

A semi-analytic model for the co-evolution of galaxies, black holes and active galactic nuclei

Rachel S. Somerville,^{1*} Philip F. Hopkins,² Thomas J. Cox,² Brant E. Robertson^{3,4†} and Lars Hernquist²

¹Max-Planck-Institut für Astronomie, Königstuhl 17, Heidelberg D-69117, Germany

²Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA

³Kavli Institute for Cosmological Physics, and the Department of Astronomy and Astrophysics, University of Chicago, 933 East 56th Street, Chicago, IL 60637, USA

⁴Enrico Fermi Institute, 5640 South Ellis Avenue, Chicago, IL 60637, USA

Accepted 2008 August 5. Received 2008 July 18; in original form 2008 February 25

ABSTRACT

We present a new semi-analytic model that self-consistently traces the growth of supermassive black holes (BH) and their host galaxies within the context of the Lambda cold dark matter (Λ CDM) cosmological framework. In our model, the energy emitted by accreting black holes regulates the growth of the black holes themselves, drives galactic scale winds that can remove cold gas from galaxies, and produces powerful jets that heat the hot gas atmospheres surrounding groups and clusters. We present a comprehensive comparison of our model predictions with observational measurements of key physical properties of low-redshift galaxies, such as cold gas fractions, stellar metallicities and ages, and specific star formation rates. We find that our new models successfully reproduce the exponential cut-off in the stellar mass function and the stellar and cold gas mass densities at $z \sim 0$, and predict that star formation should be largely, but not entirely, quenched in massive galaxies at the present day. We also find that our model of self-regulated BH growth naturally reproduces the observed relation between BH mass and bulge mass. We explore the global formation history of galaxies and black holes in our models, presenting predictions for the cosmic histories of star formation, stellar mass assembly, cold gas and metals. We find that models assuming the ‘concordance’ Λ CDM cosmology overproduce star formation and stellar mass at high redshift ($z \gtrsim 2$). A model with less small-scale power predicts less star formation at high redshift, and excellent agreement with the observed stellar mass assembly history, but may have difficulty accounting for the cold gas in quasar absorption systems at high redshift ($z \sim 3$ –4).

Key words: galaxies: evolution – galaxies: formation – cosmology: theory.

1 INTRODUCTION

It is now well established that the cold dark matter (CDM) paradigm for structure formation (Blumenthal et al. 1984), in its modern, dark energy dominated (Λ CDM) incarnation, provides a remarkably successful paradigm for interpreting a wide variety of observations, from the cosmic microwave background fluctuations at $z \sim 1000$ (Spergel et al. 2003, 2007), to the large-scale clustering of galaxies at $z \sim 0$ (Percival et al. 2002; Tegmark et al. 2004; Eisenstein et al. 2005). Other successful predictions of the Λ CDM paradigm include the cosmic shear field as measured by weak gravitational lensing (Bacon et al. 2005; Heymans et al. 2005; Hoekstra et al.

2006), the small-scale power spectrum as probed by the Lyman α forest (Desjacques & Nusser 2005; Jena et al. 2005), and the number densities of (Borgani et al. 2001), and baryon fractions within, galaxy clusters (White et al. 1993; Allen et al. 2004).

Λ CDM as a paradigm for understanding and simulating galaxy formation has had more mixed success. In this picture, as originally proposed by White & Rees (1978) and Blumenthal et al. (1984), galaxies form when gas cools and condenses at the centres of dark matter (DM) dominated potential wells, or ‘haloes’. More detailed calculations, using semi-analytic and numerical simulations of galaxy formation, have shown that this framework does provide a promising qualitative understanding of many features of galaxies and their evolution (e.g. Kauffmann, White & Guiderdoni 1993; Cole et al. 1994; Kauffmann et al. 1998; Somerville & Primack 1999; Cole et al. 2000; Somerville, Primack & Faber 2001; Baugh 2006). However, it has been clear for at least a decade now

*E-mail: somerville@mpia.de

†Spitzer Fellow.

that there is a fundamental tension between certain basic predictions of the Λ CDM-based galaxy formation paradigm and some of the most fundamental observable properties of galaxies. In this paper we focus on two interconnected, but possibly distinct, problems: (1) the ‘overcooling’ or ‘massive galaxy’ problem and (2) the star formation ‘quenching’ problem.

