Constraining quenching time-scales in galaxy clusters by forward-modelling stellar ages and quiescent fractions in projected phase space

Andrew M. M. Reeves [1,2]★ Michael J. Hudson [1,2,3] and Kyle A. Oman [4,5]

[1]*Department of Physics and Astronomy, University of Waterloo, Waterloo, ON N2L 3G1, Canada*
[2]*Waterloo Centre for Astrophysics, University of Waterloo, Waterloo, ON N2L3G1, Canada*
[3]*Perimeter Institute for Theoretical Physics, Waterloo, ON N2L 2Y5, Canada*
[4]*Institute for Computational Cosmology, Durham University, South Road, Durham DH1 3LE, UK*
[5]*Department of Physics, Durham University, South Road, Durham DH1 3LE, UK*

ABSTRACT

We forward-model mass-weighted stellar ages (MWAs) and quiescent fractions (f_Q) in projected phase space (PPS), using data from the Sloan Digital Sky Survey, to jointly constrain an infall quenching model for galaxies in $\log(M_{\rm vir}/M_\odot) > 14$ galaxy clusters at $z \sim 0$. We find the average deviation in MWA from the MWA–$M_\star$ relation depends on position in PPS, with a maximum difference between the inner cluster and infalling interloper galaxies of ~ 1 Gyr. Our model employs infall information from N-body simulations and stochastic star-formation histories from the UNIVERSEMACHINE model. We find total quenching times of $t_Q = 3.7 \pm 0.4$ Gyr and $t_Q = 4.0 \pm 0.2$ Gyr after first pericentre, for $9 < \log(M_\star/M_\odot) < 10$ and $10 < \log(M_\star/M_\odot) < 10.5$ galaxies, respectively. By using MWAs, we break the degeneracy in time of quenching onset and time-scale of star formation rate (SFR) decline. We find that time of quenching onset relative to pericentre is $t_{\rm delay} = 3.5^{+0.6}_{-0.9}$ Gyr and $t_{\rm delay} = -0.3^{+0.8}_{-1.0}$ Gyr for $9 < \log(M_\star/M_\odot) < 10$ and $10 < \log(M_\star/M_\odot) < 10.5$ galaxies, respectively, and exponential SFR suppression time-scales are $\tau_{\rm env} \leq 1.0$ Gyr for $9 < \log(M_\star/M_\odot) < 10$ galaxies and $\tau_{\rm env} \sim 2.3$ Gyr for $10 < \log(M_\star/M_\odot) < 10.5$ galaxies. Stochastic star formation histories remove the need for rapid infall quenching to maintain the bimodality in the SFR of cluster galaxies; the depth of the green valley prefers quenching onsets close to first pericentre and a longer quenching envelope, in slight tension with the MWA-driven results. Taken together these results suggest that quenching begins close to, or just after pericentre, but the time-scale for quenching to be fully complete is much longer and therefore ram-pressure stripping is not complete on first pericentric passage.

Key words: galaxies: clusters: general – galaxies: evolution – galaxies: haloes – galaxies: star formation.

1 INTRODUCTION

Galaxies have long been known to exhibit a bimodality in their star formation rates (SFRs) such that the population of bluer galaxies with higher SFRs make up the 'star-forming main sequence' and the redder galaxies with lower SFRs forming the quiescent 'red sequence' (e.g. Bell et al. 2004; Brinchmann et al. 2004; Brammer et al. 2011; Muzzin et al. 2012). The quiescent fraction increases strongly with increasing galaxy stellar mass, likely due to some internal ('mass') quenching process, as well as increasing with the density of its surrounding environment (e.g. Cooper et al. 2007; Peng et al. 2010), due to some environmental quenching process(es). Despite significant exploration of quenching trends and their statistical relation to internal and environmental properties, it remains a matter of debate which quenching mechanisms are responsible for the bulk of quenching in the densest environments across cosmic time, particu-

larly galaxy clusters (Naab & Ostriker 2017; Wechsler & Tinker 2018).

Galaxies falling into groups or clusters may have their star formation affected by a number of physical processes. Initially, upon infall into a cluster, a galaxy may experience enhanced star formation, due to ram pressure compressing and triggering collapse of a galaxies' cold gas reservoir (see e.g. Jaffé et al. 2018; Vulcani et al. 2018; Roberts et al. 2021). If strong enough, ram pressure can strip the cold gas in star-forming galaxies, triggering relatively abrupt quenching (normally <1 Gyr is assumed to be indicative of this scenario, see e.g. Boselli et al. 2016; Fossati et al. 2018). Alternatively, numerous studies support stripping of just the hot halo gas could be leading to a 'strangulation' scenario, where a star-forming galaxy continues forming stars using its cold gas reservoir, but since its cold gas supply is not replenished, the galaxy then quenches once it runs out of cold gas (such as argued by Larson, Tinsley & Caldwell 1980; Balogh, Navarro & Morris 2000; Bekki, Couch & Shioya 2002; Boselli & Gavazzi 2006; McGee et al. 2009; Taranu et al. 2014; Paccagnella et al. 2016). Such a scenario is often simply parametrized by a 'delayed-then-rapid' quenching prescription (Wetzel et al. 2013),

with a range of time until quenching (after 'infall' – definitions for which vary) of 2–7 Gyr preferred by observations at $z = 0$ (see also e.g. De Lucia et al. 2012; Moster, Naab & White 2018; Rhee et al. 2020; Oman et al. 2021, among numerous others). Other effects, like harassment from galaxies passing near each other, as well as mergers, can affect the star formation in a galaxy (Boselli & Gavazzi 2006). Complicating this picture is the 'pre-processing' of galaxies in host haloes of a lower mass groups prior to their infall into a cluster, which can additionally enhance the quiescent fraction in clusters compared to accreting isolated galaxies from the field. Although widely recognized as playing an important role in explaining the environmental dependence of star formation, work remains in characterizing the significance of pre-processing for quenching, where exactly it occurs, and how it evolves with redshift (Fujita 2004; McGee et al. 2009; De Lucia et al. 2012; Webb et al. 2020). It is at least clear that quenching is significantly enhanced relative to the field at least up to $z \sim 1.5$ in rich galaxy groups (Gerke et al. 2007; Wetzel, Tinker & Conroy 2012; Mok et al. 2013; Reeves et al. 2021) and clusters (Van Burg et al. 2018; Pintos-Castro et al. 2019; Van Burg et al. 2020).

Constraining time-scales using forward modelling of observable quantities in projected phase space (PPS) in clusters can provide significant information for discerning when and where different physical quenching mechanisms may be dominant. In observational works, galaxy properties have shown a strong link to their position in PPS (for a variety of distinct examples see e.g. Mahajan, Mamon & Raychaudhury 2011; Smith et al. 2012; Muzzin et al. 2014; Jaffé et al. 2015; Noble et al. 2016; Gavazzi et al. 2018; Jaffé et al. 2018; Kelkar et al. 2019; Kim et al. 2020; Rhee et al. 2020; Roberts et al. 2021). Little detailed forward modelling work has been done on constraining quenching time-scales in PPS, aside from that developed in the series of papers by Oman, Hudson and collaborators (Oman, Hudson & Behroozi 2013; Oman & Hudson 2016; Oman et al. 2021). The second and third of these papers use the quiescent fraction of cluster galaxies in PPS coordinates as well as corresponding distribution of infall times inferred from N-body simulations. The PPS coordinates used in this context, namely the radial distance from cluster centre and line-of-sight velocity offset from its host system, only carry magnitude (i.e. no sign) information in practice, as it's generally not possible to distinguish between a satellite receding towards a background host or a background satellite receding away from a foreground host. This method provides a nearby interloper population on the cluster outskirts (outside some 3D radial cut) against which to compare cluster satellites. This has an advantage over using a generic field population: using these interlopers in the cluster outskirts provides an already pre-processed infalling galaxy sample, allowing better isolation of cluster physics from the physics involved in pre-processing.

Whereas many previous works that modelled environmental quenching have relied on quiescent fractions, observational indicators that are sensitive to the star formation history over longer time-scales should provide more information. Luminosity- or mass-weighted ages (MWAs) in particular not only provide complementary quantitative information about the star formation history of galaxies, but are also insensitive to whether a given galaxy's star formation history is assumed to be smooth or stochastic (see discussion in Section 4.2.2).

Studies using spectroscopic indicators of stellar age have long established that more massive galaxies, or, more accurately, galaxies with deeper potential wells, have older ages (Nelan et al. 2005; Graves et al. 2007). Dependence on environment is weaker: early studies focused on the age difference between field and cluster ellipticals after controlling for mass or velocity dispersion (Thomas et al. 2005; Bernardi et al. 2006). Radial trends of the stellar ages of galaxies within clusters (at fixed velocity dispersion) were found (Smith et al. 2006) with a stronger environmental dependence on the ages of dwarf galaxies (Smith et al. 2012).

Studies of stellar age in PPS are less common. Pasquali et al. (2019) considered luminosity-weighted stellar ages in zones of PPS around low redshift clusters, finding ages that are older by ~ 2 Gyr for the innermost lower velocity regions of PPS compared to the outermost regions. Recently, Khullar et al. (2022) studied D4000-derived ages in PPS in $0.3 < z < 1.1$ clusters, again finding older ages for PPS regions associated with the core of the cluster (see also Kim et al. 2022).

There has been even less effort modelling the effect of quenching on the stellar ages. Taranu et al. (2014) used infalling subhalo orbits from *N*-body simulations to forward-model the age-sensitive Balmer line Lick indices as a function of cluster-centric radius. Their models preferred long quenching time-scales occurring for galaxies that have passed within $\sim 0.5 r_{200c}$ (i.e. preferring slow 'strangulation' over rapid ram-pressure stripping).

As little work has been done forward modelling spectroscopically derived ages in clusters and none has been carried out in full PPS with a large sample (but see Upadhyay, Oman & Trager 2021, for a first attempt with a sample of 11 galaxies in the Coma cluster), the goal of this work is to extend the modelling of Oman & Hudson (2016) and Oman et al. (2021), who studied the quiescent fraction, f_Q, in PPS. Oman et al. (2021) found that using f_Q alone is not sufficient to constrain both the time of quenching onset and the time-scale over which quenching occurs on average and so they marginalized over this latter parameter to find the average time at which 50 per cent of infalling star-forming galaxies have quenched. Extending this and similar work could make use of, for example, the entire sSFR distribution, but as we will describe, the shape of the quiescent bump is sensitive to observational systematics, while the depth of the green valley is sensitive to choice of star formation history (smooth versus stochastic). In this paper, we will focus on extending Oman et al. (2021) by forward modelling the observed MWAs of galaxies in PPS, in addition to using quiescent fractions. This additional observable will enable joint constraints on both the time of onset of infall quenching and the time-scale for the duration of quenching preferred by both observables, for the population of galaxies as a whole.

Our paper is structured as follows. In Section 2, we describe how we parametrize PPS and the SDSS-based data sets, from which we get our f_Q and mean MWAs. We then, in Section 3, introduce our forward modelling approach (including sources of star formation histories and infall histories) and present our quiescent fraction and mean MWA modelling results in PPS as well as the constraints on quenching time-scales they provide. In Section 4, we discuss the robustness of our results, how they compare to literature, and what they imply for the dominant infall quenching processes. We then conclude in Section 5.

Unless otherwise specified, the following assumptions and conventions are used. Uncertainties are estimated from the 16th–84th percentile interval (equivalent to 1-σ for a Gaussian-distributed variable). Logarithms with base 10 ($\log_{10}$) are written simply as '$\log$' throughout this work. A flat ΛCDM cosmology consistent with the Planck 2015 cosmological parameters (Planck Collaboration XII 2016) is assumed, namely $H_0 = 68$ km s^{-1} Mpc^{-1}, $\Omega_m = 0.31$, and $\Omega_\Lambda = 0.69$. We define a virial overdensity at $z = 0$ in this work as ~ 360 times the background density, $\Omega_m \rho_c$, corresponding to the density of a recently virialized spherical top-hat solution (see e.g.

Bryan & Norman 1998).[1] A Chabrier (2003) initial mass function (IMF) is assumed throughout.

2 DATA AND SIMULATIONS

2.1 PPS selection and conventions

We adopt the convention of Oman et al. (2013) for selecting objects in PPS to allow direct comparison to observations: there are two sets of cluster-centred coordinates: (r, v) for the $2 \times 3D$ physical phase space coordinates and (R, V) for the normalized PPS coordinates. Specifically, the PPS coordinates consist of the distance to cluster centre perpendicular to the line-of-sight and the velocity component along the line of sight, which we choose to be the Cartesian Z-axis of a given simulation data set. The projected radius between the cluster centre and a galaxy is then

$$R = \left(\frac{1}{r_{\mathrm{vir,3D}}}\right)\sqrt{(r_X - r_{X,\mathrm{cls}})^2 + (r_Y - r_{Y,\mathrm{cls}})^2} \quad (1)$$

and the projected velocity relative to the motion of the cluster centre of mass is

$$V = \left(\frac{1}{\sigma_{\mathrm{3D}}}\right)\left|(v_Z - v_{Z,\mathrm{cls}}) + H(Z - Z_{\mathrm{cls}})\right|. \quad (2)$$

Using an absolute value for the projected velocity definition captures an assumption that the distances to clusters and satellites are not measured with enough accuracy to determine the sign of their relative velocity. The additional H term corrects for the Hubble flow. To allow easy comparison between clusters, for both simulations and observations, we have normalized the PPS coordinates. For ease of comparison to Oman & Hudson (2016) and Oman et al. (2021), radial coordinates are normalized by dividing by the 3D virial radius of the cluster, $r_{\mathrm{vir,3D}}$, and velocity coordinates by the 3D velocity dispersion, σ_{3D}, of a given cluster. As we are averaging over many clusters in both the orbit library and observed sample, we can assume that clusters are approximately spherically symmetric. This in turn allows relation of the 3D velocity dispersion to the observable 1D velocity dispersion by $\sigma_{\mathrm{3D}} \approx \sigma_{\mathrm{1D}}/\sqrt{3}$. We note that in order to ensure consistency between the normalization factors used for R and V, they should not be independent, i.e. the velocity dispersion should be derived from the virial radius or vice-versa. See Section 2.4 for details of how these are calculated for the SDSS data.

2.2 Infall histories: *N*-body simulation orbit library

We use the PPS orbital libraries from Oman et al. (2021); they base their methodology on Oman & Hudson (2016) to produce a catalogue of orbit parameter probability distributions from an *N*-body simulation. As described in Oman et al. (2021), the 'level 0' voids-in-void-in-voids simulation (VVV; Wang et al. 2020) is used, which has a box size of 500 h^{-1}Mpc on a side, mass resolution elements of 10^9 $h^{-1}\mathrm{M}_\odot$, and a force softening scale of 4.6 h^{-1}kpc. They run the simulation to scale factor $a = 2$, corresponding to redshift $z = -0.5$ or ~ 10 Gyr into the future (i.e. negative lookback times), allowing them to find the time of first pericentre for >99.9 per cent of resolved satellite galaxies. Host halo masses and satellite halo masses

[1] For a commonly used virial overdensity definition of $200\rho_c$, this corresponds to $M_{200c}/M_{\mathrm{vir}} \approx 0.81$ and $r_{200c}/r_{\mathrm{vir}} \approx 0.73$ (Oman et al. 2021), assuming a concentration parameter of $c_{200} = r_{200c}/r_s \approx 5$ (where r_s is the NFW profile scale radius), a typical value for clusters (Ludlow et al. 2014).
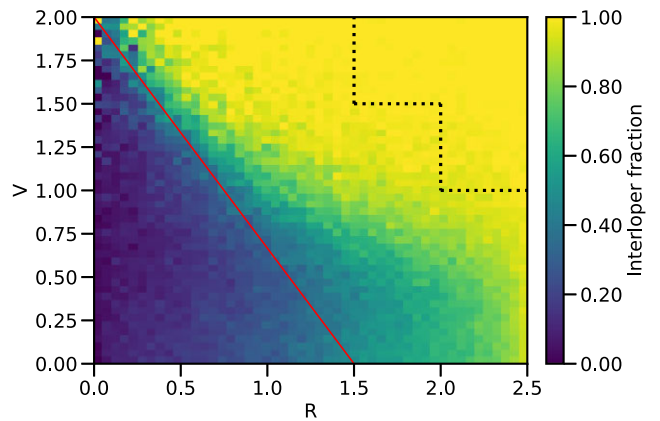

Figure 1. The fraction of interlopers in (R, V) projected phase space bins determined using the orbital library from Oman et al. (2021). R is distance from cluster centre, in units of $r_{\mathrm{vir,3D}}$ and V is the velocity offset from cluster centre, with units of σ_{3D}. The red diagonal line, $V = -(4/3)R + 2$, guides the eye as to where ~ 50 per cent of galaxies are interlopers, aside from a portion around $(R, V) \sim (1.2, 0.4)$, which has many galaxies at or close to their first apocentre after entering the cluster. The dotted line delineates a region in the upper right that is heavily dominated by the interloper population, which we use to compare between SDSS and UNIVERSEMACHINE interlopers.

are estimated using the ROCKSTAR halo finder; infall times and times of first pericentre are derived from halo merger trees generated by the CONSISTENT TREES utility. These merger trees include all haloes with >30 particles, corresponding to $M_{\mathrm{vir}} \gtrsim 4 \times 10^{10}$ M$_\odot$, allowing resolution of $M_{\mathrm{vir}} \sim 2.5 \times 10^{11}$ M$_\odot$ haloes, which host the lowest stellar mass galaxies in our observed sample – $M_\star \sim 10^{9.5}$ M$_\odot$ (at least until they are stripped of $\gtrsim 85$ per cent of their halo mass).

Satellites are identified as haloes within a 3D aperture of 2.5 r_{vir} at $z = 0$ for $\log(M_{\mathrm{vir}}/\mathrm{M}_\odot) > 12$ host systems (although we only use hosts with $\log(M_{\mathrm{vir}}/\mathrm{M}_\odot) > 14$ in this work). Satellite primary progenitors/descendants are traced backward/forward through time and orbits relative to the $z = 0$ primary progenitor/descendent of their host are recorded. The time of first pericentre is not interpolated (it is very challenging to do this properly and is unnecessary for our purposes), so it is worth noting here that this results in non-uniform output times, with a median time-step of 220 Myr and a maximum time-step of 380 Myr, sufficient to resolve the characteristic quenching and stripping time-scales fit in Oman et al. (2021). Worth noting is that the satellite halo mass, M_{sat}, is its maximum mass at $z \geq 0$, since maximum halo mass is better correlated with stellar mass in 'moderately stripped' satellites; see e.g. Conroy, Wechsler & Kravtsov (2006) and Appendix A of Wetzel et al. (2013) for further elaboration on this.

