on the inflow rate of cold gas, $\dot{M}_{\rm gas,\,in}$. We calculate $\dot{M}_{\rm gas,\,in}$ from the total accretion rate using a suppression function $\zeta(M_{\rm vir,\,1}, M_{\rm vir,\,2}, z)$ (see equation 4) to crudely account for the effects of gas heating by the UV background, stellar feedback-driven winds, AGN quenching, and ambient hot gas in the primary halo. Here $M_{\rm vir,\,1}$ and $M_{\rm vir,\,2}$ represent the mass of the primary and accreted halo, respectively.

Following Davé et al. (2012), we model ζ as a product of several different terms accounting for different processes:

$$\zeta = \zeta_{\rm photo} \times \zeta_{\rm winds} \times \zeta_{\rm quench} \times \zeta_{\rm grav}. \tag{A1}$$

$\zeta_{\rm photo}$, $\zeta_{\rm winds}$, and $\zeta_{\rm quench}$ represent the reduction in the cold gas content, relative to the cosmic baryon fraction, of a galaxy being accreted; they are functions of $M_{\rm vir,\,2}$. On the other hand, $\zeta_{\rm grav}$ represents the suppression of cold gas accretion due to the hot gas in the primary halo; it is a function of $M_{\rm vir,\,1}$.

$\zeta_{\rm photo}$ represents the decrease in the cold gas mass of low-mass halos due to photoionization heating after the epoch of reionization. It drops to 0 below the 'photosuppression mass,' $M_\gamma(z)$, which increases from $\sim 10^8\,{\rm M}_\odot$ during reionization to a few $\times\, 10^9\,{\rm M}_\odot$ at $z = 0$. Following Okamoto, Gao & Theuns (2008, their equation 1), this is parametrized as

$$\zeta_{\rm photo} = \left\{1 + \left[2^{\alpha/\beta} - 1\right]\left(\frac{M_{\rm vir,2}}{M_\gamma\,(z)}\right)^{-\alpha}\right\}^{-\beta/\alpha}, \tag{A2}$$

where $\alpha = 2$ and $\beta = 3$. We calculate $M_\gamma(z)$ by interpolating on the results of the simulations presented in Okamoto et al. (2008, their fig. 5).

$\zeta_{\rm winds}$ represents the removal of cold gas from the accreted galaxy by winds prior to its accretion. The precise form of $\zeta_{\rm winds}$ is highly uncertain; but in general, winds are expected to affect low-mass galaxies more strongly than massive galaxies and to reduce the cold gas content more at late times than at early times. $\zeta_{\rm winds}$ is