The first problem is manifested by the fact that both semi-analytic and numerical simulations predict that a large fraction of the available baryons in the Universe rapidly cools and condenses, in conflict with observations which indicate that only about ~ 5 – 10 per cent of the baryons are in the form of cold gas and stars (Bell et al. 2003a; Fukugita & Peebles 2004). This is due to the fact that gas at the densities and temperatures characteristic of DM haloes is expected to cool rapidly, and is related to the classical ‘cooling flow’ problem (Cowie & Binney 1977; Fabian & Nulsen 1977; Mathews & Bregman 1978). Direct observations of the X-ray properties of hot gas in clusters similarly imply that this gas should have short cooling times, particularly near the centre of the cluster, but the condensations of stars and cold gas at the centres of these clusters are much smaller than would be expected if the hot gas had been cooling so efficiently over the lifetime of the cluster (for reviews see Fabian 1994; Peterson & Fabian 2006). Moreover, X-ray spectroscopy shows that very little gas is cooling below a temperature of about one-third of the virial temperature of the cluster (Peterson et al. 2003).

A further difficulty is that there is a fundamental mismatch between the *shape* of the DM halo mass function and that of the observed mass function of cold baryons (cold gas and stars) in galaxies (Somerville & Primack 1999; Benson et al. 2003). The galaxy mass function has a sharp exponential cut-off above a mass of about a few $\times 10^{10} M_{\odot}$, while the halo mass function has a shallower, power-law cut-off at much higher mass ($\sim \text{few} \times 10^{13} M_{\odot}$). There is also a mismatch at the small mass end, as the mass function of DM haloes is much steeper than that of galaxies. If we are to assume that each DM halo hosts a galaxy, this then implies that the ratio between the luminosity or stellar mass of a galaxy and the mass of its DM halo varies strongly and non-monotonically with halo mass (Kravtsov et al. 2004a; Conroy, Wechsler & Kravtsov 2006; Wang et al. 2006; Moster et al., in preparation), such that ‘galaxy formation’ is much more inefficient in both small-mass and large-mass haloes, with a peak in efficiency close to the mass of our own Galaxy, $\sim 10^{12} M_{\odot}$. On very small-mass scales (below halo velocities of ~ 30 – 50 km s^{-1}), the collapse and cooling of baryons may be suppressed by the presence of a photoionizing background (Efstathiou 1992; Thoul & Weinberg 1996; Quinn, Katz & Efstathiou 1996). For larger mass haloes (up to $V_{\text{vir}} \simeq 150$ – 200 km s^{-1}), the standard assumption is that winds driven by massive stars and supernovae (SNe) are able to heat and expel gas, resulting in low baryon fractions in small-mass haloes (White & Rees 1978; Dekel & Silk 1986; White & Frenk 1991). However, stellar feedback probably cannot provide a viable solution to the overcooling problem in massive haloes (Benson et al. 2003): stars do not produce enough energy to expel gas from these large potential wells; and the massive, early-type galaxies in which the energy source is needed have predominantly old stellar populations and little or no ongoing or recent star formation. Other solutions, like thermal conduction, have been explored, but probably do not provide a full solution (Benson et al. 2003; Voigt & Fabian 2004).

The second problem is related to the correlation of galaxy structural properties (morphology) and spectrophotometric properties (stellar populations) with stellar mass. The presence of such a correlation has long been known, in the sense that more massive galaxies

tend to be predominantly spheroid-dominated, with red colours, old stellar populations, low gas fractions and little recent star formation, while low-mass galaxies tend to be disc-dominated and gas-rich, with blue colours and ongoing star formation (e.g. Roberts & Haynes 1994). More recently, with the advent of large galaxy surveys such as the Sloan Digital Sky Survey (SDSS), we have learned that the galaxy colour distribution (and that of other related properties) is strongly *bimodal* (e.g. Baldry et al. 2004), and that the transition in galaxy properties from star-forming discs to ‘dead’ spheroids occurs rather sharply, at a characteristic stellar mass of $\sim 3 \times 10^{10} M_{\odot}$ (Kauffmann et al. 2003a; Brinchmann et al. 2004). In contrast, the ‘standard’ Λ CDM-based galaxy formation models predict that massive haloes have been assembled relatively recently, and should contain an ample supply of new fuel for star formation. These models predict an *inverted* colour–mass and morphology–mass relation (massive galaxies tend to be blue and disc dominated, rather than red and spheroid dominated) and no sharp transition or strong bimodality. Thus, the standard paradigm of galaxy formation does not provide a physical explanation for the ‘special’ mass scale (a halo mass of $\sim 10^{12} M_{\odot}$, or a stellar mass of $\sim 3 \times 10^{10} M_{\odot}$) which marks both the peak of galaxy formation efficiency and the transition in galaxy properties seen in observations.