For the orbit library itself, since satellites are all galaxies within $2.5r_{\mathrm{vir}}$, interlopers are then naturally defined as all galaxies that fall within $2.5r_{\mathrm{vir}}$ in projection, but outside $2.5r_{\mathrm{vir}}$ in 3D. Only a vanishingly small fraction, $\ll 1$ per cent, of interlopers would have been classified as satellites at an earlier time. All satellites have a recorded time of first infall into the final cluster ($2.5r_{\mathrm{vir}}$ in 3D) and time of first pericentre (t_{fp}) in the final cluster; we only use t_{fp} in our modelling and analysis. We make use of their relative abundance compared to selected satellites to define the probability that an 'observed' galaxy is an interloper. For the sake of illustration and intuition for the reader, we plot the statistical fraction of galaxies that are interlopers at a given position in PPS, rather than cluster satellites, in Fig. 1. In the plot it's clear that at both high R and V the vast majority of galaxies are interlopers and vice-versa for lower

R and *V*. We note that around $(R, V) = (1.5 - 2, 0 - 0.5)$ there are many galaxies on their first infall into a cluster, whereas galaxies at around $(R, V) = (0 - 0.5, 1 - 2)$ are primarily galaxies at their first pericentre. Over several orbits galaxies settle to low (R, V). As well, we plot a straight line indicating where approximately 50 per cent of galaxies are interlopers – this should provide a helpful rule-of-thumb for the reader in the plots of various quantities in PPS that will follow.

2.3 Star formation histories: the UNIVERSEMACHINE

For individual galaxies' SFR histories we use the UNIVERSEMACHINE semi-analytic model (Behroozi et al. 2019) which parametrizes galaxy SFRs as a function of halo potential well depth, redshift, and assembly history. Their halo potentials and assembly history are derived from the *Bolshoi–Planck* dark matter simulation (Klypin et al. 2016; Rodríguez-Puebla et al. 2016), which features a periodic comoving volume $250h^{-1}$ Mpc on a side with 2048^3 ($\sim 8 \times 10^9$) particles. The simulation employs a flat ΛCDM cosmology compatible with Planck15 results (Planck Collaboration XIII 2016), with stored snapshots equally spaced in $\log(a)$ (180 intervals). The ROCKSTAR halo finder (Behroozi, Wechsler & Wu 2013a) and CONSISTENT TREES (Behroozi et al. 2013b) are used to construct merger trees, the same codes used in the construction of the merger trees for the orbit library employed in this work (Section 2.2).

For our purposes, it was important that their model reliably reproduce the observed SFRs and quiescent fractions over time, especially in overdense regions in SDSS (although not necessarily cluster cores, as we are only using UNIVERSEMACHINE interloper galaxies, described in the following paragraph). A key aspect of the UNIVERSEMACHINE model is that SFRs are stochastic – they fluctuate over time – and are linked to the merger history of the halo. Behroozi et al. (2019) found that without such scatter the quiescent fractions of satellites would be too high (relative to centrals). Specifically, higher SFRs are assigned to haloes with higher levels of halo growth, parametrized by the relative logarithmic growth in v_{max}, over the past dynamical time (where v_{max} is the maximum circular velocity of the halo). The stochasticity has two components: a short time-scale component (~ 10–100 Myr) representing internal processes affecting local galactic cold gas, and the second component linked to the dynamical time (~ 1 Gyr). This stochasticity will be a key difference in our comparison to other infall models (which use smooth star formation histories). A galaxy can transition between quiescent and star forming, as long as the overall quiescent fraction and other properties are matched to the various observations they used, in a self-consistent manner.

We construct a PPS catalogue using the UNIVERSEMACHINE simulation results at $z = 0$, assuming arbitrarily that the z-axis is the observational line-of-sight. For each of the 144 UNIVERSEMACHINE simulation sub-volumes, galaxies were selected from around central haloes ('UPID'$== -1$) with halo masses $\log(M_{\mathrm{vir}}/M_{\odot}) > 14$. All galaxies (including centrals) are selected so that their projected radius is $R < 2.5$ and projected velocity is $V < 2$, where the PPS coordinates (R, V) are defined as described previously in Section 2.1. True interlopers and satellites, as opposed to interlopers and satellites as defined in PPS, are identified with 3D cuts of $R > 2.5$ and $R < 2.5$, respectively.

2.4 SDSS data set

For the observed sample of galaxies, we build on the sample used by Oman et al. (2021) for their quenching analysis. As in Oman et al. (2021), we use the SDSS Data Release 7 catalogue (Abazajian

et al. 2009), supplemented with SFRs (Brinchmann et al. 2004; Salim et al. 2007) and stellar masses (Mendel et al. 2014), which were used in Oman et al. (2021) for estimating galaxy subhalo masses using subhalo abundance matching. We also use the spectroscopic sample of galaxies, which we note suffers from fiber collisions, although Oman et al. (2021) determine through a detailed exploration that the impact of this should be minimal.

We use the mass-weighted ages, as well as the accompanying stellar mass estimates, of Comparat et al. (2017), who performed full spectral fitting of galaxy properties using FIREFLY (Wilkinson et al. 2017). We compare their stellar masses with those of Mendel et al. (2014) and confirm that offsets in the stellar masses do not affect our results. In particular, we use the fits done using the stellar population models of Maraston & Strömbäck (2011), a Chabrier (2003) IMF, and the MILES stellar library (Sánchez-Blázquez et al. 2006; Beifiori et al. 2011; Falcón-Barroso et al. 2011) for the mass-weighted ages ('CHABRIER_MILES_age_massW'), and stellar masses ('CHABRIER_MILES_stellar_mass').

For simplicity, we focus on high-mass cluster haloes, $\log(M_{\mathrm{vir}}/M_{\odot}) > 14$, where environmental effects should be most extreme. As in Oman et al. (2021), we select candidate satellite galaxies around groups and clusters from the Lim et al. (2017) and Von Der Linden et al. (2007) catalogues, with cluster virial masses peaking at $\sim 3 \times 10^{13} M_{\odot}$ and $\sim 3 \times 10^{14} M_{\odot}$, respectively. We exclude groups with redshifts $z < 0.01$, as their members are generally too bright to be covered by SDSS spectroscopy. We note that the two group catalogues are constructed with different algorithms. Von Der Linden et al. (2007) look for overdensities of galaxies with similar colours, while Lim et al. (2017) use a friends-of-friends group finding algorithm. This does not significantly affect our analysis as we only use the cluster masses, centres, and line-of-sight velocities from the catalogues, rather than membership information. Satellite velocity offsets are normalized using cluster velocity dispersions, calculated from the virial masses of the clusters in the catalogues. The velocity dispersions are similar to the dark matter particle velocity dispersion of the systems (within $\lesssim 10$ per cent; see Munari et al. 2013, table 1); this is important since the velocity dispersion of dark matter particles in the host halo is used to normalize the velocity offsets of satellite haloes in the simulations that the Oman et al. (2021) orbit library is derived from (see Section 2.2). Host halo masses are calculated following equation (1) of Von Der Linden et al. (2007) and equation (4) of Lim et al. (2017). These are converted to virial masses assuming an NFW density profile (Navarro, Frenk & White 1997) and mean mass–concentration relation from Ludlow et al. (2014), with differences between assumed cosmologies accounted for. The mean enclosed density is then used to define the virial radii,

$$r_{\mathrm{vir}} = \left(\frac{3}{4\pi} \frac{M_{\mathrm{vir}}}{\Delta_{\mathrm{vir}}(z)\Omega_m(z)\rho_{\mathrm{crit}}(z)} \right)^{\frac{1}{3}}, \quad (3)$$

where $\Delta_{\mathrm{vir}}(z)$ is the virial overdensity in terms of the mean matter density $\Omega_m(z)\rho_{\mathrm{crit}}(z)$, with critical density $\rho_{\mathrm{crit}} = 3H^2/(8\pi G)$. Group velocity dispersions are defined following Biviano et al. (2006), but with redshift dependence from Bryan & Norman (1998); explicitly this is

$$\frac{\sigma_{\mathrm{1D}}}{\mathrm{km\ s}^{-1}} = \frac{0.0165}{\sqrt{3}} \left(\frac{M_{\mathrm{vir}}}{M_{\odot}} \right)^{\frac{1}{3}} \left(\frac{\Delta_{\mathrm{vir}}(z)}{\Delta_{\mathrm{vir}}(0)} \right)^{\frac{1}{6}} (1 + z)^{\frac{1}{2}}. \quad (4)$$

We use two stellar mass bins, $9 < \log(M_\star/M_{\odot}) < 10$ and $10 < \log(M_\star/M_{\odot}) < 10.5$. We note that for $\log(M_\star/M_{\odot}) > 10.5$ there are too few star forming galaxies to yield robust results, so we do not consider this mass range in this paper.
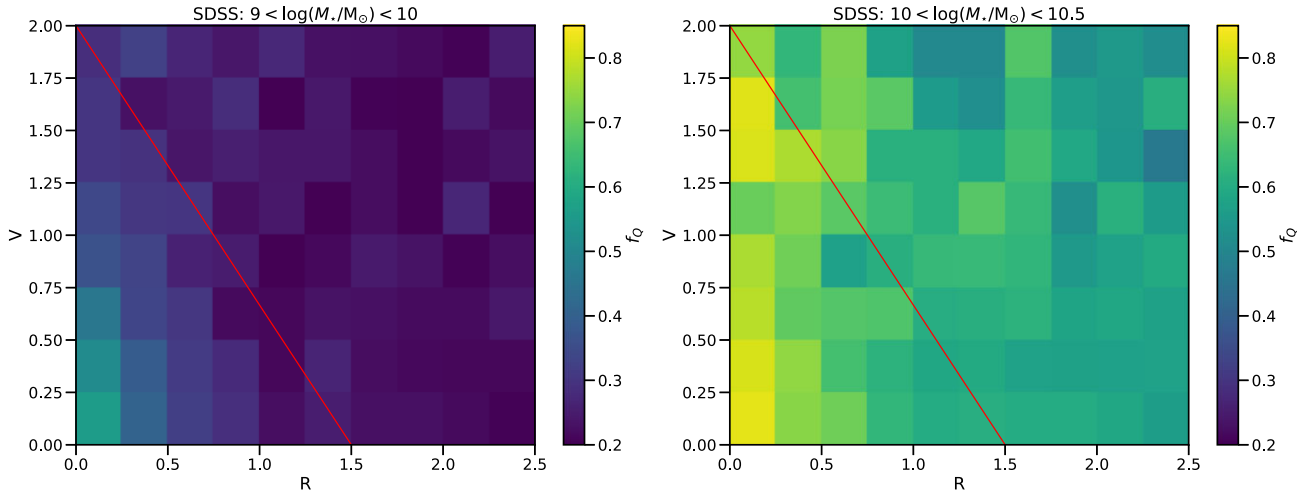
Figure 2. Quiescent fraction of SDSS galaxies in the stellar mass ranges $9 < \log(M_\star/M_\odot) < 10$ (left-hand panel) and $10 < \log(M_\star/M_\odot) < 10.5$ (right-hand panel) in clusters with host halo mass $\log(M_{vir}/M_\odot) > 14$, binned in PPS coordinates. The quiescent fraction is significantly higher in cluster centres than in their outskirts ($\gtrsim 30$ per cent). For the lower stellar mass bin, regions within the ∼50 per cent interloper (red diagonal) line, show at least modestly elevated quiescent fractions, with little evidence of enhanced quenching immediately outside of (above) this line. For the higher stellar mass bin, some higher velocity galaxies within the virial radius also show somewhat elevated quiescent fractions.

2.4.1 Quiescent fraction trends in PPS around SDSS clusters

Our first source of information for deriving quenching time-scales is the quiescent fraction in SDSS clusters. As we are using the same observational data sample as Oman & Hudson (2016) and Oman et al. (2021), we define star-forming and passive galaxies using the same cut, namely

$$\log\left(sSFR/yr^{-1}\right) = -0.4\log(M_\star/M_\odot) - 6.6, \quad (5)$$

with star-forming (quiescent) galaxies lying above (below) the specific star formation rate (sSFR) relation.

We illustrate the quiescent fraction trends in PPS for $9 < \log(M_\star/M_\odot) < 10$ galaxies in Fig. 2. In the figure we see well-known observed trends of decreasing f_Q with increasing distance from cluster centre as well as higher f_Q for galaxies with lower velocity offsets. We draw a line indicating roughly where ∼50 per cent of galaxies are interlopers, with galaxies above this line increasingly dominated by interlopers, moving from lower to higher R and V coordinates. Enhanced quenching largely falls within the region bounded by this 50 per cent interloper line in PPS. Contrasting with the field results of Oman et al. (2021, see upper right-hand panel of e.g. Fig. 10 to see star-forming fraction versus stellar mass for the SDSS field), with field defined as the average overall SDSS population, we see that the interloper quiescent fraction is higher than the field by 0.05 and 0.09 for our lower and higher stellar mass bins, respectively.

Having information on how quenching depends on both the radial position and velocity relative to host cluster centre gives us very useful information on when and where quenching occurs. It's clear in Fig. 2 that most quenching of infalling star forming galaxies occurs well within the virial radius. Modestly enhanced quenching up to high velocities indicates quenching is at least beginning to occur within the first infall or just after first pericentre. Although we have some useful information from f_Q in PPS, this information alone results in significant degeneracies between time of quenching onset and duration of quenching, as we will clearly show later in this work (see Section 3.2). To break this time-scale information degeneracy we must turn to another independent observable as a function of its position in PPS.

2.4.2 Mass-weighted age–stellar mass relation

For the stellar ages of SDSS galaxies we use a simple proxy to capture when the bulk of the stellar mass in a galaxy was formed, namely the mass-weighted age (MWA). We include only quiescent galaxies for our MWA-based measurements, since the emission lines of star-forming galaxies fill in the age-sensitive Balmer absorption lines, making them rather unreliable as an age indicator. In this subsection, we note a trend in MWA with stellar mass in both the observationally derived SDSS values and the UNIVERSEMACHINE simulated galaxies. As we are only interested in the differential change in MWA relative to an infalling population, we first describe how we control for the MWA–$M_\star$ effect in this work. After this, we will briefly illustrate observed MWA trends in PPS in Section 2.4.3.

We show the trend in MWA with stellar mass for both the observationally derived SDSS values and the UNIVERSEMACHINE (Behroozi et al. 2019) simulated galaxies side-by-side in Fig. 3. The mean and median trends that we see for SDSS and UNIVERSEMACHINE are qualitatively similar, but the UNIVERSEMACHINE galaxies are offset in overall MWA such that they are younger than SDSS MWAs by ∼3 Gyr on average. We note that mass-weighted ages derived from spectra depend on the assumed star formation history. For example, Allanson et al. (2009) showed that a range of star formation histories can fit the observed stellar absorption lines equally well in quiescent cluster galaxies, yielding mass-weighted ages that range from 5 Gyr (single burst) to 8 Gyr (exponentially declining star formation history) to 12.8 Gyr ('frosting' model) for the lowest mass cluster galaxies. Since we use ages only in a differential sense in our analysis, i.e. comparing the difference in MWA in the cluster core to the infalling regions in PPS, any systematic offset in the ages due to different star-formation history assumptions will not affect our analysis. For SDSS, the MWA–$M_\star$ relation is approximately flat, with MWA ∼ 9 Gyr for $9 < \log(M_\star/M_\odot) \lesssim 10.5$ galaxies, which then increases for $\log(M_\star/M_\odot) \gtrsim 10.5$ to ∼ 12–13 Gyr for the highest stellar mass galaxies, $\log(M_\star/M_\odot) > 11$. The increase in MWA for UNIVERSEMACHINE galaxies from lowest to highest stellar masses shown in Fig. 3 is similar in magnitude to that for SDSS (∼4 Gyr), albeit with a more steadily increasing monotonic trend,
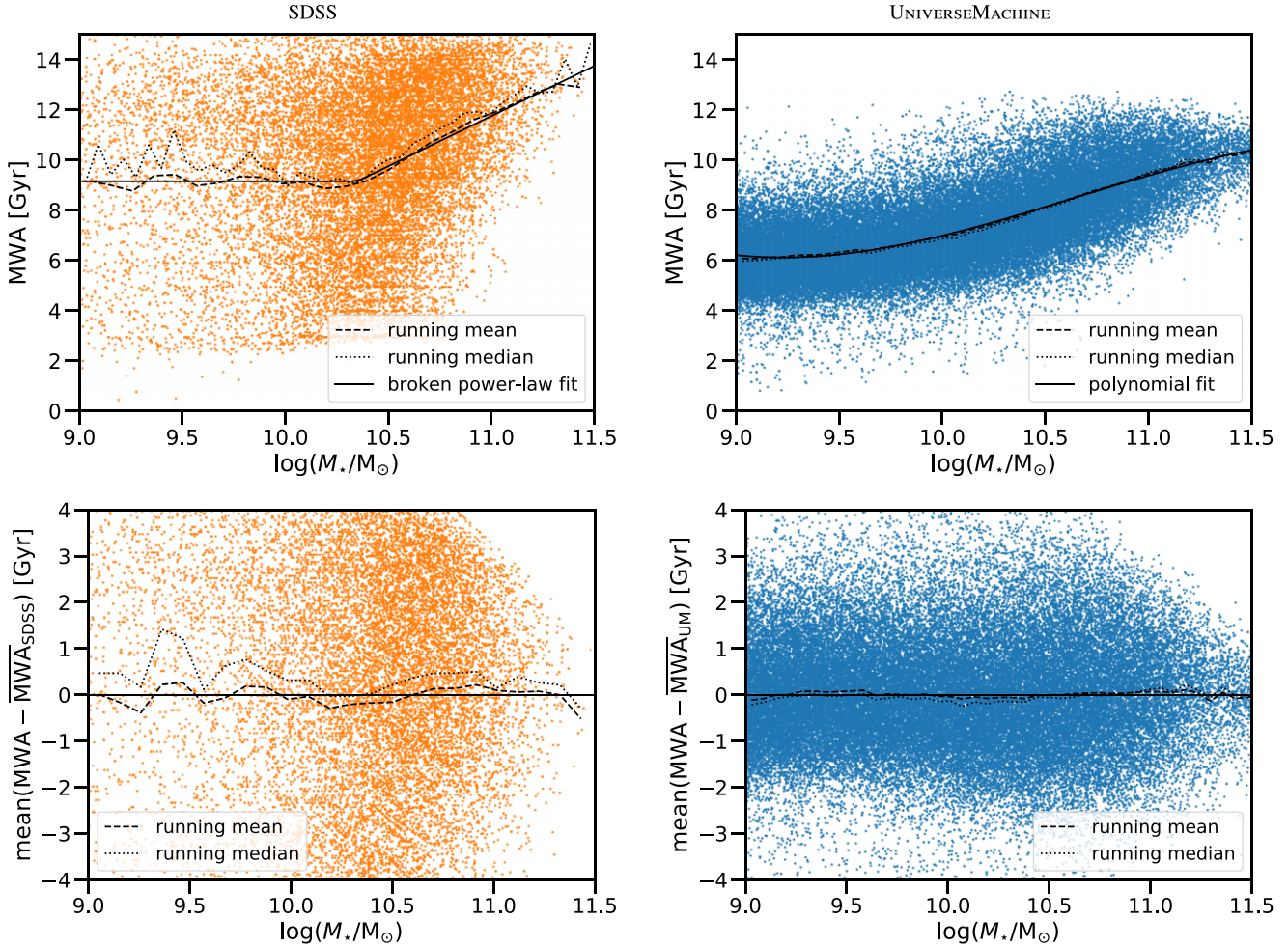
Figure 3. Top panels: SDSS observationally derived mass-weighted ages for quiescent galaxies (left-hand panel) exhibit a high degree of scatter and trend older with increasing galaxy stellar mass. UNIVERSEMACHINE simulated galaxies (right-hand panel) show a similar trend in mass-weighted age with stellar mass, albeit with a ~ 3 Gyr offset and significantly less scatter (see Section 2.4.2 for discussion). SDSS points indicate the best-fitting mass-weighted ages for SDSS galaxies from Wilkinson et al. (2017, stellar masses are from Comparat et al. 2017), with the running mean (median) shown as a dashed (dotted) line. The functional fits of the mean MWA–$M_\star$ relation used in the analysis of this work is fit to the running mean mass-weighted age at a given stellar mass and is shown with the solid black curves. Bottom panels: Residual MWA from the respective MWA–$M_\star$ functional fits for SDSS (left-hand panel) and UNIVERSEMACHINE (right-hand panel) is shown along with the running mean (dashed) and median (dotted) of the residual.

i.e. a non-flat trend in MWA below $\log (M_\star/M_\odot) < 10.5$. The scatter is also quite noticeably different – the SDSS scatter can be thought of as the natural distribution of galaxies (shown modelled here by the UNIVERSEMACHINE) convolved with the scatter induced in estimated MWAs from uncertainties in SED-fitting the galaxy spectra.