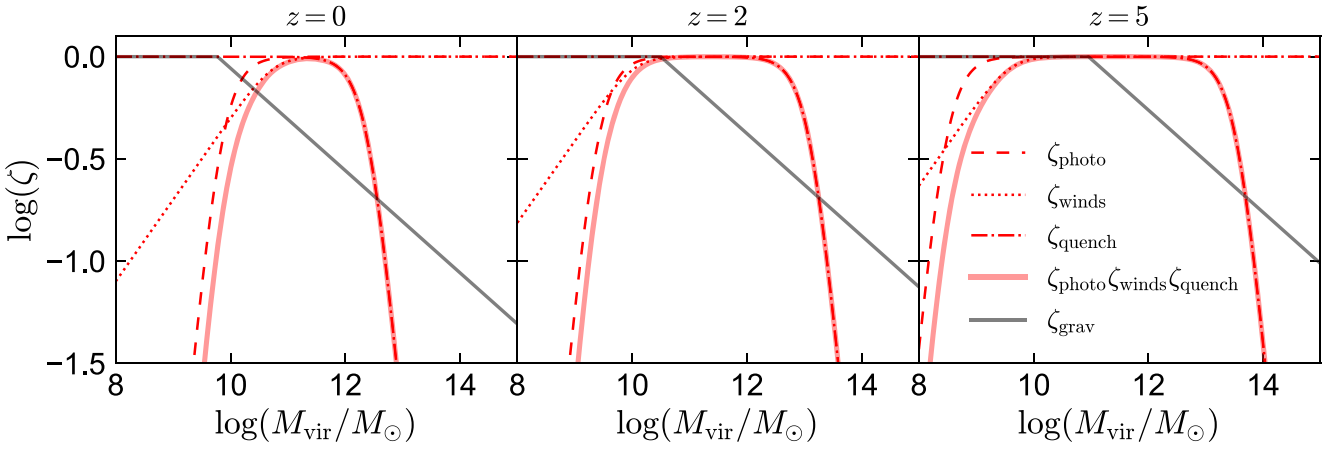

Figure A1. Components of the feedback parameter ζ that determines the cold gas mass accreted at each accretion event relative to the cosmic baryon fraction. Solid red curve represents the total reduction in the cold gas mass of a halo of mass $M_{\rm vir}$ being accreted at redshift z; dashed and dotted red lines show the contributions of individual feedback sources. Solid gray curve represents the additional suppression of cold gas accretion due to hot halo gas during an accretion event *into* a halo of mass $M_{\rm vir}$. Equations (A1)–(A5) give parametrizations of our adopted ζ.

parametrized analogously to $\zeta_{\rm photo}$:

$$\zeta_{\rm winds} = \left\{1 + \left[2^{\alpha/\beta} - 1\right]\left(\frac{M_{\rm vir,2}}{M_{\rm w}\,(z)}\right)^{-\alpha}\right\}^{-\beta/\alpha}, \quad \text{(A3)}$$

with $\alpha = 2$, $\beta = 0.4$, and $M_{\rm w}\,(z) = 10^{10}\,(1+z)^{-1.5}\ {\rm M}_\odot$.

$\zeta_{\rm quench}$ represents the effects of whatever processes heat gas in high-mass halos, likely connected to AGN. It drops to 0 above a 'quenching mass,' $M_{\rm q}(z)$, which increases at higher redshift. It is parametrized as

$$\zeta_{\rm quench} = \left\{1 + \left[2^{\alpha/\beta} - 1\right]\left(\frac{M_{\rm vir,2}}{M_{\rm q}\,(z)}\right)^{\alpha}\right\}^{-\beta/\alpha}, \quad \text{(A4)}$$

where $M_{\rm q}\,(z) = 10^{12.3}\,(1+z)^{1.47}\ {\rm M}_\odot$ and $\alpha = 2$ and $\beta = 3$.

Finally, $\zeta_{\rm grav}$ represents the suppression of cold gas accretion due to hot halo gas, which is heated by virial shocks. Following Faucher-Giguère, Kereš & Ma (2011) and Davé et al. (2012), we parametrize it as

$$\zeta_{\rm grav} = 0.47\left(\frac{1+z}{4}\right)^{0.38}\left(\frac{M_{\rm vir,1}}{10^{12}\,{\rm M}_\odot}\right)^{-0.25}. \quad \text{(A5)}$$

$\zeta_{\rm grav}$ suppresses cooling into high-mass halos, especially at lower redshifts. When equation (A5) exceeds unity, $\zeta_{\rm grav}$ is set to 1.

The combined effects of our parametrization of ζ are shown in Fig. A1. We emphasize that this model is largely phenomenological and is not expected to hold in detail for any galaxy. The main point of the model is to capture the facts that (a) cold gas fractions are higher at high redshift and in intermediate mass halos, and (b) gas accretion and cooling are suppressed in high-mass halos.

APPENDIX B: VARYING THE CRITICAL DENSITY FOR CLUSTER FORMATION

Fig. B1 shows examples of the GC formation efficiency parametrization assumed in our model for several values of β_Γ.

Throughout our analysis, we fixed the value of $\Sigma_{\rm crit}$, the critical surface density above which the GC formation efficiency plateaus, to 3000 ${\rm M}_\odot\,{\rm pc}^{-2}$. This is approximately the value predicted by analytic theory and found in idealized cloud-collapse simulations (e.g. Fall et al. 2010; Murray et al. 2010; Kim et al. 2016; Grudić et al. 2018a,b). As discussed in Section 4.3, this results in a low value of α_Γ, implying that even at asymptotically high surface densities, the fraction of stars forming in proto-GCs is low. In Fig. B2, we investigate how changing the value of $\Sigma_{\rm crit}$ changes the value of α_Γ, which is required to match the observed GC-to-halo mass relation at the high-mass end. We fix $\beta_\eta = 1/3$.