Several pieces of observational evidence provide clues to the solution to these problems. It is now widely believed that every spheroid-dominated galaxy hosts a nuclear supermassive black hole (SMBH), and that the mass of the SMBH is tightly correlated with the luminosity, mass and velocity dispersion of the stellar spheroid (Kormendy & Richstone 1995; Magorrian et al. 1998; Ferrarese & Merritt 2000; Gebhardt et al. 2000; Marconi & Hunt 2003; Häring & Rix 2004). These correlations may be seen as a kind of ‘fossil’ evidence that black holes were responsible for regulating the growth of galaxies and vice versa. This also implies that the most massive galaxies, where quenching is observed to be the most efficient, host the largest black holes, and therefore the available energy budget is greatest in precisely the systems where it is needed, in contrast to the case of stellar feedback. The integrated energy released over the lifetime of an SMBH ($\simeq 10^{60}$ – 10^{62} erg) is clearly very significant compared with galaxy binding energies (Silk & Rees 1998). In view of these facts, it seems almost inconceivable that active galactic nucleus (AGN) feedback is *not* important in shaping galaxy properties.

However, in order to build a complete, self-consistent machinery to describe the formation and growth of black holes within the framework of a cosmological galaxy formation model, and to attempt to treat the impact of the energy feedback from black holes in this context, we need to address several basic questions. (1) When, where and with what masses do seed black holes form? (2) What triggers black hole accretion, what determines the efficiency of this accretion and what shuts it off? (3) In what form is the energy produced by the black hole released, and how does this energy couple with the host galaxy and its surroundings? In order to address some of these questions, we first identify two modes of AGN activity which have different observational manifestations, probably correspond to different accretion mechanisms and have different physical channels of interaction with galaxies.

1.1 The bright mode of black hole growth

Classical luminous quasars and their less powerful cousins, optical or X-ray bright AGN, radiate at a significant fraction of their Eddington limit ($L \sim (0.1\text{--}1)L_{\text{Edd}}$; Vestergaard 2004; Kollmeier et al. 2006), and are believed to be fed by optically thick,

geometrically thin accretion discs (Shakura & Syunyaev 1973). We will refer to this mode of accretion as the ‘bright mode’ because of its relatively high radiative efficiency (with a fraction $\eta_{\text{rad}} \sim 0.1\text{--}0.3$ of the accreted mass converted to radiation). The observed space density of these quasars and AGN is low compared to that of galaxies, implying that if most galaxies indeed host an SMBH, this ‘bright mode’ of accretion is only ‘on’ a relatively small fraction of the time. Constraints from quasar clustering and variability imply that quasar lifetimes must be $\lesssim 10^{8.5}$ yr (Martini & Weinberg 2001; Martini & Schneider 2003). These short time-scales combined with the large observed luminosities immediately imply that fuelling these objects requires funnelling a quantity of gas comparable to the entire supply of a large galaxy ($\sim 10^9\text{--}10^{10} M_{\odot}$) into the central regions on a time-scale of the order of the dynamical time ($\sim \text{few} \times 10^7\text{--}10^8$ yr).

These considerations alone lead one to consider galaxy–galaxy mergers as a promising mechanism for triggering this efficient accretion on to nuclear black holes. The observational association of mergers with enhanced star formation, particularly with the most violent observed episodes of star formation exhibited by ultraluminous infrared galaxies (ULIRGS), is well established (Sanders & Mirabel 1996; Barton, Geller & Kenyon 2000; Colina et al. 2001; Farrah et al. 2001; Woods, Geller & Barton 2006; Woods & Geller 2007; Barton et al. 2007; Li et al. 2007; Lin et al. 2007). Moreover, numerical simulations have shown that tidal torques during galaxy mergers can drive the rapid inflows of gas that are needed to fuel both the intense starbursts and rapid black hole accretion associated with ULIRGS and quasars (Hernquist 1989; Barnes 1992; Mihos & Hernquist 1994; Barnes & Hernquist 1996; Mihos & Hernquist 1996; Di Matteo, Springel & Hernquist 2005; Springel, Di Matteo & Hernquist 2005b). As well, it seems that if one can probe sufficiently deep to study the spectral energy distribution beneath the glare of the quasar, one always uncovers evidence of young stellar populations indicative of a recent starburst (Brotherton et al. 1999; Canalizo & Stockton 2001; Kauffmann et al. 2003b; Jahnke et al. 2004; Sánchez et al. 2004; Vanden Berk et al. 2006). Near-equal mass (major) mergers also have the attractive feature that they scramble stars from circular to random orbits, leading to morphological transformation from disc to spheroid (Toomre & Toomre 1972; Barnes 1988, 1992; Hernquist 1992, 1993). If spheroids and black holes both arise from violent mergers, this provides a possible explanation for why black hole properties always seem to be closely associated with the spheroidal components of galaxies.