To remove the systematic trend in MWA with stellar mass and the ~ 3 Gyr offset between the SDSS observationally derived MWAs and the simulated UNIVERSEMACHINE MWAs so that we can focus solely on the effect of environment on galaxy MWAs, we look at the mean MWA residual from the mean MWA–$M_\star$ relation. This allows us to measure the differential effect between inner cluster regions and cluster outskirts (i.e. controlled for stellar mass). In particular, we use interloper galaxies as our reference infalling (already pre-processed) population.

For a simple clean proxy interloper sample that can be easily compared between UNIVERSEMACHINE and SDSS, we choose galaxies in the region of PPS given by

$$(1.5 < R < 2.5, 1.5 < V < 2) \cup (2 < R < 2.5, 1 < V < 1.5) \quad (6)$$

as illustrated in Fig. 1 (see also Section 3.1.1). Based on statistics from the orbit library of Oman et al. (2021), this region of PPS provides a sample that is made up of $\gtrsim 97$ per cent true interlopers. Correcting for the ~ 3 per cent impurity is unnecessary for our purposes.

For SDSS, we parametrize the mean MWA–$M_\star$ relation (fit and residual shown as black curve in the left-hand panels of Fig. 3) for (predominantly interloper) galaxies with $R < 2.5$ and $V < 2$ in the stellar mass range $10^9 < M_\star/M_\odot < 10^{12}$ using the function

$$\overline{\mathrm{MWA}}_{\mathrm{SDSS}} = A \log \left[\left(\frac{M_\star}{M_{\mathrm{crit}}} \right)^\alpha + \left(\frac{M_\star}{M_{\mathrm{crit}}} \right)^\beta \right] + b, \quad (7)$$

where $A = 0.1$ Gyr is the normalization, $b = 9.16$ Gyr is the vertical offset, $M_{\mathrm{crit}} = 10^{10.35} M_\odot$ is the location of the break in slopes, and $\alpha = 0$ and $\beta = 40$ specify the lower and higher stellar mass slopes, respectively. The fit is chosen to minimize the residual between the running mean MWA–$M_\star$ relation and the curve, and an offset is added so that the interloper population has mean$(\mathrm{MWA} - \overline{\mathrm{MWA}}_{\mathrm{SDSS}}) = 0$ by construction. This was done rather than fit the interloper MWA–
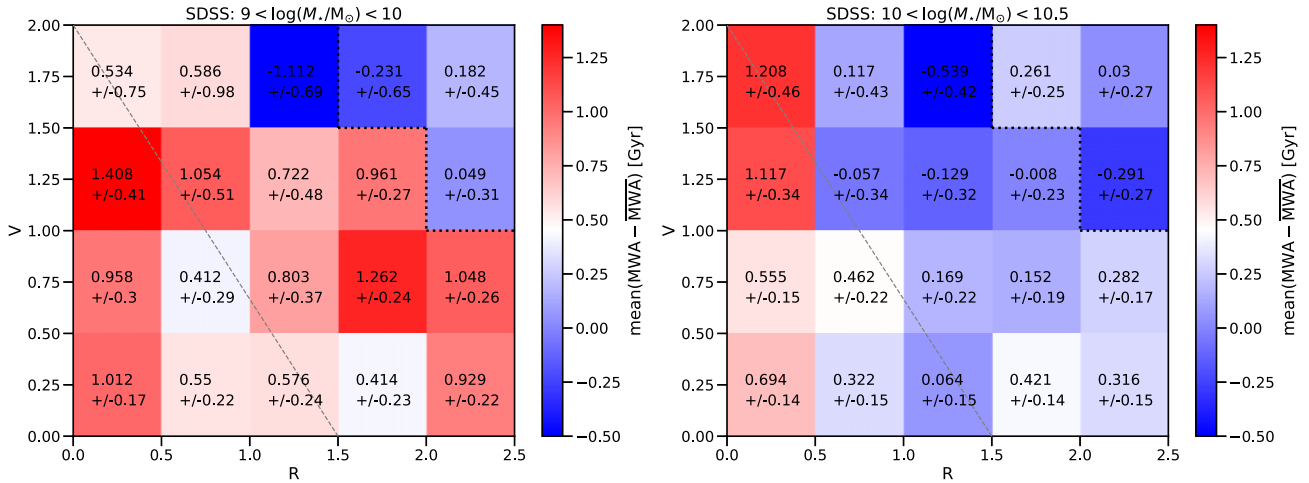
Figure 4. Mean deviation in MWA from the MWA–$M_\star$ relation of an SDSS interloper proxy sample (top right three bins; outlined by dotted black lines), mean(MWA − $\overline{\mathrm{MWA}}_{\mathrm{SDSS}}$), for quiescent $9 < \log(M_\star/\mathrm{M_\odot}) < 10$ (left-hand panel) and $10 < \log(M_\star/\mathrm{M_\odot}) < 10.5$ (right-hand panel) SDSS galaxies as a function of PPS coordinates. Errors on the mean values were calculated by bootstrapping over clusters, rather than individual galaxies. We note that the zero-point for this figure, defined by the three PPS bins in the top right-hand corner are 97 per cent interloper galaxies according to the orbit library of Oman et al. (2021). The dashed grey diagonal line indicates where approximately 50 per cent of galaxies are interlopers.

$M_\star$ relation directly since the interloper population had relatively few galaxies to give a good fit across the whole stellar mass range.

For the UNIVERSEMACHINE simulated sample, a polynomial is fit to the ∼pure interloper population (equation 6) to the running mean MWA (fit and residual shown as the black curve in the right-hand panels of Fig. 3). The polynomial fit has the following form:

$$\overline{\mathrm{MWA}}_{\mathrm{UM}} = -0.301 + 9.83x - 105x^2 + 370x^3; \tag{8}$$

where $x \equiv \log(M_\star/\mathrm{M_\odot})$ and the coefficients have units of Gyr.

2.4.3 *Trends in the mean deviation of mass-weighted age in PPS*

We present a binned map of the deviation in mass-weighted age from the mean MWA–$M_\star$ relation of the SDSS interloper proxy sample, mean(MWA − $\overline{\mathrm{MWA}}_{\mathrm{SDSS}}$) (see previous section), as a function of the PPS coordinates R and V in Fig. 4. Errors on the mean values are calculated by bootstrapping over clusters. The figure shows that mean(MWA − $\overline{\mathrm{MWA}}_{\mathrm{SDSS}}$) is enhanced for galaxies at lower radii and velocities. This is particularly clear for the higher stellar mass bin. Within the 50 per cent interloper line shown in the figure, we calculate mean(MWA − $\overline{\mathrm{MWA}}_{\mathrm{SDSS}}$) = $0.76^{+0.12}_{-0.14}$ and mean(MWA − $\overline{\mathrm{MWA}}_{\mathrm{SDSS}}$) = $0.66^{+0.08}_{-0.09}$ for the lower and higher stellar mass bins, respectively. Our higher stellar mass bin is consistent with Kim et al. (2022), who find that populations in PPS around clusters at $0.3 < z < 1.4$ identified as 'early infall' were older on average by 0.71 ± 0.4 Gyr than comparable 'recent infall' galaxies.

In the following section, we will show that the distribution of modelled mean(MWA − $\overline{\mathrm{MWA}}_{\mathrm{UM}}$), ΔMWA, in PPS differs from that of f_Q (shown previously in Fig. 2, Section 2.4.1). These differences in observed f_Q and ΔMWA distributions will provide us with a distinct set of preferred values for time of quenching onset and the duration of quenching time-scales.

3 MODELLING AND RESULTS

Our goal is to model the effects of infall quenching on two SDSS observables, f_Q and mean ΔMWA, in PPS around $z \sim 0$ galaxy clusters, and then use these results to constrain the associated

quenching time-scales. In this section, we first describe our infall quenching model in detail in Section 3.1. We show our model results for f_Q in Section 3.2 and for ΔMWA in Section 3.3. Finally, we combine our modelling results for both f_Q and ΔMWA to find a joint constraint on our model's two quenching time-scales in Section 3.4.

3.1 The model

To model the effects of infall quenching on both f_Q and mean ΔMWA, we use the UNIVERSEMACHINE semi-analytic model, as described in Section 2.3, as a representative model of star formation histories for galaxies in the infall region around a given cluster. A simple, smooth parametric model could be used, but such models do not have any intrinsic scatter in their star formation histories. We will see that this scatter has an impact on the bimodality in the star formation rates of the galaxy population (see Section 4.2). The UNIVERSEMACHINE simulation also provides us with a pre-processed infall sample.

Using a galaxy cluster PPS orbit library from Oman et al. (2021), as described in Section 2.2, we sample the infall time distribution of galaxies in PPS in galaxy clusters, including the interloper fraction across PPS. With these tools in hand, we describe in Section 3.1.1 how we construct an infall quenching model to predict f_Q and ΔMWAs as a function of position in PPS. For the sake of intuition, we visually illustrate some basic trends in f_Q and ΔMWA with our quenching parameters in Section 3.1.2.

3.1.1 *Description of our satellite quenching model*

We take the previously defined interloper sample as the infalling population, which will naturally include any pre-processing of galaxies in infalling structures (e.g. groups or filaments), as captured by the UNIVERSEMACHINE model. Our model assumes that star formation histories of $z = 0$ cluster satellite galaxies are the same as the $z = 0$ infalling population of galaxies, but with the addition of a quenching mechanism on infall, as described below. Galaxies that fall into the cluster begin suppression of their SFR when a time interval of t_{delay} has elapsed after their first pericentre passage. Defining the model this way means that we are modelling the mean quenching time-scales

for the population of galaxies as a whole, over all possible infall orbits (e.g. circular versus eccentric/plunging), giving us robust mean observed properties, insensitive to the choice between stochastic or smooth star formation histories (see Section 4.2.2 for discussion of this point). For UNIVERSEMACHINE galaxies, we use the same sSFR cut to separate quiescent/star forming galaxies as Behroozi et al. (2019), namely

$$\log\left(\text{sSFR/yr}^{-1}\right) = -11,\tag{9}$$

where star-forming (quiescent) galaxies are defined as lying above (below) the cut. The UNIVERSEMACHINE interloper population and SDSS interloper population differ slightly in their quiescent fractions (see e.g. Fig. A1), but we are only interested in the differential quenching effect (f_Q in the cluster versus some infall population). To find the time-scales associated with infall quenching, we will fit for the infalling f_Q. In Appendix A, we will confirm that our results are robust to adjusting the UNIVERSEMACHINE infalling star formation histories such that the infalling galaxies closely match the SDSS $f_Q - M_\star$ trend.

In the literature, some variation of a 'delayed-then-rapid' quenching scenario is common, like that proposed and explored in Wetzel et al. (2013). Exponentially declining SFR (e.g. Wetzel et al. 2013; Roberts et al. 2019) or linearly declining quiescent fraction (such as in Oman et al. 2021) are common assumptions. To facilitate comparison with previous work, we choose a model with an exponential suppression of the SFR on a time-scale, τ_{env}, that starts after some time delay, t_{delay}, relative to the time of first pericentre t_{fp} (we note that this allows for negative values of t_{delay}, i.e. quenching onset prior to t_{fp}).

Galaxies with a given stellar mass and time of first pericentre are Monte Carlo sampled from satellites in the orbital library of Oman et al. (2021). This gives, for each part of PPS, some mix of interloper galaxies, as well as in-cluster galaxies, which have a distribution of infall times depending on position in PPS. The star formation history for a random interloper with the closest stellar mass to that just drawn is then also drawn from UNIVERSEMACHINE, which gives the star formation history of this Monte Carlo-simulated galaxy. Each of these simulated satellite galaxies have quenching applied by starting the decline in the SFR at the time $t_{\text{fp}} + t_{\text{delay}}$, i.e. $\text{SFR}_{\text{sat}} = \text{SFR}_i q(t)$, where $q(t)$ is the multiplicative quenching envelope,

$$q(t; t_{\text{delay}}, \tau_{\text{env}}) = \begin{cases} 1, & t < t_{\text{fp}} + t_{\text{delay}}, \\ e^{-(t - t_{\text{delay}} - t_{\text{fp}})/\tau_{\text{env}}}, & t \geq t_{\text{fp}} + t_{\text{delay}}, \end{cases}\tag{10}$$

where i refers to a given Monte Carlo-sampled interloper (infalling galaxy) star formation history. In Fig. 5, we illustrate the effect of the multiplicative envelope on the stochastic star formation histories of the UNIVERSEMACHINE galaxies.

We then define our model of the quiescent fraction in a given PPS bin as

$$f_{Q,\text{model}}(t_{\text{delay}}, \tau_{\text{env}}, f_{Q,\text{infall}}) = \Delta f_Q(t_{\text{delay}}, \tau_{\text{env}}) + f_{Q,\text{infall}},\tag{11}$$

where Δf_Q is the increase in quenching from our model and $f_{Q,\text{infall}}$ is left as a free parameter of the infalling population's quiescent fraction to account for systematic differences between UNIVERSEMACHINE and the SDSS data.

To determine the stellar masses of any galaxy at $z = 0$, the UNIVERSEMACHINE SFRs are integrated numerically according to equation (21) and equation (22) in Behroozi et al. (2019). Explicitly, this gives a final stellar mass of

$$M_\star(t_{\text{now}}) = \int_0^{t_{\text{now}}} \frac{\mathrm{d}M}{\mathrm{d}t}\, \mathrm{d}t\tag{12}$$
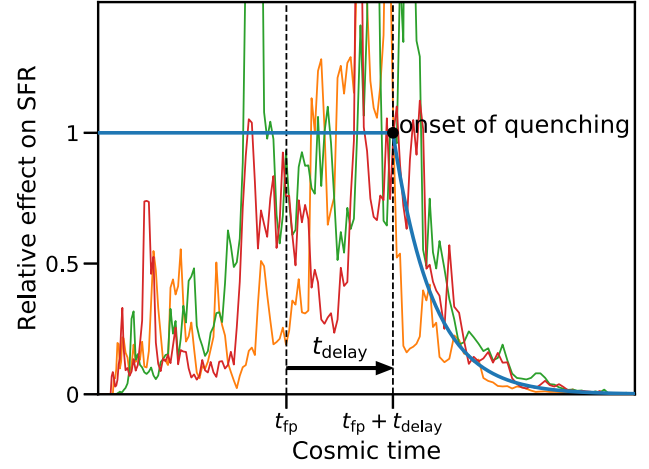
Figure 5. Illustration of the effect of our delayed-then-exponential envelope (thick blue line) on SFR, where at some time delay, t_{delay}, after time of first pericentre upon infall into a cluster, t_{fp}, SFR is quenched exponentially by a multiplicative envelope with time-scale, τ_{env}.

$$= \int_0^{t_{\text{now}}} \text{SFR}(t)(1 - f_{\text{loss}}(t_{\text{now}} - t))\, \mathrm{d}t,\tag{13}$$

where $f_{\text{loss}}(t) = 0.05\ln\left(1 + t/1.4\text{Myr}\right)$ and t_{now} is the age of the Universe at $z = 0$.

Integrated true (simulated) stellar masses are offset to give 'observed' stellar masses, according to the prescription in Behroozi et al. (2019, equation 25 and equation 27), namely by offsetting them by $\mu = \text{SM}_{\text{obs}} - \text{SM}_{\text{true}} = \mu_0 + \mu_a(1 - a)$ [dex] (where $a = 1/(1 + z)$ is the cosmological scale factor) and additionally adding Gaussian scatter of $\sigma_{\text{SM, obs}} = \min(\sigma_{\text{SM}, z}(1 + z), 0.3)$ [dex]. Best-fitting values associated with UNIVERSEMACHINE catalogues available online[2] were used for these stellar mass adjustments, namely $\mu_0 = 5.6 \times 10^{-3}$ and $\mu_a = -0.03$ used for the offset and $\sigma_{\text{SM}, z} = 0.069$ for the Gaussian scatter parameter, respectively.

Mass-weighted ages are integrated numerically from the stellar mass calculations, using the midpoints between the simulation snapshot time-steps and $\mathrm{d}M/\mathrm{d}t$ between two snapshots' time-steps,

$$\text{MWA} = t_{\text{now}} - \frac{1}{M_{\text{true}}} \int_0^{t_{\text{now}}} t\left(\frac{\mathrm{d}M}{\mathrm{d}t}\right) \mathrm{d}t,\tag{14}$$

where M_{true} is the total simulated stellar mass at the present time without any corrections for observational effects. Note that MWA is expressed as a lookback time.