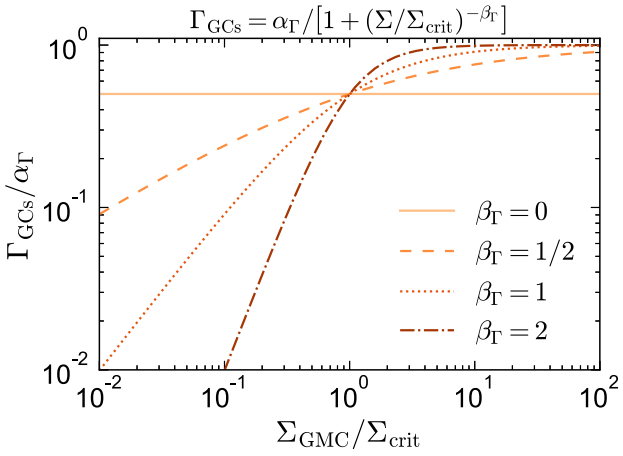

Figure B1. Parametric form of the GC formation efficiency, $\Gamma_{\rm GCs} = \dot{M}_{\rm GCs}/{\rm SFR}$, assumed in our model (equation 9). $\Gamma_{\rm GCs}$ dictates the fraction of star formation that occurs in massive bound clusters that survive until $z = 0$. In our model, $\Gamma_{\rm GCs}$ goes to 0 at $\Sigma_{\rm GMC} \ll \Sigma_{\rm crit}$ and plateaus at a value α_Γ at $\Sigma_{\rm GMC} \gg \Sigma_{\rm crit}$; the free parameter β_Γ determines the steepness with which $\Gamma_{\rm GCs}$ falls off at low surface densities.

As expected, a higher value of $\Sigma_{\rm crit}$ requires a higher value of α_Γ, because a smaller fraction of star-forming events have $\Sigma_{\rm GMC} \gtrsim \Sigma_{\rm crit}$. However, even for large values of β_Γ (i.e. a sharp truncation of GC formation at $\Sigma_{\rm GMC} < \Sigma_{\rm crit}$), it is necessary to increase $\Sigma_{\rm crit}$ by more than an order of magnitude in order to bring α_Γ to order unity. Such high values of $\Sigma_{\rm crit}$ significantly exceed those predicted by idealized cloud-collapse simulations.

In Fig. B3, we show the effects of increasing $\Sigma_{\rm crit}$ on the predicted cosmic GC formation rate. The solid line is the same as the solid line in the top panel of Fig. 8, since the cluster formation efficiency is independent of $\Sigma_{\rm crit}$ for $\beta_\Gamma = 0$. With $\Sigma_{\rm crit} = 10^5\,{\rm M}_\odot\,{\rm pc}^{-2}$, the cosmic GC formation rate varies more strongly with β_Γ, and in-

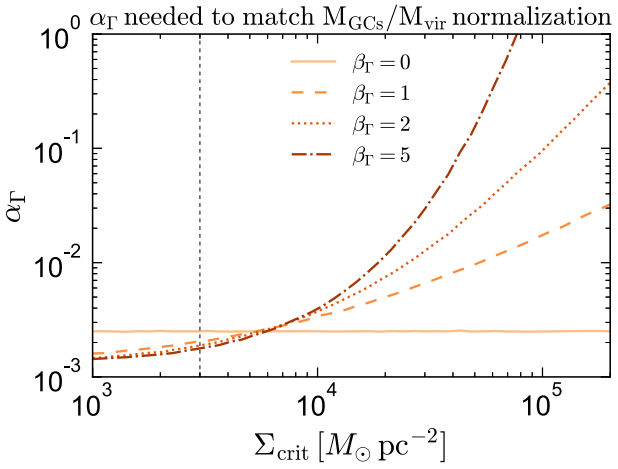

Figure B2. Normalization of the cluster formation efficiency, α_Γ (equation 9), versus Σ_{crit}, the critical density above which the cluster formation efficiency Γ_{GCs} plateaus. For each β_Γ and Σ_{crit}, we plot the value of α_Γ that matches the normalization of the observed GC-to-halo mass relation at the high-mass end (Fig. 2). Choosing a much higher value of Σ_{crit} than our default value of 3000 $M_\odot$ pc^{-2} increases the value of α_Γ.