What impact does the energy associated with this rapid, bright mode growth have on the galaxy and on the growth of the black hole itself? Long thought to be associated only with a small subset of objects [e.g. broad absorption line (BAL) quasars], high-velocity winds have been detected in a variety of different types of quasar systems (de Kool et al. 2001; Chartas, Brandt & Gallagher 2003; Pounds et al. 2003; Pounds & Page 2006), and are now believed to be quite ubiquitous (Ganguly & Brotherton 2008). However, their impact on the host galaxy remains unclear, as the mass outflow rates of these winds are difficult to constrain (though see Steenbrugge et al. 2005; Chartas et al. 2007; Krongold et al. 2007). Recently, numerical simulations of galaxy mergers including black hole growth found that depositing even a small fraction (~ 5 per cent) of the energy radiated by the BH into the interstellar medium (ISM) can not only halt the accretion on to the BH, but can drive large-scale winds (Di Matteo et al. 2005). These winds sweep the galaxy nearly clean of cold gas and halt further star formation, leaving behind a rapidly reddening, spheroidal remnant (Springel, Di Matteo & Hernquist 2005a).

To study how the interplay between feedback from SMBH accretion and SNe, galaxy structure, orbital configuration and gas dissipation combine to determine the properties of spheroidal galaxies formed through mergers, hundreds of hydrodynamical simulations were performed by Robertson et al. (2006a,b,c) and Cox et al. (2006a,b) using the methodology presented by Di Matteo et al. (2005) and Springel et al. (2005b). Robertson et al. and Cox et al. analysed the merger remnants to study the redshift evolution of the BH mass– σ relation, the Fundamental Plane, phase-space density and kinematic properties. This extensive suite has been supplemented by additional simulations of minor mergers from Cox et al. (2008). Throughout the rest of this paper, when we refer to ‘the merger simulations’, we refer to this suite.

Based on their analysis of these simulations, Hopkins et al. (2008) have outlined an evolutionary sequence from galaxy–galaxy merger, to dust-enshrouded starburst and buried AGN, blow-out of the dust and ISM by the quasar- and starburst-driven winds, to classical (unobscured) quasar, post-starburst galaxy and finally ‘dead’ elliptical. Hopkins et al. (2007a) find that in the merger simulations, the accretion on to the BH is eventually halted by a pressure-driven outflow. Because the depth of the spheroid’s potential well determines the amount of momentum necessary to entrain the infalling gas, Hopkins et al. (2007a) find that this leads to a ‘black hole fundamental plane’, a correlation between the final black hole mass and sets of spheroid structural/dynamical properties (mass, size, velocity dispersion) similar to the one seen in observations (Marconi & Hunt 2003; Hopkins et al. 2007b).

Furthermore, Hopkins et al. (2005a,b,d) have shown that the self-regulated nature of black hole growth in these simulations leads to a characteristic form for quasar light curves. As the galaxies near their final coalescence, the accretion rises to approximately the Eddington rate. After the critical black hole mass is reached and the outflow phase begins, the accretion rate enters a power-law decline phase. Although most of their growth occurs in the near-Eddington phase, quasars spend much of their time in the decline phase, and this implies that many observed low-luminosity quasars are actually relatively massive black holes in the last stages of their slow decline. Hopkins et al. (2006d) found that when these light curves are convolved with the observed mass function of merging galaxies, the predicted AGN luminosity function is consistent with observations. Moreover, Hopkins et al. (2005b,c, 2006a,c) have shown that this picture reproduces many quasar and galaxy observables that are difficult to account for with more simplified assumptions about QSO light curves, such as differences in the quasar luminosity function in different bands and redshifts, Eddington ratio and column density distributions, the X-ray background spectrum, and relic red, early-type galaxy population colours and distributions.