The MWA predicted by our quenching model is then

$$\text{MWA}_Q = t_{\text{now}} - \frac{1}{M_{\text{true}}} \int_0^{t_{\text{now}}} t\left(\frac{\mathrm{d}M}{\mathrm{d}t}\right) q(t; t_{\text{delay}}, \tau_{\text{env}})\, \mathrm{d}t,\tag{15}$$

Similar to f_Q in equation (11) above, we define the ΔMWA observable of our model in a given PPS bin for each galaxy as

$$\Delta\text{MWA}(t_{\text{delay}}, \tau_{\text{env}}, \delta\text{MWA}_{\text{infall}}) = \text{MWA}_{Q,\text{model}}(t_{\text{delay}}, \tau_{\text{env}})\tag{16}$$

$$-\overline{\text{MWA}}_{\text{UM}}$$

$$+\delta\text{MWA}_{\text{infall}},\tag{17}$$

[2]UNIVERSEMACHINE data release 1 catalogues using the *Bolshoi-Planck* dark matter simulation, including catalogues for complete star formation histories, are available on the following page of Peter Behroozi's website: https://www.peterbehroozi.com/data.html.
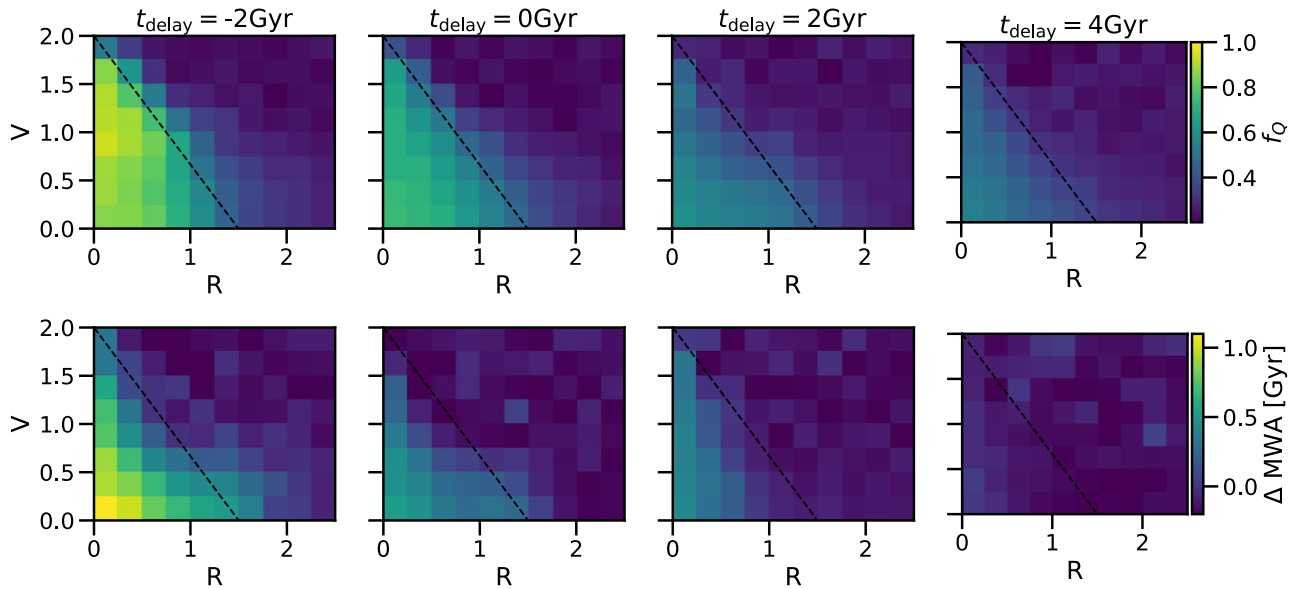
Figure 6. Delayed-then-rapid (instantaneous quenching, i.e. $\tau_{\rm env} = 0$) model predictions for a range of delay times, $t_{\rm delay}$, after time of first pericentre (a negative $t_{\rm delay}$ corresponds to quenching prior to $t_{\rm fp}$). For these illustrative plots, we use the UNIVERSEMACHINE 'observed' interloper proxy sample as defined in Section 2.4.2, for clusters with $\log(M_{\rm vir}/M_\odot) > 14$. Top row: predicted quiescent fractions, f_Q, in PPS. The diagonal dashed line, $V = -(4/3)R + 2$, marks the location where $\sim$50 per cent of galaxies are interlopers. Bottom row: quiescent mass-weighted age predictions in PPS.

where $\mathrm{MWA}_Q(t_{\rm delay}, \tau_{\rm env})$ is the MWA from our quenching model, $\delta\mathrm{MWA}_{\rm infall}$ is a floating offset to allow for the possibility that the infalling population's MWA may be different than that of the interloper phase space bins in equation (6), and $\overline{\mathrm{MWA}_{\rm UM}}$ is the MWA expected from the mean MWA–$M_\star$ relation as defined in equation (8) for UNIVERSEMACHINE galaxies. We use the mean $\Delta\mathrm{MWA}$ in a given PPS bin in the rest of this work.

3.1.2 Simple illustration of our quenching model

Now that we have described our infall quenching model, we illustrate our predictions for trends in f_Q and $\Delta\mathrm{MWA}$ in PPS. In particular, we start by exploring the effect of varying the quenching time delay relative to a satellite galaxy's first pericentre passage, $t_{\rm delay}$, assuming instantaneous quenching (i.e. $\tau_{\rm env} = 0$). We first illustrate this visually in Fig. 6 for $t_{\rm delay}$ ranging from $t_{\rm delay} = -2$ Gyr (i.e. prior to first pericentre) to $t_{\rm delay} = 4$ Gyr. We include contaminating interlopers.

For $t_{\rm delay} = -2$ Gyr, quiescent fractions are very high, $\gtrsim 0.75$ throughout most of the region dominated by satellites (i.e. below the $\sim$50 per cent interloper line, indicated by the same diagonal line as in Figs 2 and 4). $\Delta\mathrm{MWA}$ reaches $\sim$1.1 Gyr for $V < 0.5$ in the cluster core ($R < 0.5$). f_Q and $\Delta\mathrm{MWA}$ values increase with decreasing radius and also generally increase with decreasing velocity offset (albeit less strongly than with radius).

As $t_{\rm delay}$ increases, there is a decrease in both f_Q and $\Delta\mathrm{MWA}$, as intuitively expected. f_Q is still clearly enhanced relative to the interloper population within the $\sim$50 per cent interloper line, even for a $t_{\rm delay}$ as long as 4 Gyr. $\Delta\mathrm{MWA}$ reaches $\lesssim 0.25$ Gyr at all (R, V) bins for $t_{\rm delay} = 4$ Gyr. For an even longer $t_{\rm delay} \gtrsim 6$ Gyr (not shown), f_Q continues to drop and $\Delta\mathrm{MWA}$ drops to $\sim$0 Gyr.

With this physical intuition in mind, we now carry out a quantitative analysis of our model for f_Q and $\Delta\mathrm{MWA}$ in turn, as well as the joint constraint that observed SDSS f_Q and $\Delta\mathrm{MWA}$ values place on our model. We will see in the subsections that follow, that

differences in the trends for these two observables will enable us to break degeneracy in the $t_{\rm delay}$ and $\tau_{\rm env}$ model parameters.

3.2 Quiescent fraction: comparison of models and data

We plot f_Q model predictions as a function of time of quenching onset relative to the first pericentre, $t_{\rm delay}$, for a range of $\tau_{\rm env}$ values (different coloured curves) in Fig. 7 for $9 < \log(M_\star/M_\odot) < 10$ galaxies. For the equivalent plot for $10 < \log(M_\star/M_\odot) < 10.5$ galaxies at $z \sim 0$, see Fig. B1 in Appendix B. For reference, we show the mean f_Q (grey horizontal line) and the bootstrapped error on the mean (shaded grey region) for the SDSS sample in each PPS bin. When finding the best-fitting $t_{\rm delay}$ values (see Section 3.4 for this), we also fit for the infalling population's f_Q as a nuisance parameter, to account for possible systematic differences between the quenching from UNIVERSEMACHINE and the actual infall population, that we will then marginalize over. For the infalling galaxies, our fits prefer $f_{Q,\,\rm infall} = 0.16 \pm 0.01$ ($f_{Q,\,\rm infall} = 0.48 \pm 0.01$) for $9 < \log(M_\star/M_\odot) < 10$ ($10 < \log(M_\star/M_\odot) < 10.5$) galaxies, shown in Table 1, which are offset -0.06 ± 0.01 (0.09 ± 0.01) relative to the infall/interloper population's MWAs predicted by the UNIVERSEMACHINE model.

We now turn our attention to the trends in f_Q with $t_{\rm delay}$ and $\tau_{\rm env}$, by focusing on the region of PPS where the infall quenching effect in our model is most pronounced, namely the innermost parts of clusters. In e.g. the $R < 0.5$, $V < 0.5$ bin we see no effect on f_Q for very long quenching times ($t_{\rm delay} \sim 10$ Gyr), as expected. f_Q then steadily increases with decreasing time delay up until a point at which it begins decreasing again, at e.g. $t_{\rm delay} \sim -3$ Gyr. This turnover feature comes about because the star formation histories of infalling galaxies are being truncated so aggressively that many quenched galaxies are dropping out of our stellar mass range, with increasing numbers of quenched galaxies having $\log(M_\star/M_\odot) < 9$. Since there are fewer high stellar mass galaxies, truncated higher stellar mass galaxies dropping into the bin are not enough to compensate for those dropping below $\log(M_\star/M_\odot) = 9$. The effect is most pronounced for
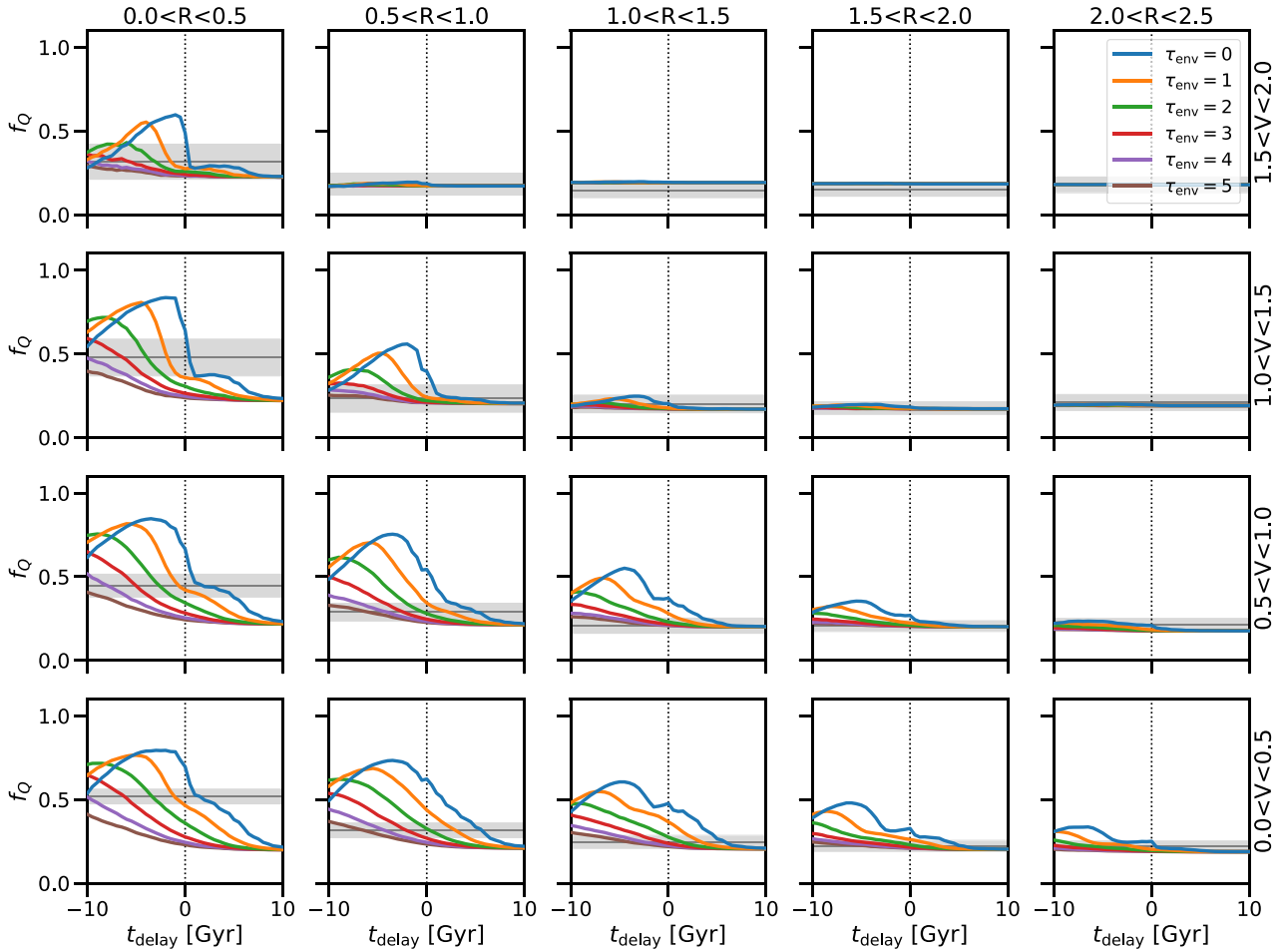
Figure 7. Quiescent fraction, f_Q, predictions for $9 < \log(M_\star/M_\odot) < 10$ galaxies for a range of models where galaxies quench after some delay time, t_{delay}, relative to the time of first pericentre. We show models for a range of exponential suppression time-scales. The SDSS mean f_Q values are shown as grey lines, with the shaded regions showing the bootstrapped (over clusters) uncertainty in the mean.

Table 1. Best-fitting quenching model parameters, given as the 50th percentile for each respective marginal posterior probability distribution. Upper and lower uncertainties are the 84th and 16th percentiles of the marginal distributions. Top rows: The time delay until the onset of quenching relative to the time of a galaxy's first pericentre (t_{delay}), and the exponential star formation suppression time-scale after the onset of quenching (τ_{env}). These are the best-fitting values indicated by the black diamond markers for t_{delay} and τ_{env} on the subplots of Fig. 9. Bottom rows: Best-fitting infalling population values, as described for f_Q in equation (11) and ΔMWA in equation (16). † indicates that the value is a 95th- percentile upper limit.

Parameter	$9 < \log(M_\star/M_\odot) < 10$	$10 < \log(M_\star/M_\odot) < 10.5$
t_{delay}	$3.5^{+0.6}_{-0.9}$ Gyr	$-0.3^{+0.8}_{-1.0}$ Gyr
τ_{env}	≤ 1.0 Gyr†	$2.3^{+0.5}_{-0.4}$ Gyr
$f_{Q, infall}$	0.16 ± 0.01	0.48 ± 0.01
δMWA$_{infall}$	0.59 ± 0.12	0.47 ± 0.07

instantaneous quenching, $\tau_{env} = 0$, and less pronounced for longer star formation suppression time-scales, $\tau_{env} > 0$; τ_{env} acts to smooth out dependencies of f_Q on t_{delay}. This delay in quenching due to longer τ_{env} values also results in an offset between the maximum possible f_Q as a function of t_{delay}. This offset effect is such that for

higher values of τ_{env} the maximum possible f_Q (the turnover point) requires much earlier quenching, significantly before the time of first pericentre. Features on shorter time-scales (i.e. bumps and wiggles) shown in our model curves are due to the discrete snapshot times of the *N*-body simulation used for the orbit library in Oman et al. (2021).

We contrast this with the methodology of Oman et al. (2021), where stellar masses were not truncated by quenching, since for simplicity they did not make use of star formation histories. In that case, there is no turnover effect in f_Q as a function of t_{delay} since galaxies are not dropping out of the stellar mass range. That said, the f_Q data do not prefer excessively negative values of t_{delay} and so the shift in preferred delay time compared to our choice of truncating stellar masses is rather minimal: it simply causes an offset of ~ 0.5 Gyr (earlier) relative to our model.

In PPS, the models reproduce the general trends in f_Q visible in the SDSS data, but there is clearly significant degeneracy in the preferred model (i.e. which t_{delay}, τ_{env} combination is preferred). In order to break this degeneracy, we need another observable beyond the quiescent fraction to constrain t_{delay} and τ_{env}. As mentioned earlier, we will use observationally derived stellar ages of galaxies for this purpose. We explore model predictions for stellar ages of galaxies in PPS in the next section.
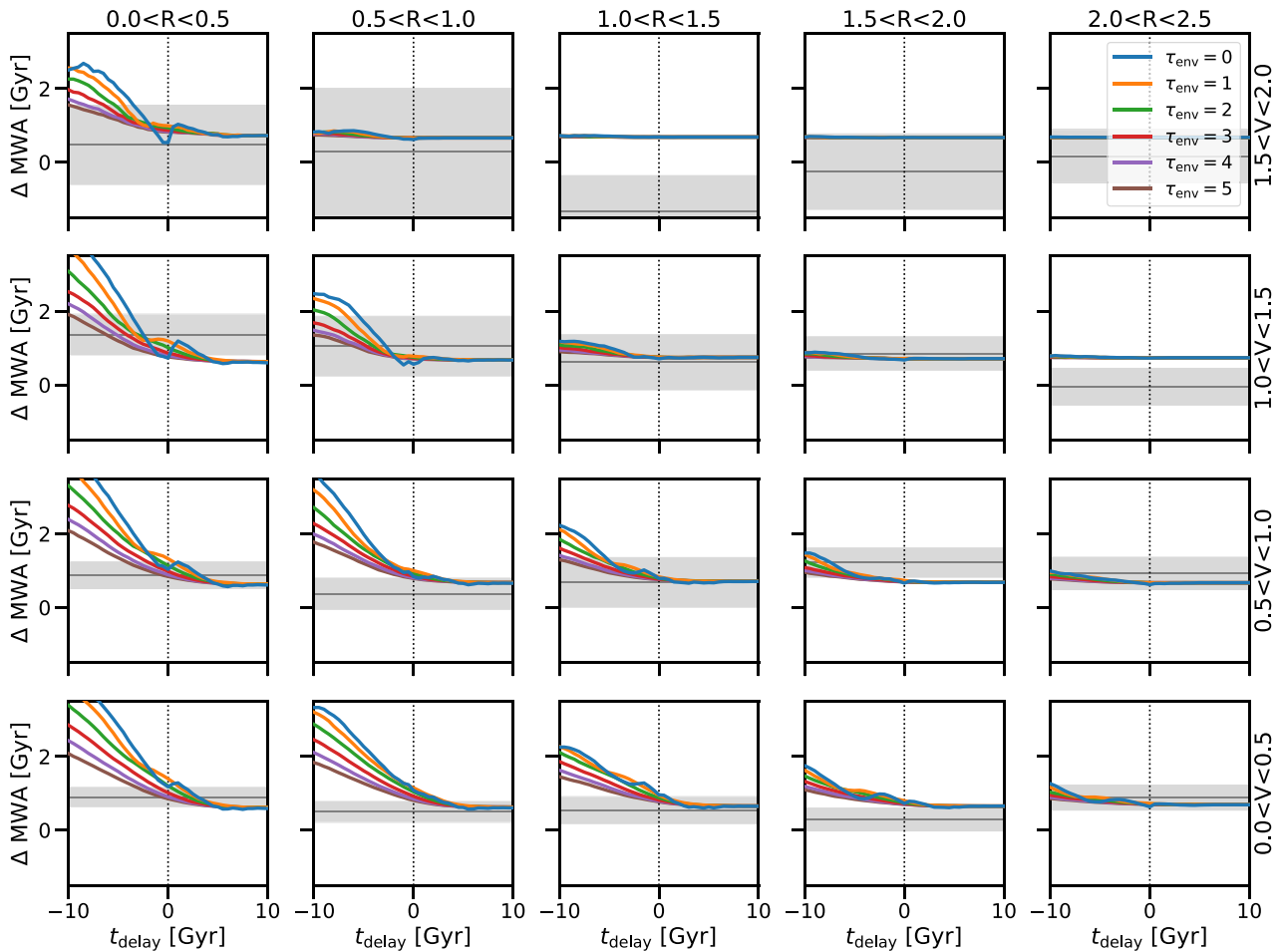
Figure 8. Mean ΔMWA predictions for quiescent $9 < \log(M_\star/M_\odot) < 10$ galaxies for a range of models where galaxies quench after some delay time, relative to time of first pericentre. We show models run for a range of exponential suppression time-scales. The SDSS mean ΔMWA values are shown as grey lines, with the shaded regions showing the bootstrapped (over clusters) uncertainty in the mean. Note that our models include an additional offset to allow for the possibility that the infalling population may be different than that of the interloper phase space bins in equation (6).