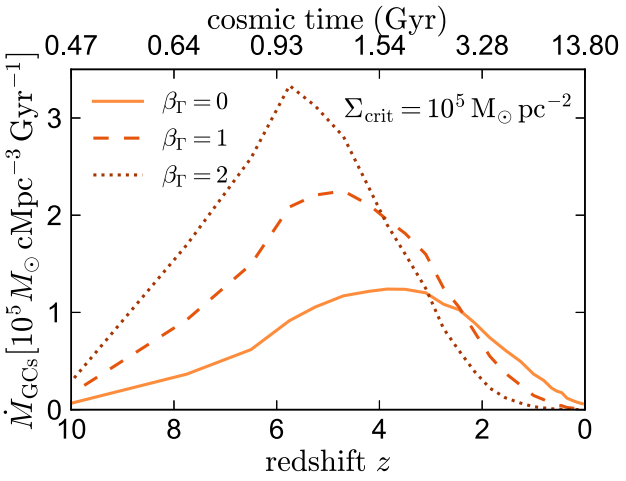

Figure B3. Cosmic GC formation rate (similar to the top panel of Fig. 8) using a higher critical density for GC formation. We fix $\beta_\eta = 1/3$. Increasing Σ_{crit} causes GCs to form earlier and makes the epoch of GC formation more sensitive to β_Γ.

creasing β_Γ causes the epoch of GC formation to move to earlier times. This makes GCs contribute more to the UV luminosity density during reionzation, though still only at the ∼5 per cent level. Although we consider a value of Σ_{crit} as high as 10^5 $M_\odot$ pc^{-2} unlikely, it is possible that approximations in our model, such as the adopted merger time-scale or the assumption of $\Sigma_{GMC} = 5 \times \Sigma_{gas}$, overestimate the true value of Σ_{GMC}. Using a higher value of Σ_{crit} has exactly the same effect on the predicted GC population as using a longer τ_{merger} (equation 10) or a lower value of Σ_{GMC} relative to Σ_{gas}.

APPENDIX C: STELLAR METALLICITY DISTRIBUTION

Changing the value of β_η changes the metallicity distribution for both GCs and field stars. The effect of varying β_η on the metallicity

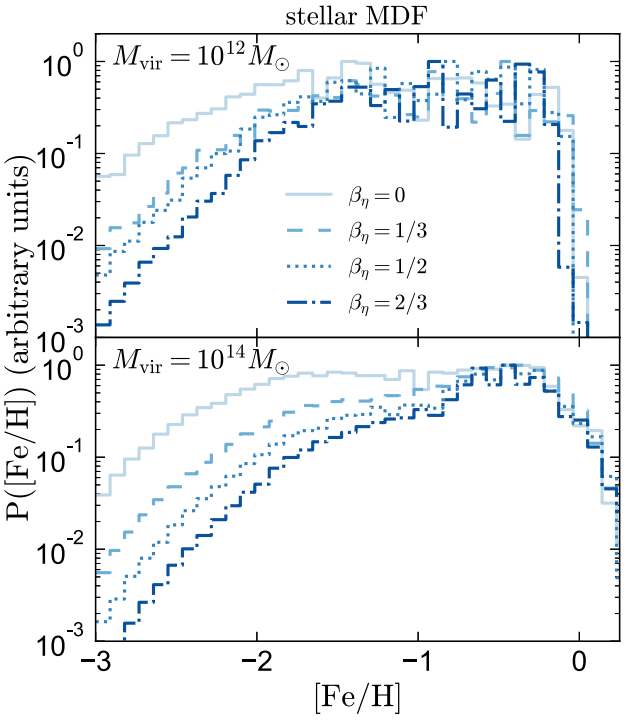

Figure C1. Metallicity distribution of all stars in a halo with $M_{vir} = 10^{12}$ $M_\odot$ (top) and $M_{vir} = 10^{14}$ $M_\odot$ (bottom) at $z = 0$, for different scalings of η with halo mass. Increasing β_η decreases the fraction of low-metallicity stars. Beyond the GC colour distribution (Fig. 10), the implied metallicity distribution for all stars provides a secondary constraint on β_η. The fiducial model with $\beta_\eta = 1/3$ somewhat underestimates the mean metallicity of field stars.