1.2 The radio mode

The second mode of AGN activity is much more common, and in general less dramatic. A fairly large fraction of massive galaxies (particularly galaxies near the centres of groups and clusters) are detected at radio wavelengths (Best et al. 2005, 2007). Most of these radio sources do not have emission lines characteristic of classical optical or X-ray bright quasars (Best et al. 2005; Kauffmann, Heckman & Best 2008), and their accretion rates are believed to be a small fraction of the Eddington rate (Rafferty et al. 2006). They are extremely radiatively inefficient (Birzan et al. 2004), and thought to be fuelled by optically thin, geometrically thick accretion as expected in ADAF and ADIOS models such as those proposed by Narayan & Yi (1994) and Blandford & Begelman (1999). Because

these objects are generally identified via their radio emission, we refer to this mode of accretion and BH growth as the ‘radio mode’ (following Croton et al. 2006).

Although these black holes seem to be inefficient at producing radiation, they can apparently be quite efficient at producing kinetic energy in the form of relativistic jets. Intriguingly, the majority of cooling flow clusters host these active radio galaxies at their centres (Dunn & Fabian 2006, 2008), and X-ray maps reveal that the radio lobes are often spatially coincident with cavities, thought to be bubbles filled with relativistic plasma and inflated by the jets (McNamara & Nulsen 2007, and references therein). The observations of these bubbles can be used to estimate the work required to inflate them against the pressure of the hot medium (Bîrzan et al. 2004; Allen et al. 2006; Rafferty et al. 2006), and hence obtain lower limits on the jet power.

While the idea that radio jets provide a heat source that could counteract cooling flows has been discussed for many years (e.g. Binney & Tabor 1995; Churazov et al. 2002; Fabian et al. 2003; Binney 2004; Omma et al. 2004), these observations now make it possible to investigate more quantitatively whether the heating rates are sufficient to offset the cooling rates in groups and clusters. Several studies conclude that in the majority of the systems studied, the AGN heating traced by the power in the X-ray cavities alone is comparable to or in excess of the energy being radiated by the cooling gas (Best et al. 2006; Fabian et al. 2006; McNamara et al. 2006; Rafferty et al. 2006; McNamara & Nulsen 2007; Dunn & Fabian 2008). Moreover, the net cooling rate is correlated with the observed star formation rate (SFR) in the central cD galaxy, indicating that there may be a self-regulating cycle of heating and cooling (Rafferty et al. 2006).

Several other physical processes that could suppress cooling in large-mass haloes have been suggested and explored, such as thermal conduction (Benson et al. 2003; Voigt & Fabian 2004), multi-phase cooling (Maller & Bullock 2004), or heating by substructure or clumpy accretion (Khochfar & Ostriker 2007; Naab et al. 2007; Dekel & Birnboim 2008). While some or all of these processes may well be important, in this paper we will investigate whether it is plausible that ‘radio mode’ heating alone can do the job.

1.3 A unified model for black hole activity and AGN feedback

All of this begs the question: what determines whether a black hole accretes in the ‘bright mode’ or ‘radio mode’ state? An interesting possible answer comes from an analogy with X-ray binaries (Jester 2005; K rding, Jester & Fender 2006). Observers can watch X-ray binaries in real time as they transition between two states: the ‘low/hard’ state, in which a steady radio jet is present and a hard X-ray spectrum is observed, and the ‘high/soft’ state, in which the jet disappears and the X-ray spectrum has a soft, thermal component (Maccarone, Gallo & Fender 2003; Fender, Belloni & Gallo 2004). The transition between the two states is thought to be connected to the accretion rate itself: the ‘high/soft’ state is associated with accretion rates of $\gtrsim (0.01-0.02) \dot{m}_{\text{Edd}}$ and the existence of a classical thin accretion disc, while the ‘low/hard’ state is associated with lower accretion rates and radiatively inefficient ADAF/ADIOS accretion (Fender et al. 2004).

Recently, Sijacki et al. (2007) have applied this idea in cosmological hydrodynamic simulations, by assuming that when the accretion rate exceeds a critical value, ‘bright mode’ feedback occurs (AGN-driven winds), while when the accretion rate is lower, ‘radio mode’ (mechanical bubble feedback) is implemented. The results of their simulations appear promising –they produced black hole

and stellar mass densities in broad agreement with observations. In addition, they found that their implementation of AGN feedback was able to suppress strong cooling flows and produce shallower entropy profiles in clusters, and to quench star formation in massive galaxies. However, the very large dynamic range required to treat the growth of black holes and galaxies in a cosmological context – from the sub-pc scales of the BH accretion disc to the super-Mpc scales of large-scale structure – means that numerical techniques such as these will likely need to be supplemented by semi-analytic or subgrid methods for some time to come.