3.3 Stellar ages: comparison of models and data

We now compare our models of stellar age to the SDSS data. As in the previous section, in Fig. 8, we plot the mean ΔMWAs as a function of $t_{\rm delay}$ in bins of PPS coordinates R and V for $9 < \log(M_\star/M_\odot) <$ 10. For the equivalent plot for $10 < \log(M_\star/M_\odot) < 10.5$ galaxies, see Fig. B1 in Appendix B. We note that for the infalling galaxies, we find that our fits prefer offsets of $\delta{\rm MWA}_{\rm infall} = 0.59 \pm 0.12$ Gyr and $\delta{\rm MWA}_{\rm infall} = 0.47 \pm 0.07$ Gyr for $9 < \log(M_\star/M_\odot) < 10$ and $10 < \log(M_\star/M_\odot) < 10.5$ galaxies (shown in Table 1), which correspond to infalling MWAs of $6.70^{+0.11}_{-0.13}$ Gyr and 7.63 ± 0.07 Gyr.

As in the previous section, looking at e.g. the innermost, low-velocity bin ($R < 0.5$, $V < 0.5$), we see no effect on ΔMWA for very long quenching delay times, with the effect on ΔMWA usually increasing with decreasing $t_{\rm delay}$ (earlier onset of quenching). This general trend, for the most part, is similar for different values of $\tau_{\rm env}$, but with shallower slopes for higher values of $\tau_{\rm env}$. Or, put another way, a longer time-scale for the suppression of star formation (a higher value of $\tau_{\rm env}$) would require an earlier onset of quenching to have the same degree of impact on the stellar age. Interestingly, the trend in higher ΔMWA for earlier onset of quenching relative to time of first pericentre is not monotonic for all values of $\tau_{\rm env}$: for example, for $\tau_{\rm env} \sim 0$ (instantaneous or almost instantaneous suppression of star formation) around $t_{\rm fp}$ that there is a small ~0.1−0.3 Gyr dip. This

dip is most pronounced for the higher velocity bins, as this is where galaxies reaching their first pericentre will preferentially be located in PPS – close to the cluster core and moving at a high velocity. Since the figure is only showing the mean ΔMWA of quiescent galaxies, and this region of PPS is dominated by galaxies on their first infall, the influx of recently quenched galaxies lowers the mean ΔMWA relative to values of $t_{\rm delay}$ slightly earlier/later than $t_{\rm delay} \sim -1$ Gyr (the deepest part of the dip). Longer suppression time-scales blur out this effect. Since the lower velocity regions in the cluster core ($R < 0.5$) are dominated by galaxies that entered the cluster a long time ago, this effect is washed out.

Although the uncertainties on ΔMWA are larger than for f_Q, our model still predicts clear trends for ΔMWA across PPS. Because the ΔMWA trends with the time-scale parameters differ from those obtained from f_Q, we explore the constraints that arise from combining f_Q and ΔMWA in the next section.

3.4 Fitting and joint constraints of time delay and exponential quenching time-scale

By running our model across a range of $t_{\rm delay}$ and $\tau_{\rm env}$ values we can jointly constrain these time-scale parameters by using the information contained in the trends of f_Q and ΔMWA as a function of
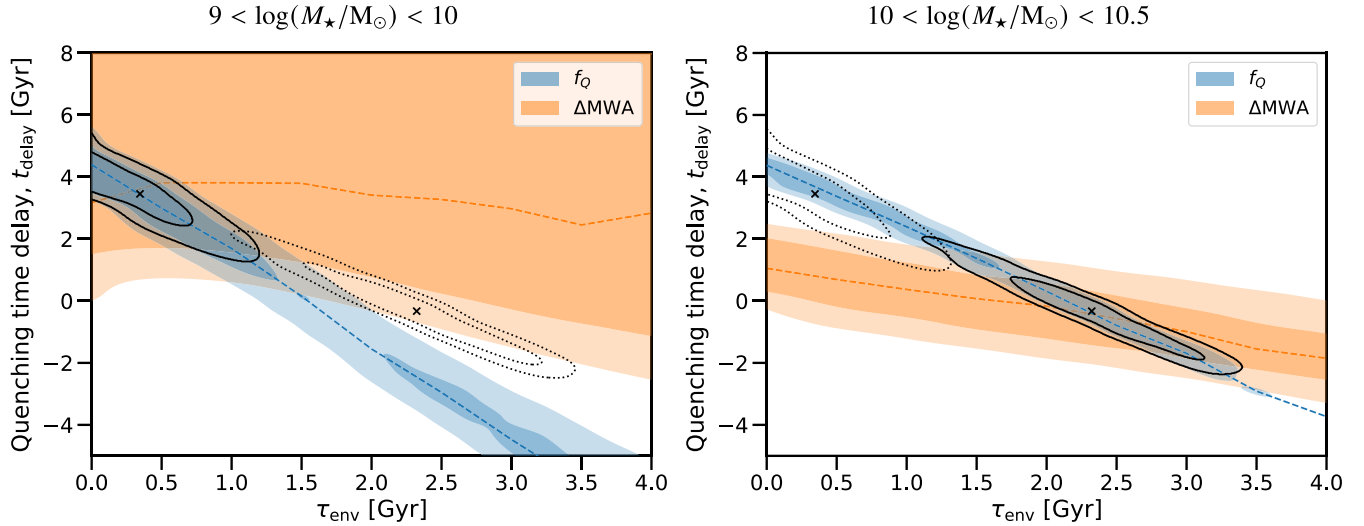
Figure 9. Marginal best-fitting parameter values on our model's quenching time-scales for the two observables, f_Q (blue), mean ΔMWA for quiescent galaxies (orange), as well as joint (black lines/grey shading) for $9 < \log(M_\star/M_\odot) < 10$ galaxies (left-hand panel) and $10 < \log(M_\star/M_\odot) < 10.5$ (right-hand panel). Uncertainties on the best-fitting parameters are shown with corresponding shaded regions (darker and lighter shaded regions corresponding to the 68 per cent and 95 per cent confidence regions, respectively). The parameters fit here are the quenching delay time relative to time of first pericentre, t_{delay}, and the exponential suppression of star formation time-scale, τ_{env}. The dashed lines indicate the best-fitting t_{delay} at a given τ. The joint 68 per cent and 95 per cent confidence region from both observables for t_{delay} and τ_{env} are shown overlaid with black contours. The joint best-fitting model, i.e. the location with the peak joint probability, is indicated with a cross symbol. For each plot, the other stellar mass bin's joint constraint is overlaid (black cross and dotted contours). Marginal best-fitting values and uncertainties from the joint constraints are tabulated in Table 1.

the PPS coordinates. For each τ_{env} we perform χ^2 fitting to determine the best-fitting t_{delay} value, by minimizing

$$\chi^2 = \sum_i \left(\frac{y_{obs,i} - y_{model,i}}{\sigma_i} \right)^2. \tag{18}$$

$y_{obs,i}$ and σ_i are the mean observed SDSS data value and its uncertainty, respectively, for either f_Q or MWA, in a given PPS bin i. $y_{model,i}$ is the modelled value, as defined in equation (11) and equation (16), respectively, and is a function of t_{delay}, τ_{env}, and the infall population. The sum is over all PPS bins. We find that the uncertainties on the data may be underestimated, since for the $20 - 3 = 17$ degrees of freedom (20 PPS bins, three free parameters) we find the following reduced χ^2 values: $\chi^2_{red}(f_Q) \approx 2.75$ and $\chi^2_{red}(\Delta MWA) \approx 2.47$ for $9 < \log(M_\star/M_\odot) < 10$. Similarly, for $10 < \log(M_\star/M_\odot) < 10.5$ we find $\chi^2_{red}(f_Q) \approx 1.38$ and $\chi^2_{red}(\Delta MWA) \approx 1.25$. A high χ^2 leads to parameter constraints that may be too tight, so to be conservative we inflate the uncertainties on the observed SDSS mean f_Q and mean ΔMWA in each PPS bin by a factor of $\sqrt{\chi^2_{red}}$. We note that uncertainties on the models are negligible as they can be made arbitrarily small simply by sampling more UNIVERSEMACHINE galaxies.

We present our best-fitting values for t_{delay} as a function of τ_{env} for our observables f_Q and ΔMWA in Fig. 9, for both $9 < \log(M_\star/M_\odot) < 10$ (left-hand panel) and $10 < \log(M_\star/M_\odot) < 10.5$ (right-hand panel) stellar mass bins. Consistent with the PPS results presented in Section 3.3, ΔMWA provides a constraint on t_{delay} that depends only weakly on τ_{env}. For f_Q there is a steeper trend between the fit t_{delay} for a given τ_{env}. These relations make sense intuitively: as f_Q depends on the distributions of SFRs at a given moment in time, it can be greatly impacted by any change in the time-scale of SFR suppression, τ_{env}. ΔMWA, on the other hand, will not change significantly with τ_{env} since the bulk of stellar mass growth occurred in the past and will mainly depend on the time at which a galaxy has quenched.

Also shown in Fig. 9 are contours showing the confidence intervals for t_{delay} and τ_{env} jointly when combining f_Q and ΔMWA. We list our best-fitting joint values with uncertainties (16th and 84th percentiles) defined by the t_{delay} and τ_{env}'s marginal probability distributions in Table 1.

Looking first at the $9 < \log(M_\star/M_\odot) < 10$ bin in Fig. 9 (left-hand panel), quenching time delay relative to time of first pericentre is preferred to be $t_{delay} = 3.5^{+0.6}_{-0.9}$ Gyr. We find a relatively rapid suppression of star formation once quenching has started, with a best-fitting $\tau_{env} \leq 1.0$ Gyr (95 per cent confidence level used as an upper bound). We choose to use a one-sided upper limit for τ_{env} for the lower stellar mass bin, as the probability distribution for τ_{env} is not Gaussian, peaking at $\tau_{env} \sim 0$ and the 50th-percentile occurring at $\tau_{env} = 0.3$ Gyr.

Turning to the higher stellar mass bin, $10 < \log(M_\star/M_\odot) < 10.5$ (bottom panel), a different fit is preferred: an earlier onset of quenching, $t_{delay} = -0.3^{+0.8}_{-1.0}$ Gyr, corresponding to the onset of quenching beginning close to the time of first pericentre. The corresponding best-fitting τ_{env} is $2.3^{+0.5}_{-0.4}$ Gyr, indicating a significantly slower suppression of star formation.

4 DISCUSSION

We now discuss the meaning of our modelling results and how they contrast with previous literature. For a discussion of the robustness of our results to changes in the overall UNIVERSEMACHINE star formation histories, see Appendix A.

We first compare with the work that we have directly built upon, namely that of Oman et al. (2021) (see also: Oman et al. 2013; Oman & Hudson 2016). We then compare our results to the alternative framework proposed in the seminal work of Wetzel et al. (2013), followed by other observational studies: Taranu et al. (2014) which made use of stellar age-related spectral indices and Rhee et al. (2020) which used PPS information. Finally, we compare to a study

Table 2. Comparison of time-scales from various literature results to our work; in each table section they are listed in the order in which they appear in Section 4. 'Lower' and 'higher' stellar mass bins refer to $9 < \log(M_\star/M_\odot) < 10$ and $10 < \log(M_\star/M_\odot) < 10.5$, respectively. Upper section: average time t_Q until quenching, relative to time of first pericentre. Lower section: exponential time-scales τ compared to our exponential suppression best-fitting time-scale as well as with the conditional assumption (∗) that quenching begins at time of first pericentre ($t_{\text{delay}} \equiv 0$), using the f_Q contour only. † indicates use of the upper 95 per cent confidence level as an upper bound. Note that definitions of τ in the literature vary; the respective discussion sections should be consulted when making comparisons between τ values.

Study	Quantity	Lower $M_\star$ bin	Higher $M_\star$ bin
This work (best)	t_Q	**4.3 ± 0.4 Gyr**	**3.9 ± 0.2 Gyr**
This work ($t_{\text{delay}} \equiv 0$)∗	t_Q	**3.7 ± 0.4 Gyr**	**4.0 ± 0.2 Gyr**
Oman + 2021	t_Q	3.7 ± 0.2 Gyr	3.4 ± 1.0 Gyr
Wetzel + 2013	t_Q	2.4 Gyr	1.5 Gyr
Taranu + 2014	t_Q	–	5.0 Gyr
Rhee + 2020	t_Q	4.2 Gyr	2.5 Gyr
Wright + 2022	t_Q	1.1 Gyr	2.3 Gyr
This work (best)	τ_{env}	**≤ 1.0 Gyr†**	**$2.3^{+0.5}_{-0.4}$ Gyr**
This work ($t_{\text{delay}} \equiv 0$)∗	τ_{env}	**1.6 ± 0.2 Gyr**	**2.2 ± 0.1 Gyr**
Wetzel + 2013	$\tau_{\text{Q, fade}}$	0.8 ± 0.2 Gyr	0.5 ± 0.2 Gyr
Taranu + 2014	τ_{post}	–	3 Gyr
Rhee + 2020	τ_{cluster}	$1.7^{+0.2}_{-0.3}$ Gyr	$1.1^{+0.3}_{-0.2}$ Gyr
Wright + 2022	τ	0.4 Gyr	0.9 Gyr

examining the predictions of hydrodynamical simulations (Lotz et al. 2019; Wright et al. 2022).

For the purposes of our discussion, we define a quantity, t_Q, as the average time for a star-forming galaxy with median sSFR (of a star-forming galaxy) to exponentially decline in SFR until it crosses the UNIVERSEMACHINE quenching threshold of 10^{-11} yr^{-1}. For our lower (higher) stellar mass bin this difference in sSFR is ∼1.0 dex (∼0.8 dex). Assuming a constant stellar mass for simplicity, the quenching time-scale is $t_Q = t_{\text{delay}} + 2.33\tau_{\text{env}}$ ($t_Q = t_{\text{delay}} + 1.82\tau_{\text{env}}$). These time-scales are summarized in Table 2 and their contours are also plotted in Fig. 10.

4.1 Oman et al. (2021)

The satellite quenching model of Oman et al. (2021), based on modelling quiescent fractions in PPS, employs a maximum-likelihood model constraining four parameters: the quiescent fraction of an infalling population (f_{before}), the final quiescent fraction of galaxies after satellite quenching has taken place (f_{after}, set to zero for their core analysis), the time at which half of the drop in quiescent fraction is complete (t_{mid}, relative to time of first pericentre), and the time-scale to go from the initial to final quiescent fraction (Δt). For the purpose of consistency across the following discussion sections, we will refer to the quenching time as t_Q rather than t_{mid}.

We've previously discussed how the preferred time of quenching in Oman et al. (2021) is offset to be ≲ 0.5 Gyr later than ours, since for simplicity they assumed galaxy star formation histories are not truncated by premature quenching.

They find a nearly flat t_Q relation with stellar mass (excluding $\log(M_\star/M_\odot) \gtrsim 10.5$), with $t_Q = 3.7 \pm 0.2$ Gyr and $t_Q = 3.4 \pm 1.0$ Gyr for our lower and higher stellar mass bins, respectively. We find consistent overall quenching times ($t_Q(t_{\text{delay}} = 0) = 3.7 \pm 0.4$ Gyr and $t_Q(t_{\text{delay}} = 0) = 3.9 \pm 0.2$ Gyr, respectively). Relaxing this assumption, our model is still consistent with theirs, although the lower stellar mass bin t_Q increases slightly, by ∼0.5 Gyr.

Oman et al. (2021) were unable to provide strong constraints on their Δt parameter; they find that it tends to be underestimated, e.g. with the median value (based on the probability density from the model) underestimated by up to ∼1 Gyr at lower stellar masses. They note that this effect is due to resolution issues in the N-body simulation resulting in low-mass satellite haloes being disrupted too early. Values of $\Delta t \sim 1 - 1.5$ Gyr were preferred by their modelling, but with very large uncertainties, which they marginalized over for their primary analysis to get their tight constraints on t_Q and the quiescent fraction parameters. They note that the resolution effect could mask a decreasing trend in t_Q, which could likewise occur in our modelling.

4.2 Wetzel et al. (2013)

Wetzel et al. (2013) found that satellites falling into a cluster experience 'delayed-then-rapid' quenching. One difference between their model and ours is that although Wetzel et al. (2013) fits a delayed-then-rapid model with an exponentially declining sSFR (specifically, their parameters are: time delay, $t_{\text{Q, delay}}$, and 'fading time-scale', τ_{fade}), our model relies on the star formation histories of UNIVERSEMACHINE which are stochastic rather than a simple smooth analytical function.