of all stars in a halo (i.e. the result of integrating equation (3) over all nodes in the merger tree) is shown in Fig. C1. Increasing the value of β_η suppresses star formation in low-mass halos and thus reduces the fraction of low-metallicity stars. This serves as an additional constraint on β_η: although models with $\beta_\eta = 0$ can produce plausible GC metallicities, ages, and colour distributions (Figs 9 and 10), $\beta_\eta \gtrsim 1/3$ is required for a plausible stellar metallicity distribution.

The fiducial model somewhat underestimes the typical metallicity of field stars. At $M_{vir} = 10^{12}$ $M_\odot$, it predicts a mean metallicity for all stars in the halo of [Fe/H] ≈ −1; we find the same value for the simulated MW-mass galaxies studied in El-Badry et al. (2018b) to be [Fe/H] ≈ −0.4. Perhaps relatedly, the fiducial model predicts the cosmic star formation rate to peak at $z \sim 3.5$ (Fig. 8), which is earlier than the value $z \sim 2$ found observationally. This may indicate that our approximations for the suppression of cold gas accretion (Appendix A) prevent accretion at late times too strongly, or that the accretion-based model overestimates the SFR at early times due to a gas accumulate phase (see also Rodríguez-Puebla et al. 2016a).

APPENDIX D: EFFECTS OF CLUSTER DISRUPTION

Here we test the effects of a simplified model for GC mass loss and disruption due to both tidal fields and two-body evaporation. We use the model from Choksi et al. (2018, their equation 9). The model attempts to account for both disruption due to tidal fields (averaging over all spatial distributions) and two-body evaporation; tidal effects are dominant for all but the least massive clusters. In

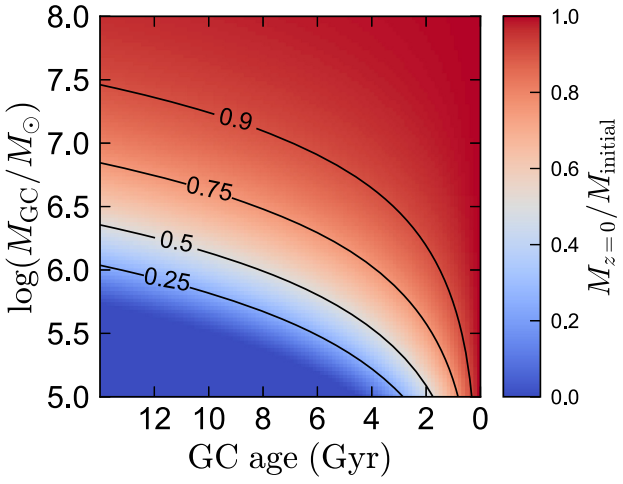

Figure D1. Effective disruption model taken from Choksi et al. (2018) and implemented in Fig. D2. Colour scale shows the fraction of a single GC's initial mass that survives at $z = 0$; GCs in dark blue regions of parameter space are disrupted completely.