Our approach is in many respects very similar in spirit to that of Sijacki et al. (2007), although of course we are forced to implement both modes of AGN feedback in an even more schematic manner because we are using a semi-analytic model rather than a numerical simulation. We adopt fairly standard semi-analytic treatments of the growth of DM haloes via accretion and mergers, radiative cooling of gas, star formation, SN feedback and chemical evolution. We then adopt the picture of self-regulated black hole growth and bright mode feedback in mergers discussed in Section 1.1, and implement these processes in our model using the results extracted from the merger simulations described above. We assume that the radio mode is fuelled instead by hot gas in quasi-hydrostatic haloes, and that the accretion rate is described by Bondi accretion from an isothermal cooling flow as proposed by Nulsen & Fabian (2000, hereafter NF00). We calibrate the heating efficiency of the associated radio jets against direct observations of bubble energetics in clusters.

A number of authors have previously explored the formation of black holes and AGN in the context of CDM-based semi-analytic models of varying complexity (Efstathiou & Rees 1988; Kauffmann & Haehnelt 2000; Wyithe & Loeb 2002; Volonteri, Haardt & Madau 2003; Bromley, Somerville & Fabian 2004; Scannapieco & Oh 2004; Volonteri & Rees 2005), and recently several studies have also investigated the impact of AGN feedback on galaxy formation using such models (Bower et al. 2006; Cattaneo et al. 2006; Croton et al. 2006; Kang, Jing & Silk 2006; Menci et al. 2006; Schawinski et al. 2006; Monaco, Fontanot & Taffoni 2007). The models that we present here differ from previous studies of which we are aware, in two main respects: (1) we implement detailed modelling of self-regulated black hole growth and bright mode feedback based on an extensive suite of numerical simulations of galaxy mergers and (2) we adopt a simple but physical model for radio mode accretion and heating, and calibrate our model against direct observations of accretion rates and radio jet heating efficiencies. We present a broader and more detailed comparison with observations than previous works, and highlight some remaining problems that have not previously been emphasized. As well, unlike most previous studies, we calibrate our models and make our comparisons in terms of ‘physical’ galaxy properties such as stellar mass and SFR, which can be estimated from observations, rather than casting our results in terms of observable properties such as luminosities and colours. Our results are therefore less sensitive to the details of dust and stellar population modelling, and easier to interpret in physical terms.

The goals of this paper are to present our new models in detail, and to test and document the extent to which they reproduce basic galaxy observations at $z = 0$ and the global cosmic histories of the main baryonic components of the Universe. The structure of the rest of this paper is as follows. In Section 2, we describe the ingredients of our models and provide a table of all of the model parameters. In Section 3, we present predictions for key properties of galaxies at $z \sim 0$ and for the global history of the main baryonic components

Table 1. Summary of cosmological parameters.

Parameter	Description	C- Λ CDM	WMAP3
Cosmological parameters			
Ω_m	Present-day matter density	0.30	0.2383
Ω_Λ	Cosmological constant	0.70	0.7617
H_0	Hubble parameter (km s ⁻¹ Mpc ⁻¹)	70.0	73.2
f_b	Cosmic baryon fraction	0.14	0.1746
σ_8	Power spectrum normalization	0.9	0.761
n_s	Slope of primordial power spectrum	1.0	0.958

of the Universe: star formation, evolved stars, cold gas, metals and black holes. We conclude in Section 4.

2 MODEL

Our model is based on the semi-analytic galaxy formation code described in Somerville & Primack (1999, SP99) and Somerville et al. (2001, hereafter SPF01), with several major updates and important new ingredients, which we describe in detail here. Unless specified otherwise, we adopt the ‘concordance’ Λ CDM model (C- Λ CDM), with the parameters given in Table 1, which has been used in many recent semi-analytic studies of galaxy formation. In Section 3.5, we also consider a model that uses the set of parameters obtained from the three year results of the *Wilkinson Microwave Anisotropy Probe* (WMAP) by Spergel et al. (2007), which are also specified in Table 1. We refer to this as the ‘WMAP3’ model. We assume a universal Chabrier stellar initial mass function (IMF; Chabrier 2003), and where necessary we convert all observations used in our comparisons to be consistent with this IMF. In Table 2, we provide a summary of the galaxy formation parameters and the section in which they are defined.