Another important difference in the models is the treatment of pre-processing of galaxies prior to entering the cluster. In our model, a simulated infalling population is used to account for pre-processing, whereas in Wetzel et al. (2013) the infall quenching 'clock' for a given galaxy starts on first infall into any larger halo, i.e. when a central becomes a satellite. Because of this, one would expect their (average) quenching time delay to be longer, as their quenching time delay includes the time spent in smaller groups prior to infalling into the main cluster progenitor. The magnitude of the median difference between time of first infall into any halo and infall into the final cluster is 3.2 Gyr (their fig. 2; assuming $\log M_{200c/M_\odot}(z = 0) = 14.2$ – our median host halo mass).

4.2.1 Quenching delay time

To compare their model with ours, we must account for the two populations of galaxies falling into clusters which will have different quenching times relative to time of first pericentre in the $z \sim 0$ cluster: those that are already satellites in some smaller halo which have had their quenching clock start at some earlier time (A) and those that are falling into a larger halo for the first time (B). For case (A), according to the fits of Wetzel et al. (2013) for a $9 < \log(M_\star/M_\odot) < 10$ galaxy, the full time until quenching for a galaxy that became a satellite in a $\log(M_{\text{vir}}/M_\odot) > 14$ cluster is ∼5.7 Gyr (interpolated). Taking the median time between first infall and infall into the cluster to be ∼3.2 Gyr (see fig. 2 of Wetzel et al. 2013) and time from r_{vir} to first pericentre in the final $\log(M_{\text{vir}}/M_\odot) > 14$ cluster to be ∼1 Gyr (Oman et al. 2013), we would expect such galaxies to quench $\bar{t} \sim 1.5$ Gyr after first pericentre on average. We note that these pre-processed galaxies are quenching earlier than they would in a scenario where quenching occurs only due to infall into the final cluster host. For case (B), according to their fits for galaxies falling into $\log(M_{\text{vir}}/M_\odot) > 14$ clusters, the full time until quenching is $t_Q \sim 4.8$ Gyr for $9 < \log(M_\star/M_\odot) < 10$ galaxies, or about ∼3.8 Gyr after first pericentre. The relative proportions of the (A) and (B) scenarios are ∼0.6 and ∼0.4, respectively (see Wetzel et al. 2013, fig. 1).

These approximations give an average quenching time, relative to time of first pericentre, of $t_Q \sim 0.6(1.5$ Gyr$) + 0.4(3.8$ Gyr$) \sim$

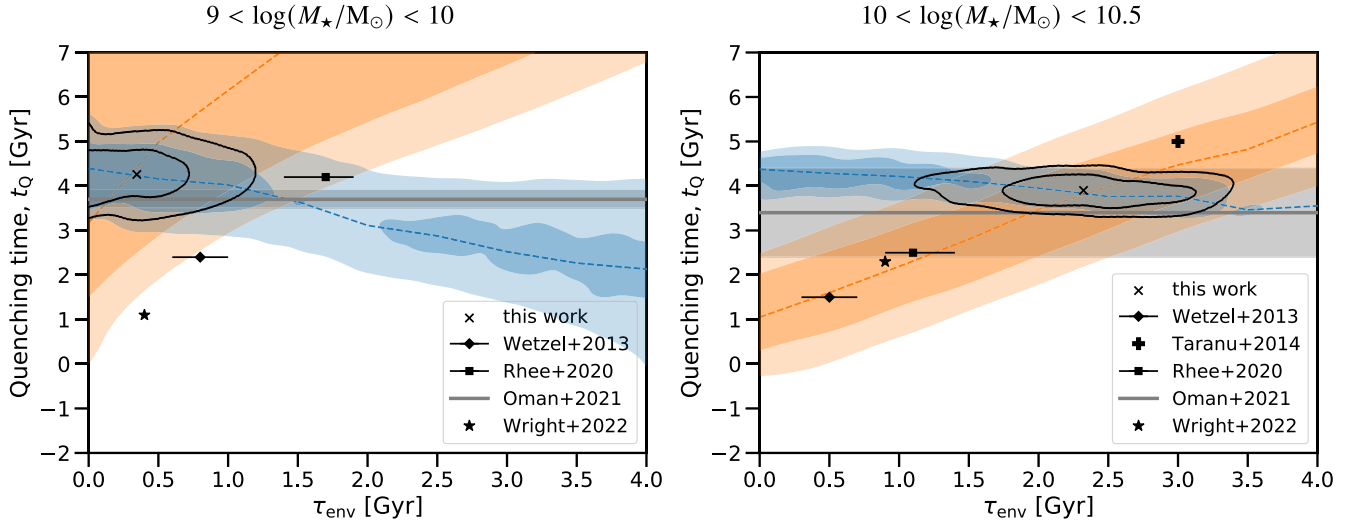**Figure 10.** Marginal and joint best-fitting parameter values on overall quenching time, t_Q, and τ_{env} for the two observables, f_Q (blue) and quiescent ΔMWA (orange) for $9 < \log(M_\star/M_\odot) < 10$ galaxies (left-hand panel) and $10 < \log(M_\star/M_\odot) < 10.5$ (right-hand panel), analogous to Fig. 9. As well, we have overlaid the t_Q literature values from Table 2 with the various points and lines shown in the legend. We omit error bars where it was not straightforward to determine them.

2.4 Gyr. The same calculation for our higher stellar mass bin gives $t_Q \sim 1.5$ Gyr. These are shorter, by ~1.9 Gyr and ~2.4 Gyr, than our quenching times of $t_Q = 4.3 \pm 0.4$ Gyr and $t_Q = 3.9 \pm 0.2$ Gyr, for our lower and higher stellar mass bins, respectively.

4.2.2 Fading time and sSFR bimodality

Wetzel et al. (2013) claim that in order for the bimodality in the sSFR distribution to persist across environments, including in the most massive clusters, infall-induced quenching requires a short exponential time-scale, $\tau_{Q,\,fade} < 1$ Gyr (they find $\tau_{Q,\,fade} = 0.8 \pm 0.2$ Gyr and $\tau_{Q,\,fade} = 0.5 \pm 0.2$ Gyr for our lower and higher stellar mass bins, respectively). For longer SFR fading time-scales (~2 times their best-fitting τ_{fade}), their model predicts no bimodality in the sSFR distribution, but rather a unimodal distribution peaking near the centre of the 'green valley' ($\log(sSFR) \sim -11.2$ for $\log(M_\star/M_\odot) < 10.5$ galaxies), which we illustrate in Fig. 11. This occurs because their model assumes a smooth exponentially declining star formation history for the infalling galaxies, and a gentle fading leads to galaxies passing through the green valley too slowly to maintain the bimodality.

We find our lower stellar mass bin exponential time-scale ($\tau_{env} \leq 1.0$ Gyr) is consistent with theirs, but our higher stellar mass bin value of $\tau_{env} = 2.3^{+0.5}_{-0.4}$ Gyr is significantly longer than their $\tau_{Q,\,fade} \sim 0.5 \pm 0.2$ Gyr. However, these fading times are different quantities: $\tau_{Q,\,fade}$ in Wetzel et al. (2013) represents the fading time of an individual galaxy, whereas our τ_{env} is a population-wide fading 'envelope' which multiplies the stochastic star formation histories of individual UNIVERSEMACHINE galaxies. The stochasticity in star formation histories is due to significant variations in SFR of UNIVERSEMACHINE galaxies, which occur by design: in this model, SFR tracks the accretion rate of baryonic matter, as well having a random variability on short time-scales ($\sim10-100$ Myr). This stochasticity allows the sSFR distribution to be bimodal even in the case of a long τ_{env}. Because of the rapid changes in SFR for UNIVERSEMACHINE galaxies, they do not spend significant amounts of time in the green valley, whether or not their SFRs are forced to decline by our exponential suppression envelope (visually illustrated
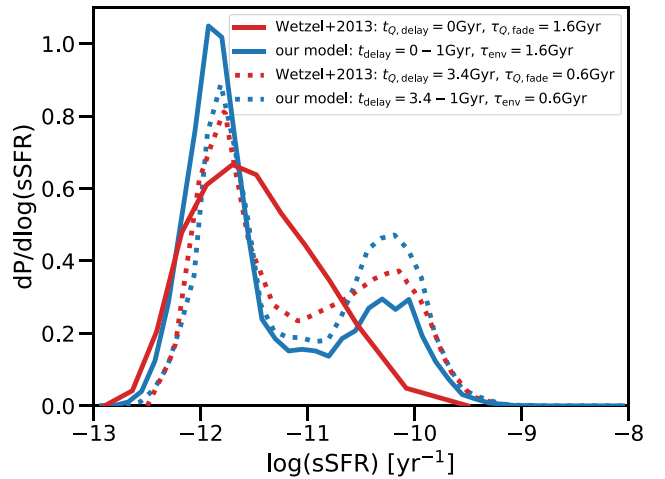

Figure 11. Our model displays a persistent bimodality in sSFR distribution, regardless of SFR suppression time-scale, whereas Wetzel et al. (2013) finds a unimodal distribution for long suppression times (contrast the red solid line for their model with the blue solid line for our model). This arises because of our use of a stochastic star formation history, whereas they use smooth star formation histories. In particular, we are contrasting with fig. 9 in Wetzel et al. (2013), for our $10 < \log(M_\star/M_\odot) < 10.5$ stellar mass bin and their $10.1 < \log(M_\star/M_\odot) < 10.5$ stellar mass bin. A correction of -1 Gyr is added to be comparable to their average infall time, which has the time zero-point set at r_{vir}. Additionally, for ease of comparison with Wetzel et al. (2013), who claimed need for a short exponential suppression time-scale to reproduce the bimodality, we include only galaxies within $r_{200c} \approx 0.73 r_{360m}$ and add log-normal scatter with mean $\log(sSFR/yr^{-1}) = -12$ and 0.25 dex variance to all galaxies with $\log(sSFR/yr^{-1}) < -12$.

in Fig. 5). The SFRs of infalling galaxies undergoing satellite quenching drop below the sSFR quenched threshold much more abruptly than a suppressed smooth star formation history, so longer fading/suppression envelopes are allowed.

From a qualitative comparison of the scenarios in Fig. 12, a longer $\tau_{env} \gtrsim 2.3$ Gyr is preferred by the depth of the green valley for $9 < \log(M_\star/M_\odot) < 10$ galaxies, rather than a short $\tau_{env} \leq 1.0$ Gyr
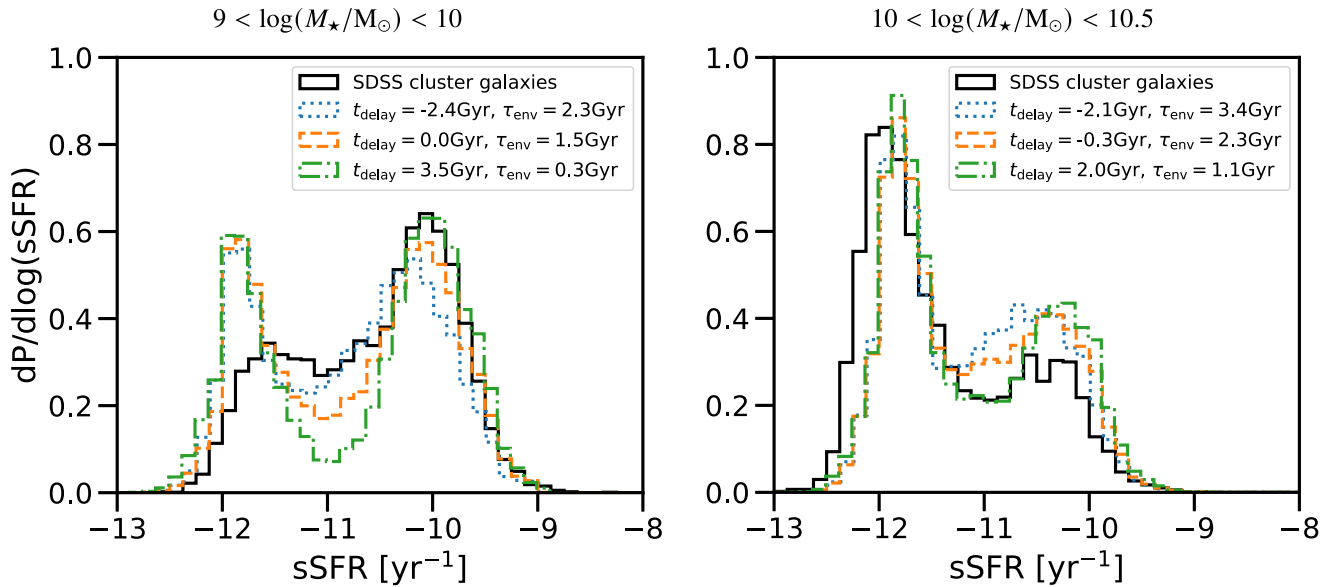
Figure 12. Similar to Fig. 11, except we now demonstrate how we can reproduce the bimodality in the sSFR distributions of the SDSS data (solid black histogram) for a range of relevant models with different exponential quenching time-scales (coloured lines) for our two stellar mass bins, $9 < \log(M_\star/M_\odot) < 10$ (left-hand panel) and $10 < \log(M_\star/M_\odot) < 10.5$ (right-hand panel). The left-hand panel shows three scenarios, which are (in the order that they are listed in the legend): (i) long τ equal to that preferred for our higher stellar mass bin, but t_{delay} taken from along the f_Q error contour, (ii) a scenario where $t_{delay} = 0$, with τ_{env} chosen to minimize χ^2 (see Fig. 9), (iii) 50th percentile of our time-scale parameters. The right-hand panel is similar, showing: (i) rightward end of the 95 per cent confidence level ellipse, (ii) best-fitting time-scale parameters, (iii) leftward end of the 95 per cent confidence level ellipse. It is worth noting that there is an offset in f_Q not factored into the above model histograms, equal to that described in Section 3.2, and that the quenching sSFR cut for SDSS has some stellar mass dependence, rather than a simple sSFR $= 10^{-11}$ yr^{-1} cut.

preferred by the f_Q and ΔMWA constraints. We note that the location of the star forming peak of the sSFR distribution is better fit by e.g. $t_{delay} = 0$ Gyr and $\tau_{env} \sim 1.5$ Gyr. For $10 < \log(M_\star/M_\odot) < 10.5$ galaxies, we find that the relative depth of the green valley implies a preference for longer τ_{env}, e.g. $\tau_{env} \gtrsim 2.3$ Gyr, which is our best-fitting result.

A robust quantitative comparison of infall quenching models using additional sSFR distribution features, aside from f_Q, clearly requires controlling for type of star formation history. The depth of the green valley depends on whether the assumed star formation histories are smooth or stochastic and the specific choice of stochasticity made in UNIVERSEMACHINE is not unique. Additionally, measured star formation rates of quiescent galaxies have high systematic uncertainties regardless of how they are obtained (Brinchmann et al. 2004; Salim et al. 2007; Wetzel et al. 2012; Hayward et al. 2014), meaning that much of the shape of the quiescent bump is determined by observational systematics, rather than quenching physics. As such, we do not utilize additional features of the sSFR distribution beyond f_Q, as they do not appear informative or robust in constraining our infall quenching model and comparing it to e.g. Wetzel et al. (2013).

4.3 Taranu et al. (2014)

Using an approach broadly similar to our work, Taranu et al. (2014) combined a star formation and quenching model with information from a library of subhalo orbits from *N*-body simulations of 4 rich clusters with $\log(M_{200c}/M_\odot) > 15$, with median stellar masses in their sample of $\log(M_\star/M_\odot) = 10.4$. They compared model predictions to observed bulge and disc colours and are one of the only previous studies (see also Upadhyay et al. 2021, for an attempt with a spectroscopic sample of 11 galaxies) to make use of age-sensitive stellar absorption line-strength indices for constraining

cluster infall quenching times using radial information. Since galaxy disc colours are particularly affected by environment, they can be used to constrain the quenching time-scale and location of quenching onset in the cluster using their median trends in cluster-centric radius. They used the age-sensitive Balmer lines of luminous passive cluster galaxies, which are more sensitive to older stellar populations than disc colours.

From their set of models, they found that quenching starting shortly before the time of first pericentre (within a quenching radius of $0.5r_{200c}$, which is $\sim$0.5 Gyr prior to the time of first pericentre for a galaxy on first infall) produced the best fits to disc colours as a function of radius. They found that delayed-then-rapid models lead to an excessively large slope in galaxy disc colour versus cluster-centric radius, and also overpredict the strength of the Balmer lines. Instead, their preferred model is one with a gentle exponential quenching time-scale of $\tau_{post} \sim 3$ Gyr, consistent with the τ_{env} that we find for the higher stellar mass bin. This corresponds to a total quenching time of $t_Q \sim 5$ Gyr, longer than ours by $\sim$0.7 Gyr (see Table 2).

4.4 Rhee et al. (2020)

Rhee et al. (2020) fit a quenching model to disc galaxies in $z = 0.08$ clusters. Infall is defined as when a galaxy first crosses $1.5R_{vir}$. They follow an approach similar to ours, parametrizing quenching with a time delay followed by an exponential suppression of star formation, but instead of modelling the quiescent fraction they model the full SFR distribution of galaxies as a function of position in PPS.

They find that the time delay from infall until the onset of quenching is 2 Gyr for all stellar masses $\log(M_\star/M_\odot) > 9.5$. The time for a galaxy to fall from their infall definition of $1.5R_{vir}$ to first pericentre is $\sim$1.3 Gyr. We can then subtract this from their quenching time since infall (as presented as t_Q in their table 1

for similar stellar mass bins), giving $5.45\,\mathrm{Gyr} - 1.3\,\mathrm{Gyr} \sim 4.2\,\mathrm{Gyr}$ and $\sim 3.8\,\mathrm{Gyr(interpolated)} - 1.3\,\mathrm{Gyr} = 2.5\,\mathrm{Gyr}$ for our lower and higher stellar mass bins, respectively. This is consistent with our lower stellar mass bin (we find $t_Q \sim 4.3 \pm 0.4\,\mathrm{Gyr}$) but earlier for our higher stellar mass bin by $\sim 1.4\,\mathrm{Gyr}$ (our $t_Q \sim 3.9 \pm 0.2\,\mathrm{Gyr}$).