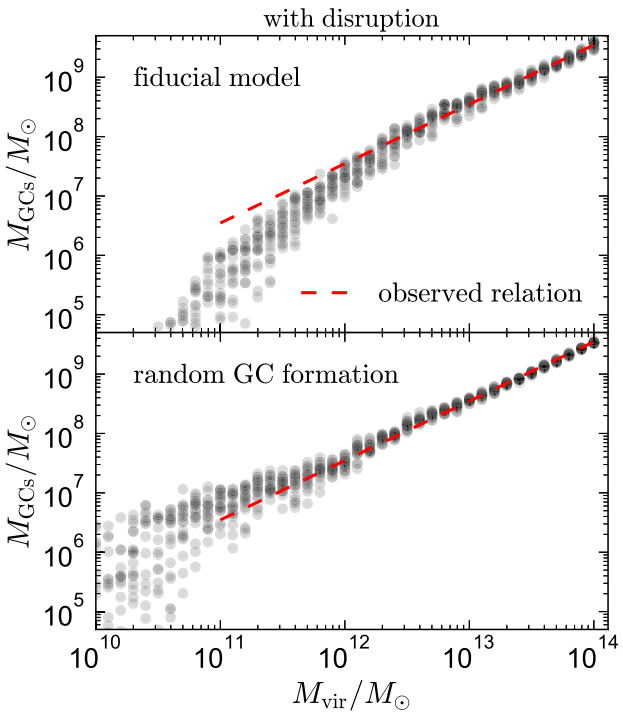

Figure D2. GC-to-halo mass relation for our fiducial model (top) and random GC formation (bottom), but now including the age- and mass-dependent GC disruption model from Choksi et al. (2018) (see Fig. D1). Including disruption has little effect on the linearity of the GC-to-halo mass relation (compare to Fig. 2).

this model, the mass of a GC at $z = 0$ depends only on its age and initial mass.

The combined total effects of these processes as we implement them are illustrated in Fig. D1. In this model, old GCs with initial masses $m \lesssim 10^6$ M$_\odot$ are disrupted or lose a dominant fraction of their mass by $z = 0$; GCs that form at later times or are more massive lose a smaller fraction of their initial mass. The disruption model was calibrated by Choksi et al. (2018) to produce a realistic $z = 0$ cluster mass function. Choksi et al. (2018) also showed that the combination of this disruption model and their sampling procedure for drawing GCs masses (which we also adopt) reproduces the observed 'blue tilt'; i.e. the trend of more massive blue GCs to be more metal rich on average (e.g. Strader et al. 2006).

Fig. D2 shows the GC-to-halo mass relation predicted by our model after implementing the disruption model. Applying the GC disruption model causes the normalization of the GC-to-halo mass ratio at the high mass end to drop by a factor of 2.6 for the fiducial model and a factor of 4 for the random model, so we increase the free parameter α_Γ by a factor of 2.6 and increase the probability of each halo hosting a GC formation event in the random model by a factor of 4. After these adjustments are made, both the fiducial and random GC formation models produce the a similar constant GC-to-halo mass relation as when no disruption is included. Disruption primarily affects old and low-mass GCs, but the fraction of GCs in a halo at $z = 0$ that are old or low mass is not a strong function of halo mass, so in this model disruption has little effect beyond changing the overall normalization of the total GC mass formed.

Although we do not explore the other effects of this GC disruption model in detail, we find that applying it also does not change our conclusion from Section 3.3.2 that bimodal GC colour and metallicity distributions can be produced for a wide range of model parameters. The model preferentially disrupts old GCs, which are bluer than average, so when β_Γ and β_η are held fixed, including disruption tends to increase the mass fraction of GCs that are red.

APPENDIX E: DISTRIBUTIONS OF GC PROPERTIES

Fig. E1 shows distributions of several GC properties for the GC populations of halos at three different mass scales. Each distribution is an ensemble for the GC populations of 20 merger tree realizations. We show predictions for three different values of β_η.

The distributions of most GC properties are similar in halos of different $z = 0$ masses. However, the GC colour and metallicity distributions evolve with halo mass due to the adopted mass–metallicity relation. At high halo masses, GC colour distributions are often more strongly bimodal than GC metallicity distributions. This is a result of the coincidental alignment of GCs in the age-metallicity plane along lines of constant colour. Increasing β_η makes GC formation more efficient in higher-mass halos and thus increases the fraction of the GC population that is in the metal-rich, red mode.

Most GCs form in discs with $\Sigma_{\rm GMC} \sim 10^4$ M$_\odot$ pc^{-2}, higher than our adopted critical density $\Sigma_{\rm crit} = 3000$ M$_\odot$ pc^{-2}. The majority of GCs form in major mergers. This is not an explicit requirement of our GC formation model, but is simply a consequence of the fact that the gas accretion rate, SFR, and gas surface density are all highest during major mergers. For this reason, many of the predictions of our model are similar to those of models in which GC formation is explicitly tied to major mergers (e.g. Muratov & Gnedin 2010; Li & Gnedin 2014; Choksi et al. 2018).