2.1 Dark matter haloes, merger trees and substructure

We compute the number of ‘root’ DM haloes as a function of mass at a desired output redshift using the model of Sheth & Tormen (1999), which has been shown to agree well with numerical simulations. Then, for each ‘root’ halo of a given mass M_0 and at a given output redshift, we construct a realization of the merger history based on the method described in Somerville & Kolatt (1999, hereafter SK99). We have introduced a modification to the SK99 algorithm, which we find leads to better agreement with N -body simulations. We choose the time-step Δt by requiring that the *average* number of progenitors \bar{N}_p be close to two, by inverting the equation for $N_{\text{prog}}(M_0, \Delta t)$ (see SK99). We then select progenitors as described in SK99, but do not allow the number of progenitors to exceed $\bar{N}_p + \sqrt{\bar{N}_p} + 1$. We follow halo merging histories down to a minimum progenitor mass of $10^{10} M_\odot$, and our smallest ‘root’ haloes have a mass of $10^{11} M_\odot$. We have also implemented our models within N -body based merger trees, and do not find any significant changes to our results.

We assign two basic properties to every DM halo in each of our merger trees: the angular momentum or spin parameter, and the concentration parameter, which describes the matter density profile. We express the angular momentum in terms of the dimensionless spin parameter $\lambda \equiv J_h |E_h|^{1/2} G^{-1} M_{\text{vir}}^{-5/2}$ (Peebles 1969), where E_h is the total energy of the halo and M_{vir} is the virial mass. Numerical N -body simulations have demonstrated that λ is uncorrelated with the halo’s mass and concentration (Bullock et al. 2001a; Macciò et al. 2007) and does not evolve with redshift. The distribution of λ is lognormal, with mean $\bar{\lambda} = 0.05$ and width $\sigma_\lambda = 0.5$ (Bullock

et al. 2001a). We assign each top-level halo a value of λ by selecting values randomly from this distribution, assuming that it is not correlated with any other halo properties or with redshift. The halo at the next stage of the merger tree inherits the spin parameter of its largest progenitor.

We assume that the initial density profile of each halo is described by the Navarro–Frenk–White (NFW) form (Navarro, Frenk & White 1997), and compute the characteristic concentration parameter c_{NFW} for the appropriate mass and redshift using a fitting formula based on numerical simulations (Bullock et al. 2001b). We adopt the updated normalization of $c_{\text{NFW}}(M_{\text{vir}})$ from Macciò et al. (2007). We neglect the scatter in c_{NFW} at fixed mass, as well as the known correlation between c_{NFW} and halo merger history (Wechsler et al. 2002).

At each stage in the merging hierarchy, one or more haloes merge together to form a new, virialized DM halo. The merged haloes (hereafter referred to as ‘subhaloes’) and their galaxies, however, can survive and continue to orbit within the potential well of the parent DM halo for some time. The time it takes for the satellite to lose all of its angular momentum due to dynamical friction and merge with the central galaxy is typically modelled with some variant of the Chandrasekhar formula (see e.g. section 2.8 of Somerville & Primack 1999). Here, we use an updated version of this formula from Boylan-Kolchin, Ma & Quataert (2008), which accounts for the tidal mass-loss of subhaloes as they orbit within the host halo, as well as the dependence on the energy and angular momentum of the orbit. Because the merger time is proportional to $M_{\text{host}}/M_{\text{sat}}$, accounting for this mass-loss increases the time it takes for small-mass satellites to merge.

These satellites may eventually lose so much of their mass that they become tidally disrupted. Based on the results of Taylor & Babul (2004) and Zentner & Bullock (2003), we assume that satellites lose ~ 30 – 40 per cent of their mass per orbital period, and that when the mass has been stripped down to the mass within the NFW scale radius $r_s \equiv r_{\text{vir}}/c_{\text{NFW}}$, we consider the satellite to be tidally destroyed. Subhaloes that survive until they reach the centre of the parent halo are assumed to merge with the central object. Subhaloes that are tidally destroyed before they can merge are assumed to contribute their stars to a ‘diffuse stellar component’, which may be associated with the stellar halo or the intracluster light. We have verified that our model reproduces the conditional multiplicity function of subhaloes over the relevant range of host halo masses. Details and tests of our new algorithm for the treatment of substructure will be presented in Maulbetsch et al. (in preparation).