This difference in trend with stellar mass between our results and those of Rhee et al. (2020) is not due to a difference in delay times, but rather is due to their SFR suppression time-scale, $\tau_{\mathrm{cluster}}(z = 0)$, which declines with increasing stellar mass over our stellar mass range, whereas we find the opposite trend. We note that they add a redshift-dependent factor to their quenching time-scale $\tau_{\mathrm{cluster}}(z_{\mathrm{inf}}) = \tau_{\mathrm{cluster},\,0}(1 + z)^{-\alpha}$ and to the quenching time delay time-scale (due to the redshift evolution of the dynamical time in clusters), $t_{\mathrm{d}}(z_{\mathrm{inf}}) = t_{\mathrm{d},\,0}(1 + z_{\mathrm{inf}})^{-1.5}$. This only results in a small correction (a decrease by ~ 10 per cent if we use e.g. $z = 0.5$, the time at which an average galaxy has fallen into the cluster). For $z = 0$, they find $\tau_{\mathrm{cluster},0} = 1.7^{+0.2}_{-0.3}\,\mathrm{Gyr}$ and $\tau_{\mathrm{cluster},0} \sim 1.1^{+0.3}_{-0.2}\,\mathrm{Gyr}$ for our lower and higher stellar mass bins, respectively. This is consistent with our lower stellar mass bin's $\tau_{\mathrm{env}}(t_{\mathrm{delay}} = 0) = 1.6 \pm 0.2\,\mathrm{Gyr}$, but more rapid than our higher stellar mass bin's $\tau_{\mathrm{env}}(t_{\mathrm{delay}} = 0) = 2.2 \pm 0.1\,\mathrm{Gyr}$, respectively.

4.5 Comparison with hydrodynamical simulations

A study directly comparable to ours, which examines quenching in the EAGLE hydrodynamical simulations, is that of Wright et al. (2022). Their work provides an orbital analysis of galaxies' gas inflow, stripping, star formation, and quenching. Stripping of infalling galaxies' hot gas begins at 2–3 virial radii from the host and takes longer for high-mass satellites ($\log (M_\star/\mathrm{M}_\odot) > 10$). This begins the process of starvation and the removal of the hot gas 'buffer', resulting in galaxies becoming vulnerable to cold gas stripping. They refer to the hot gas halo as having a protective effect, as they observe that the onset of significant cold gas stripping only begins after stripping of the hot gas halo is complete. In their work, they include both HI and molecular hydrogen in the mass of cold gas in a galaxy.

For low-mass satellites ($\log (M_\star/\mathrm{M}_\odot) < 10$), they find that suppression of gas cooling onto the galaxy becomes permanent after hot gas stripping, normally occurring around the time of first pericentre. Some high-mass satellites, on the other hand, retain small hot gas reservoirs, and continue to cool gas for star-formation after first pericentre. Cold gas stripping is shown to be periodic, being strongest for galaxies near pericentre, as that is when density and velocity is at a maximum, hence maximizing the ram-pressure force ($P_{\mathrm{ram}} \propto \rho_{\mathrm{ICM}} v^2$). All of this results in the following: low mass satellites experience very efficient ram-pressure stripping of cold gas, leading to rapid quenching, whereas high mass satellites experience less efficient stripping and a more gradual starvation-like scenario after their first pericentre.

In terms of quenching, we note that Wright et al. (2022) remove pre-processed infalling satellites, namely those galaxies which were satellites in a host halo with $\log (M_{200}/\mathrm{M}_\odot) \geq 12$ prior to falling into the current host (≈ 30 per cent of infalling galaxies). This handling of pre-processing is somewhat different from ours, as we fit the pre-processed quiescent fraction for infalling galaxies and focus on the differential quenching (relative to the infalling population) in the final $z = 0$ cluster.

To compare our quenching time-scales with Wright et al. (2022), we compare the time required for a galaxy of average sSFR (see their fig. 7) to cross the quenching threshold and assume $t_{\mathrm{delay}} = 0$ to find an approximate average quenching time relative to the time of first pericentre. Wright et al. (2022) finds median quenching times $0.25\langle T_{\mathrm{orb}}\rangle = 0.25(4.5\,\mathrm{Gyr}) = 1.1\,\mathrm{Gyr}$ and $0.65\langle T_{\mathrm{orb}}\rangle =$

$0.65(3.5\,\mathrm{Gyr}) = 2.3\,\mathrm{Gyr}$ relative to the time of first pericentre, for their $9 < \log (M_\star/\mathrm{M}_\odot) < 10$ and $10 < \log (M_\star/\mathrm{M}_\odot) < 11$ bins, respectively. For the higher stellar mass galaxies, the majority (~ 80 per cent) are quenched by second pericentre. We find longer quenching times of $t_Q = 3.7 \pm 0.4\,\mathrm{Gyr}$ for $9 < \log (M_\star/\mathrm{M}_\odot) < 10$ galaxies and $t_Q = 4.0 \pm 0.2\,\mathrm{Gyr}$ for $10 < \log (M_\star/\mathrm{M}_\odot) < 10.5$ galaxies. Using similar reasoning, their SFR suppression is equivalent to an exponential suppression time of $\tau \sim 0.4\,\mathrm{Gyr}$ and $\sim 0.9\,\mathrm{Gyr}$ for our lower and higher stellar mass bins, respectively. Taking $t_{\mathrm{delay}} = 0$, our τ_{env} is significantly longer, by 1.2 and 1.3 Gyr, respectively. Relaxing $t_{\mathrm{delay}} = 0$, we find little to no change in preferred t_Q, but results in our preferred exponential suppression time-scale being consistent with theirs for lower stellar mass galaxies.

Based on this discussion, we conclude that we prefer significantly longer total quenching time-scales than Wright et al. (2022). If we assume $t_{\mathrm{delay}} = 0$, as their models predict, then we find that our star formation suppression time-scales, τ_{env}, are longer than theirs by 1.2–1.3 Gyr.

A similar previous study is that of Lotz et al. (2019), which instead examined $z \sim 0$ quiescent fractions in the Magneticum Pathfinder hydrodynamical simulation. They found that most $\log (M_\star/\mathrm{M}_\odot) < 10.5$ galaxies are quenched within $\sim 1\,\mathrm{Gyr}$ of crossing r_{200c} (i.e. around the time of first pericentre), with the relatively small fraction of galaxies with tangential orbits and very high stellar masses able to maintain star formation after first pericentre. This quenching is significantly earlier than we find for both of our stellar mass bins, and all of the results examined in our discussion above.

4.6 Towards a consistent model of quenching

There is ample observational evidence that ram pressure stripping of the cold gas starts at or just before first pericentre. For example, Smith et al. (2010) found that ram-pressure stripped tails of Coma cluster galaxies were prevalent within half of the cluster virial radius and that most of these tails pointed away from the cluster, indicating that the stripping was occurring on infall, i.e. just before pericentre. Studies of the distribution of HI abundance as a function of the PPS coordinates (Jaffé et al. 2015; Oman et al. 2021) find that HIdepletion begins close to pericentre, i.e within $0.5r_{200c}$, in agreement with Smith et al. (2010). However, quenching is not instantaneous and likely proceeds from outside inwards if it is due to ram-pressure stripping (Boselli, Fossati & Sun 2022). Owers et al. (2019) studied cluster galaxies with spatially resolved spectroscopy and uncovered a population of galaxies with strong Hδ absorption – indicative of recent quenching – in their outskirts, but these same galaxies had ongoing star formation in their centres. They modelled the distribution of these galaxies in PPS, finding indications that galaxies with a recent quenching event in their outskirts are within 1 Gyr of entering within $0.5r_{200c}$ of the cluster centre.

That quenching should start at (or just before) pericentre is supported by our results for our higher stellar mass bin ($t_{\mathrm{delay}} = -0.3\pm^{+0.8}_{-1.0}\,\mathrm{Gyr}$) but is at odds with our results for low stellar mass galaxies. For these we find that the onset of quenching occurs well past pericentre $t_{\mathrm{delay}} = 3.5\pm^{+0.6}_{-0.9}\,\mathrm{Gyr}$ and with a fast quenching envelope ($\tau_{\mathrm{env}} \leq 1.0\,\mathrm{Gyr}$). Such a short quenching time-scale, however, would predict a deep 'green valley' in the sSFR distribution, inconsistent with the observed shallow depth (see Fig. 12). For the lower stellar mass bin, quenching starting at pericentre ($t_{\mathrm{delay}} = 0$) is permitted by the quiescent fraction but disfavoured by the stellar ages at the $\gtrsim 2\sigma$ level (see Fig. 9). This preference is driven by there being little-to-no gradient in ΔMWA between galaxies in the cluster core and those infalling, as shown in Fig. 4. This result, however, appears

to be driven partly by the very lowest stellar mass galaxies: if we restrict analysis to $9.5 < \log(M_\star/M_\odot) < 10$, we find somewhat better compatibility with $t_{\text{delay}} = 0$. If we fix $t_{\text{delay}} = 0$, as suggested by other observational and theoretical evidence, then, for the lower stellar mass bin, we find slower quenching with $\tau_{\text{env}} = 1.6 \pm 0.2$ Gyr. This time-scale is in better agreement with the sSFR distribution in Fig. 12. For the higher stellar mass sample, the quenching time-scale is longer, with $\tau_{\text{env}} = 2.2 \pm 0.1$ Gyr, and is in reasonable agreement with the sSFR distribution. While the SFR suppression time-scales that we find are longer than those found in hydrodynamical simulations (see Section 4.5), they are shorter than the gas depletion time-scales of 3.5–4 Gyr for field galaxies of comparable stellar mass (Boselli et al. 2014).

Taking all of these results together suggests a picture in which ram pressure stripping starts close to pericentre and is effective in a satellite galaxy's outskirts (where the restoring force is low compared to the force due to ram pressure) but may not be fully effective in their more tightly bound inner regions. For gas in the inner regions, while there may be no inflow of new cold gas (due to the complete stripping of the hot gas in the halo), the remaining cold gas will then be consumed by star formation. This consumption time-scale is shorter than in the field for two reasons: first, because there is less cold gas available due to the stripping in the outskirts; and secondly, because ram pressure stripped galaxies have modestly enhanced star formation rates at fixed stellar mass, compared to similar galaxies in the field, of 0.2–0.3 dex (Roberts & Parker 2020; Roberts et al. 2022).

5 CONCLUSIONS

We have combined SDSS photometry and spectroscopy, orbital information from tracking haloes in an *N*-body simulation, and simulated galaxies from the UNIVERSEMACHINE empirical model to constrain a simple model that suppresses star-formation histories of galaxies that have fallen into clusters. In the model, an infalling galaxy will have its star formation rate suppressed by an exponential envelope with time-scale τ_{env} after some delay time t_{delay}, relative to the time of the first pericentre. The parameter fits from this modeling give the mean quenching time-scales for the infalling population as a whole. We jointly fit these model parameters using both the quiescent fraction and mean deviation from the mean MWA–$M_\star$ relation (ΔMWA) of the infalling population in projected phase space. Doing so allows us to break the degeneracy between time of quenching onset and quenching duration that was present in previous models. The method accounts for interloper galaxies (which appear in projection, but are not physically in the cluster), and the pre-processing of infalling galaxies, allowing us to isolate the quenching effect of the (most recent) infall into a massive cluster. Our main results can be summarized as follows:

(i) The mean mass-weighted stellar age depends on location in projected phase space, with cluster member-dominated regions being older (by $\lesssim 1$ Gyr) relative to an interloper-dominated region.

(ii) Overall quenching times for our two stellar mass bins are driven by the quiescent fraction and are consistent with Oman et al. (2021), whose methodology we build on directly: $t_Q = 4.3 \pm 0.4$ Gyr and $t_Q = 3.9 \pm 0.2$ Gyr for our lower and higher stellar mass bins, respectively. We find longer overall quenching time-scales than other works in the literature where only star-formation rates are modelled, but agree with Taranu et al. (2014), who make additional use of the age-sensitive Balmer lines of quiescent galaxies.

(iii) Using mass-weighted ages allows us to break the degeneracy between t_{delay} and τ_{env}. We find that the onset of quenching occurs at $t_{\text{delay}} = 3.5^{+0.6}_{-0.9}$ Gyr and $t_{\text{delay}} = -0.3^{+0.8}_{-1.0}$ Gyr, relative to time of first pericentre, for galaxies in our $9 < \log(M_\star/M_\odot) < 10$ and $10 < \log(M_\star/M_\odot) < 10.5$ stellar mass bins, respectively. The models prefer a short SFR suppression time-scale, $\tau_{\text{env}} \leq 1.0$ Gyr (consistent with ram-pressure stripping), for our lower stellar mass bin, and a longer $2.3^{+0.5}_{-0.4}$ Gyr (consistent with strangulation) for our higher stellar mass bin.

(iv) In contrast to Wetzel et al. (2013), our model is able to reproduce the SFR bimodality even with long exponential suppression time-scales, thanks to the stochasticity of the UNIVERSEMACHINE star formation histories that we employ (as opposed to using smooth analytic star formation histories). We note that, for our lower stellar mass bin, the depth of the green valley prefers values of $\tau_{\text{env}} \gtrsim 1.5$ Gyr and $t_{\text{delay}} \sim 0$, in slight tension with the later quenching onset preferred by ΔMWAs.

Based on these findings and on our detailed discussion of the literature, we argue that satellites infalling into clusters experience ram pressure stripping of cold gas starting close to pericentre, which is only effective in the galaxy's outskirts, at least on the first pericentre passage. This leaves reduced cold gas available for continued star formation, resulting in star formation suppression time-scales of $\tau_{\text{env}} \sim 2$ Gyr – longer than if galaxies were fully stripped on their first pericentre passage, but shorter than a simple starvation scenario.

Future surveys like the Bright Galaxy Survey of the Dark Energy Spectroscopic Instrument (DESI) will provide a large increase in sample size over that of the one million galaxies in the SDSS DR14 main galaxy sample (Ruiz-Macias et al. 2021). An increase in sample size of galaxies with spectroscopically derived MWAs should reduce the errors on ΔMWA significantly. Using spectroscopically derived quantities more sensitive to recent SFR suppression or quenching, such as the time at which 90 per cent of the stellar mass has formed (Webb et al. 2020; Upadhyay et al. 2021), could also provide additional constraining power on infall-related and general quenching models. The models could also be improved by using a physically motivated model of ram-pressure stripping, rather than a generic time-scale for SFR decline. Such a model could involve radius and velocity at first pericentre, as suggested in a simple model by Owers et al. (2019, see also Roberts et al. 2019 for a quenching model depending on ICM density). With these various improvements, tighter constraints on time of quenching onset and duration via infall quenching as a function of stellar mass should be possible in the future.

DATA AVAILABILITY

The VVV simulation's initial conditions and snapshots are not currently publicly available. If they are needed, please ask the VVV authors for them. Alternatively, any *N*-body simulation of reasonably similar resolution and cosmology should yield the same results, statistically speaking. For satellite orbit and interloper data tables based on this simulation, request them from KAO (kyle.a.oman@durham.ac.uk). Alternatively, they may be created for any *N*-body simulation using the publicly available ROCKSTAR (bitbucket.org/gfcstanford/rockstar), CONSISTENT TREES (bitbucket.org/pbehroozi/consistent-trees), and ORBITPDF (github.com/kyleaoman/orbitpdf) codes.

The UNIVERSEMACHINE $z \sim 0$ simulation results using the *Bolshoi-Planck* dark matter simulation, including full star formation histories for all galaxies, are available in catalogue form via download from peterbehroozi.com/data.html.

The SDSS data used in this work is derived from Oman et al. (2021) and is available publicly as follows: Data Release 7 used in this work is available at skyserver.sdss.org/dr7; catalogues with stellar masses are available from VizieR (vizier.u-strasbg.fr), catalogue entry J/MNRAS/379/867; SFRs are available from wwwmpa.mpa-garching.mpg.de/SDSS/; the two group catalogues are separately available through VizieR (catalogue entry J/MNRAS/379/867) and gax.sjtu.edu.cn/data/Group.html. The mass- and luminosity-weighted ages and their accompanying stellar mass estimates are available from sdss.org/dr16/spectro/galaxy_firefly.

REFERENCES

Abazajian K. N. et al., 2009, ApJS, 182, 543
Allanson S. P., Hudson M. J., Smith R. J., Lucey J. R., 2009, ApJ, 702, 1275
Balogh M. L., Navarro J. F., Morris S. L., 2000, ApJ, 540, 113
Behroozi P. S., Wechsler R. H., Wu H.-Y., 2013a, ApJ, 762, 109
Behroozi P. S., Wechsler R. H., Wu H.-Y., Busha M. T., Klypin A. A., Primack J. R., 2013b, ApJ, 763, 18
Behroozi P., Wechsler R. H., Hearin A. P., Conroy C., 2019, MNRAS, 488, 3143
Beifiori A., Maraston C., Thomas D., Johansson J., 2011, A&A, 531, A109
Bekki K., Couch W. J., Shioya Y., 2002, ApJ, 577, 651
Bell E. F. et al., 2004, ApJ, 600, L11
Bernardi M., Nichol R. C., Sheth R. K., Miller C. J., Brinkmann J., 2006, AJ, 131, 1288
Biviano A., Murante G., Borgani S., Diaferio A., Dolag K., Girardi M., 2006, A&A, 456, 23
Boselli A., Gavazzi G., 2006, PASP, 118, 517
Boselli A., Cortese L., Boquien M., Boissier S., Catinella B., Lagos C., Saintonge A., 2014, A&A, 564, A66
Boselli A. et al., 2016, A&A, 596, A11
Boselli A., Fossati M., Sun M., 2022, A&AR, 30, 3
Brammer G. B. et al., 2011, ApJ, 739, 24
Brinchmann J., Charlot S., White S. D. M., Tremonti C., Kauffmann G., Heckman T., Brinkmann J., 2004, MNRAS, 351, 1151
Bryan G. L., Norman M. L., 1998, ApJ, 495, 80
Chabrier G., 2003, PASP, 115, 763
Comparat J. et al., 2017, preprint (arXiv:1711.06575)
Conroy C., Wechsler R. H., Kravtsov A. V., 2006, ApJ, 647, 201
Cooper M. C. et al., 2007, MNRAS, 376, 1445
De Lucia G., Weinmann S., Poggianti B. M., Aragón-Salamanca A., Zaritsky D., 2012, MNRAS, 423, 1277
Falcón-Barroso J., Sánchez-Blázquez P., Vazdekis A., Ricciardelli E., Cardiel N., Cenarro A. J., Gorgas J., Peletier R. F., 2011, A&A, 532, A95
Fossati M. et al., 2018, A&A, 614, A57
Fujita Y., 2004, PASJ, 56, 29
Gavazzi G., Consolandi G., Gutierrez M. L., Boselli A., Yoshida M., 2018, A&A, 618, A130
Gerke B. F. et al., 2007, MNRAS, 376, 1425
Graves G. J., Faber S. M., Schiavon R. P., Yan R., 2007, ApJ, 671, 243
Hayward C. C. et al., 2014, MNRAS, 445, 1598
Jaffé Y. L., Smith R., Candlish G. N., Poggianti B. M., Sheen Y.-K., Verheijen M. A. W., 2015, MNRAS, 448, 1715
Jaffé Y. L. et al., 2018, MNRAS, 476, 4753
Kelkar K., Gray M. E., Aragón-Salamanca A., Rudnick G., Jaffé Y. L., Jablonka P., Moustakas J., Milvang-Jensen B., 2019, MNRAS, 486, 868
Khullar G. et al., 2022, ApJ, 934, 177
Kim S., Contini E., Choi H., Han S., Lee J., Oh S., Kang X., Yi S. K., 2020, ApJ, 905, 12
Kim K. J. et al., 2022, preprint (arXiv:2207.12491)
Klypin A., Yepes G., Gottlöber S., Prada F., Heß S., 2016, MNRAS, 457, 4340
Larson R. B., Tinsley B. M., Caldwell C. N., 1980, ApJ, 237, 692
Lim S. H., Mo H. J., Lu Y., Wang H., Yang X., 2017, MNRAS, 470, 2982
Lotz M., Remus R.-S., Dolag K., Biviano A., Burkert A., 2019, MNRAS, 488, 5370
Ludlow A. D., Navarro J. F., Angulo R. E., Boylan-Kolchin M., Springel V., Frenk C., White S. D. M., 2014, MNRAS, 441, 378
Mahajan S., Mamon G. A., Raychaudhury S., 2011, MNRAS, 416, 2882
Maraston C., Strömbäck G., 2011, MNRAS, 418, 2785
McGee S. L., Balogh M. L., Bower R. G., Font A. S., McCarthy I. G., 2009, MNRAS, 400, 937
Mendel J. T., Simard L., Palmer M., Ellison S. L., Patton D. R., 2014, ApJS, 210, 3
Mok A. et al., 2013, MNRAS, 431, 1090
Moster B. P., Naab T., White S. D. M., 2018, MNRAS, 477, 1822
Munari E., Biviano A., Borgani S., Murante G., Fabjan D., 2013, MNRAS, 430, 2638
Muzzin A. et al., 2012, ApJ, 746, 188
Muzzin A. et al., 2014, ApJ, 796, 65
Naab T., Ostriker J. P., 2017, ARA&A, 55, 59
Navarro J. F., Frenk C. S., White S. D. M., 1997, ApJ, 490, 493
Nelan J. E., Smith R. J., Hudson M. J., Wegner G. A., Lucey J. R., Moore S. A. W., Quinney S. J., Suntzeff N. B., 2005, ApJ, 632, 137