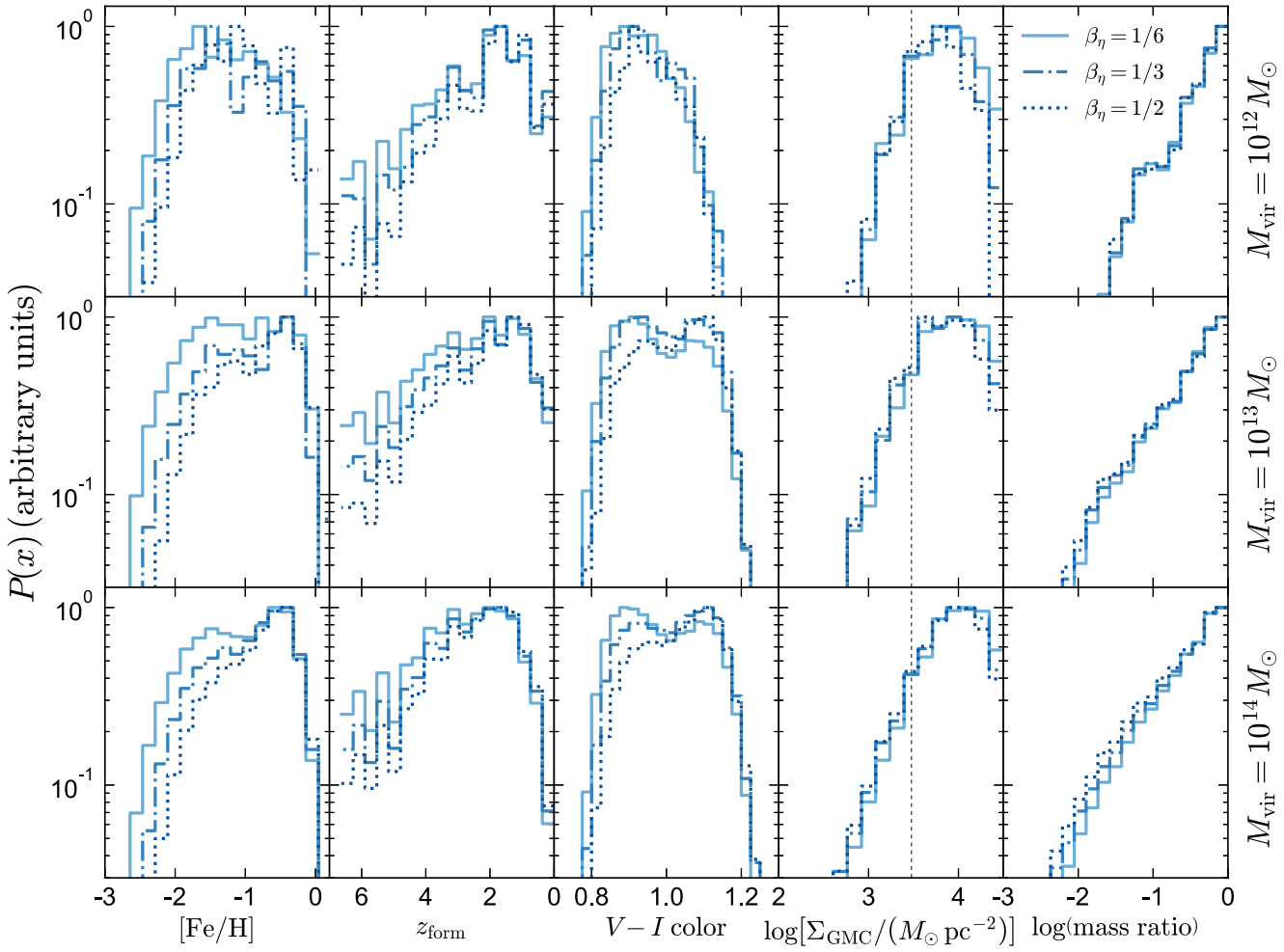

Figure E1. Mass-weighted distributions of GC metallicity, formation redshift, colour, GMC surface density at the time of formation, and the mass ratio of the merger event in which the GC formed. Different histograms show three values of β_η (equation 5); different rows show different halo masses at $z = 0$. All models assume $\beta_\Gamma = 1$ (equation 9).

This paper has been typeset from a TEX/LATEX file prepared by the author.