Noble A. G., Webb T. M. A., Yee H. K. C., Muzzin A., Wilson G., Van Burg R. F. J., Balogh M. L., Shupe D. L., 2016, ApJ, 816, 48

Oman K. A., Hudson M. J., 2016, MNRAS, 463, 3083

Oman K. A., Hudson M. J., Behroozi P. S., 2013, MNRAS, 431, 2307

Oman K. A., Bahé Y. M., Healy J., Hess K. M., Hudson M. J., Verheijen M. A. W., 2021, MNRAS, 501, 5073

Owers M. S. et al., 2019, ApJ, 873, 52

Paccagnella A. et al., 2016, ApJ, 816, L25

Pasquali A., Smith R., Gallazzi A., De Lucia G., Zibetti S., Hirschmann M., Yi S. K., 2019, MNRAS, 484, 1702

Peng Y.-j. et al., 2010, ApJ, 721, 193

Pintos-Castro I., Yee H. K. C., Muzzin A., Old L., Wilson G., 2019, ApJ, 876, 40

Planck Collaboration XIII, 2016, A&A, 594, A13

Reeves A. M. M. et al., 2021, MNRAS, 506, 3364

Rhee J., Smith R., Choi H., Contini E., Jung S. L., Han S., Yi S. K., 2020, ApJS, 247, 45

Roberts I. D., Parker L. C., 2020, MNRAS, 495, 554

Roberts I. D., Parker L. C., Brown T., Joshi G. D., Hlavacek-Larrondo J., Wadsley J., 2019, ApJ, 873, 42

Roberts I. D. et al., 2021, A&A, 650, A111

Roberts I. D. et al., 2022, MNRAS, 509, 1342

Rodríguez-Puebla A., Behroozi P., Primack J., Klypin A., Lee C., Hellinger D., 2016, MNRAS, 462, 893

Ruiz-Macias O. et al., 2021, MNRAS, 502, 4328

Salim S. et al., 2007, ApJS, 173, 267

Sánchez-Blázquez P. et al., 2006, MNRAS, 371, 703

Smith R. J., Hudson M. J., Lucey J. R., Nelan J. E., Wegner G. A., 2006, MNRAS, 369, 1419

Smith R. J. et al., 2010, MNRAS, 408, 1417

Smith R. J., Lucey J. R., Price J., Hudson M. J., Phillipps S., 2012, MNRAS, 419, 3167

Taranu D. S., Hudson M. J., Balogh M. L., Smith R. J., Power C., Oman K. A., Krane B., 2014, MNRAS, 440, 1934

Thomas D., Maraston C., Bender R., Mendes de Oliveira C., 2005, ApJ, 621, 673

Upadhyay A. K., Oman K. A., Trager S. C., 2021, A&A, 652, A16

Van Burg R. F. J., McGee S., Aussel H., Dahle H., Arnaud M., Pratt G. W., Muzzin A., 2018, A&A, 618, A140

Van Burg R. F. J. et al., 2020, A&A, 638, A112

Von Der Linden A., Best P. N., Kauffmann G., White S. D. M., 2007, MNRAS, 379, 867

Vulcani B. et al., 2018, ApJ, 866, L25

Wang J., Bose S., Frenk C. S., Gao L., Jenkins A., Springel V., White S. D. M., 2020, Nature, 585, 39

Webb K. et al., 2020, MNRAS, 498, 5317

Wechsler R. H., Tinker J. L., 2018, ARA&A, 56, 435

Wetzel A. R., Tinker J. L., Conroy C., 2012, MNRAS, 424, 232

Wetzel A. R., Tinker J. L., Conroy C., van den Bosch F. C., 2013, MNRAS, 432, 336

Wilkinson D. M., Maraston C., Goddard D., Thomas D., Parikh T., 2017, MNRAS, 472, 4297

Wright R. J., Lagos C. d. P., Power C., Stevens A. R. H., Cortese L., Poulton R. J. J., 2022, MNRAS, 516, 2891

APPENDIX A: ROBUSTNESS OF PARAMETER CONSTRAINTS TO CHANGES IN STAR FORMATION HISTORY AND FLOATING INFALL PARAMETERS

An essential assumption for our modelling results and conclusions to be robust is that they are not particularly sensitive to changes in overall star formation history, for both f_Q as well as ΔMWA. Throughout this work, we simply allowed our infalling/interloper UNIVERSEMACHINE galaxies' f_Q to float to match that of SDSS. As a test of these assumptions, we now consider a modified model where the UNIVERSEMACHINE star formation histories are adjusted such that the infalling UNIVERSEMACHINE galaxies have the same quiescent fraction as for SDSS.

The mismatch between the level of quenching observed in the infall region of UNIVERSEMACHINE and SDSS clusters is particularly noticeable at stellar masses $\log(M_\star/M_\odot) > 10$. To be specific, we
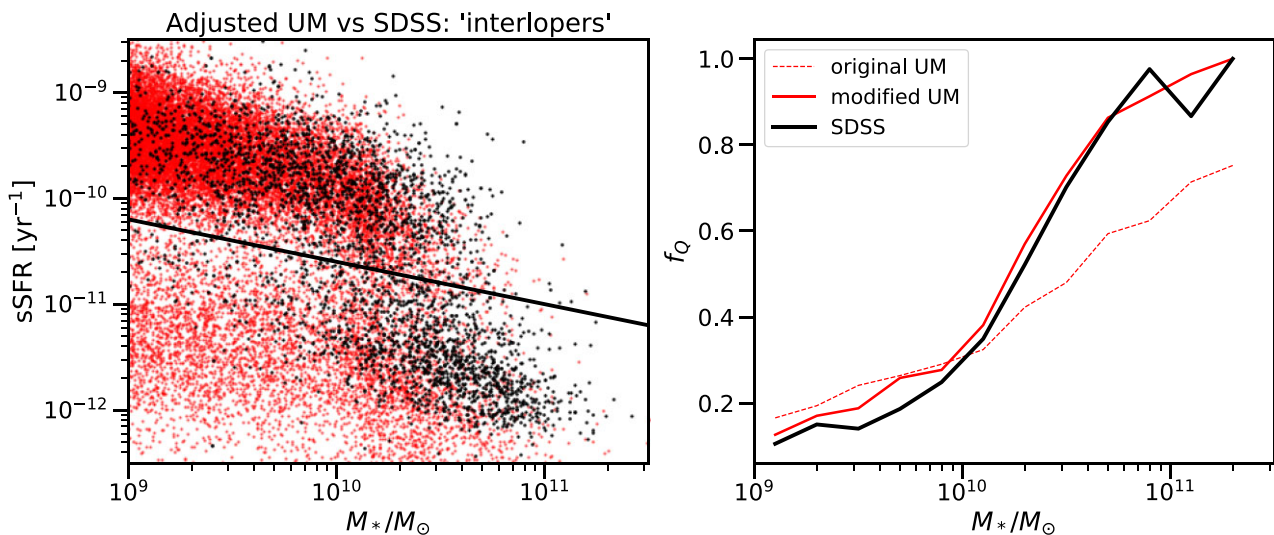

Figure A1. Plots illustrating the adjusted UNIVERSEMACHINE interlopers (red) relative to the SDSS data (black). Above $\log(M_\star/M_\odot) = 10.2$, galaxies have had their star formation history suppressed to increase their f_Q to match the SDSS data, whereas galaxies below $\log(M_\star/M_\odot) = 10.2$ have had their star formation boosted. Left-hand panel: specific star formation rate plotted against stellar mass, with the adjusted UNIVERSEMACHINE sSFR multiplied by 2.1 (vertical shift) for the purposes of comparing with SDSS data on this plot. The quiescent/star-forming cut used for both is shown with the solid black line. Right-hand panel: quiescent fraction versus stellar mass. The adjusted UNIVERSEMACHINE interloper sSFRs increase their f_Q (dashed red) up to that of the SDSS data for higher stellar masses. There is a small offset in the two curves as only the shape of the $f_Q - M_\star$ relation was fit in this stellar mass binning and the absolute f_Q values were fit for stellar mass bins with edges at $\log(M_\star/M_\odot) = (9, 10, 10.5, 11, 11.5)$.

now check whether we need to have star formation histories that match the $z \sim 0$, $f_Q - M_\star$ relation of infalling UNIVERSEMACHINE galaxies to that found for SDSS. For the purposes of this test we consider interlopers to be galaxies in the three upper-right bins in our PPS plots, namely the bins that define the infall region in e.g. Fig. 1.

Our quiescent fraction is several per cent too high at low stellar masses and several per cent too low at higher stellar masses, as apparent in Fig. A1. Our floating $f_{Q, \text{infall}}$ parameter removes this overall offset, leaving behind just the differential trends (across PPS) in f_Q. This resolves possible issues due to differences in completeness between star forming and quiescent galaxies, particularly in the lower stellar mass bin. However, as a robustness check to see how much differences in f_Q might affect our measured time-scales, we run a test where we modify the UNIVERSEMACHINE star formation histories such that we closely match the $f_Q - M_\star$ trend observed in SDSS. Star formation histories are suppressed by multiplying by an exponential function in time such that the UNIVERSEMACHINE sSFR–$M_\star$ slope better matches the SDSS data and also better matches the SDSS f_Q. Explicitly, this modified star formation history in terms of the unmodified UNIVERSEMACHINE star formation history (SFH) as a function of cosmic time, t, is expressed as:

$$\text{SFH}'(t) = \exp\left(\frac{t - t_{z=0}}{\tau}\right)\text{SFH}(t), \tag{A1}$$

where $t_{\text{now}} = 13.8$ Gyr, $\tau = 1/k$, and $k = (\ln(10)/t_{\text{now}}) \times -0.4(\log(M_\star/\text{M}_\odot) - 10)$, with -0.4 coming from the slope of the SDSS sSFR cut. The value of this correction at $z = 0$ is such that the increase or decrease in SFR forms a broken power law with slope -1 for $\log(M_\star/\text{M}_\odot) > 10$ and slope -0.7 for $\log(M_\star/\text{M}_\odot) < 10$, with

the value of the correction set to 1.0 at $\log(M_\star/\text{M}_\odot) = 10.3$. This corresponds to a boost in the star formation histories for $\log(M_\star/\text{M}_\odot) < 10.3$ and suppression for $\log(M_\star/\text{M}_\odot) > 10.3$. The correction is applied to the 'true' rather than 'observed' UNIVERSEMACHINE SFHs as well as the 'true' final SFRs and stellar masses (see Behroozi et al. 2019, for these definitions). A shift plus scatter is then applied to approximate observed versions of the quantities afterwards, according to the prescription used in Behroozi et al. (2019).

We find that this adjustment only results in minor effects on the best-fitting t_{delay} and τ_{env} parameters. Specifically, the preferred t_{delay} changes by $+0.2$ Gyr and τ_{env} by -0.2 Gyr for our lower stellar mass bin, or in terms of t_Q, a shift of 0.5 Gyr. For the $10 < \log(M_\star/\text{M}_\odot) < 10.5$ stellar mass bin, the corresponding shifts in t_{delay} and τ_{env} are $+0.1$ Gyr and -0.5 Gyr, respectively. Given our uncertainties of ~ 0.5 Gyr for τ_{env} and $\lesssim 1.0$ Gyr for t_{delay} (see Table 1), we conclude that any impact on our parameter constraints from changes in overall star formation history is negligible and therefore does not influence our discussion and qualitative conclusions.

APPENDIX B: DETAILED MODELLING PREDICTIONS FOR HIGHER STELLAR MASS BIN

Analogous to Figs 7 and 8, we show our detailed model predictions for $10 < \log(M_\star/\text{M}_\odot) < 10.5$ galaxies in bins of PPS for f_Q and ΔMWA in Fig. B1. Similar to our lower stellar mass bin, a somewhat different range of t_{delay} and τ_{env} values are preferred by f_Q and ΔMWA. The joint best fit for these parameters using both observables is given in Table 1, with the contours shown in Fig. 9.
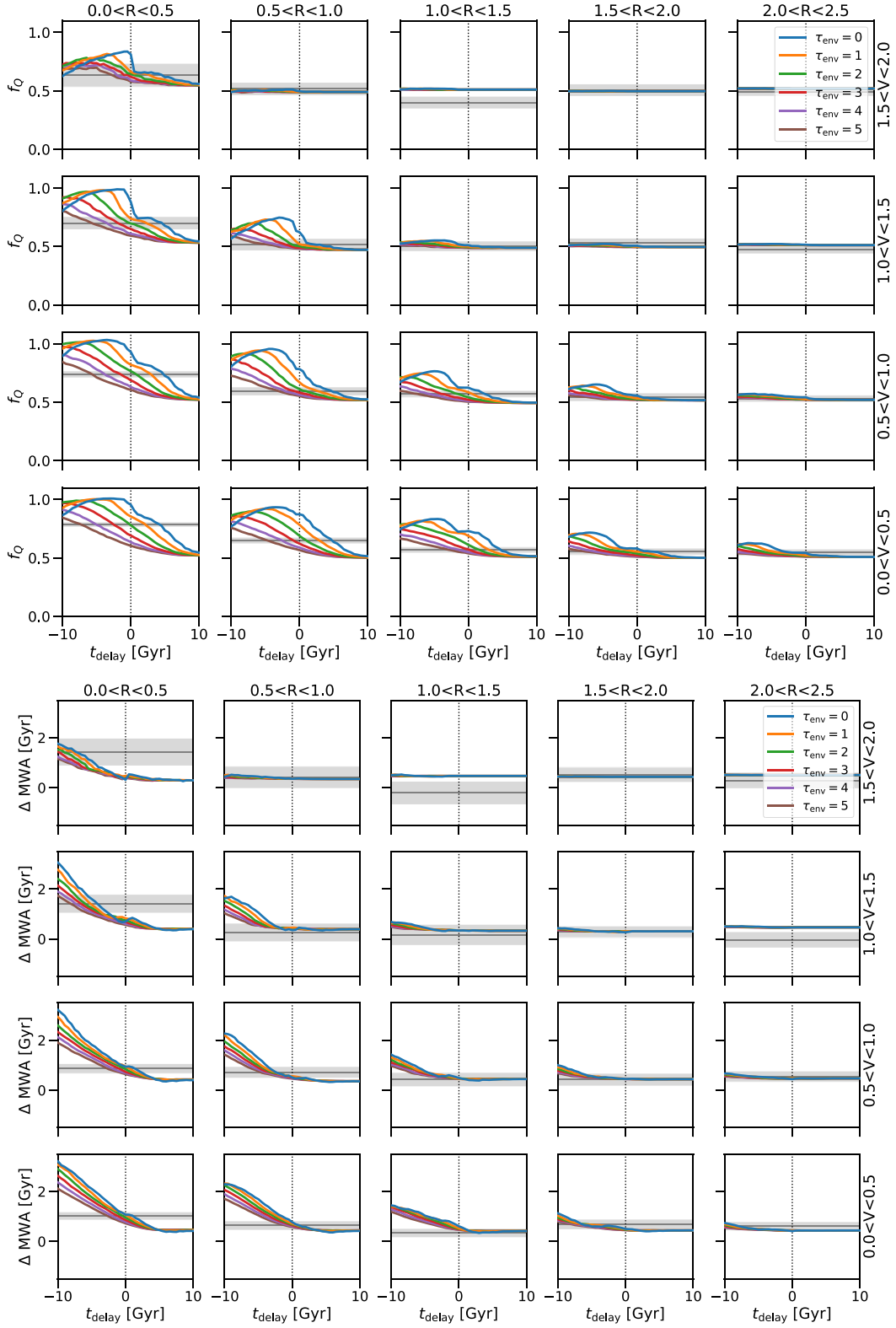
Figure B1. f_Q (upper multi-paneled figure) and mean ΔMWA (lower figure) predictions for $10 < \log(M_\star/M_\odot) < 10.5$ galaxies for a range of models where galaxies quench after some delay time, relative to time of first pericentre. We show models run for a range of exponential suppression time-scales. The SDSS mean values are shown with grey lines, with the shaded region showing the bootstrapped (over clusters) error on the mean. Analogous to Figs 7 and 8.

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