2.2 Cooling

The rate of gas condensation via atomic cooling is computed based on the model originally proposed by White & Frenk (1991), and utilized in various forms in virtually all semi-analytic models. Here

Table 2. Summary of the galaxy formation parameters in our ‘fiducial’ model. We also specify the section in the paper where a more detailed definition of each set of parameters can be found, and whether the parameter is considered to be fixed based on direct observations or numerical simulations (F), or adjusted to match observations (A).

Parameter	Description	Fiducial value	Fixed/adjusted
Photoionization squelching (Section 2.3)			
$z_{\text{overlap}}, z_{\text{re-ionize}}$	Redshift of overlap/re-ionization	11, 10	F
Quiescent star formation (Section 2.5.1)			
A_{Kenn}	Normalization of Kennicutt law ($M_{\odot} \text{ yr}^{-1} \text{ kpc}^{-2}$)	8.33×10^{-5}	A
N_{K}	Power-law index in Kennicutt law	1.4	F
χ_{gas}	Scale radius of gas disc, relative to stellar disc	1.5	A
Σ_{crit}	Critical surface density for star formation ($M_{\odot} \text{ pc}^{-2}$)	6.0	A
Burst star formation (Section 2.5.2)			
μ_{crit}	Critical mass ratio for burst activity	0.1	F
$\epsilon_{\text{burst},0}$	Burst efficiency for 1:1 merger	Equation (9)	F
γ_{burst}	Dependence of burst efficiency on mass ratio	Equation (8)	F
τ_{burst}	Burst time-scale	Equation (10)	F
Merger remnants and morphology (Section 2.6)			
f_{sph}	Fraction of stars in spheroidal remnant	Equation (11)	A
f_{scatter}	Fraction of scattered satellite stars	0.4	A
SN feedback (Section 2.7)			
ϵ_{SN}^0	Normalization of reheating function	1.3	A
α_{rh}	Power-law slope of reheating function	2.0	A
V_{eject}	Velocity scale for ejection of reheated gas (km s^{-1})	120	A
$\chi_{\text{re-infall}}$	Time-scale for re-infall of ejected gas	0.1	A
Chemical evolution (Section 2.8)			
y	Chemical yield (solar units)	1.5	A
R	Recycled fraction	0.43	F
Black hole growth (Section 2.9)			
η_{rad}	Efficiency of conversion of rest mass to radiation	0.1	F
M_{seed}	Mass of seed BH (M_{\odot})	100	F
$f_{\text{BH,final}}$	Scaling factor for mass of BH at end of merger	2.0	A
$f_{\text{BH,crit}}$	Scaling factor for ‘critical mass’ of BH	0.4	F
AGN-driven winds (Section 2.1.0)			
ϵ_{wind}	Effective coupling factor for AGN-driven winds	0.5	F
Radio mode feedback (Section 2.1.1)			
κ_{radio}	Normalization of ‘radio mode’ BH accretion rate	3.5×10^{-3}	A
κ_{heat}	Coupling efficiency of radio jets with hot gas	1.0	F

we use a slightly different implementation of the cooling model than that used in SP99 and subsequent papers, which we find is numerically better behaved. We first compute the ‘cooling time’, which is the time required for the gas to radiate away all of its energy, assuming that it all starts out at the virial temperature:

$$t_{\text{cool}} = \frac{(3/2)\mu m_p kT}{\rho_g(r)\Lambda(T, Z_h)}. \quad (1)$$

Here, μm_p is the mean molecular mass, T is the virial temperature $T_{\text{vir}} = 35.9[V_{\text{vir}}/(\text{km s}^{-1})]^2 \text{ K}$, $\rho_g(r)$ is the radial density profile of the gas, $\Lambda(T, Z_h)$ is the temperature and metallicity-dependent cooling function (Sutherland & Dopita 1993), and Z_h is the metallicity of the hot halo gas. We assume that the gas density profile is described by that of a singular isothermal sphere: $\rho_g(r) = m_{\text{hot}}/(4\pi r_{\text{vir}} r^2)$. Substituting this expression for $\rho_g(r)$, we can solve for the cooling radius r_{cool} , which is the radius within which all of the gas can cool within a time t_{cool} . Writing the expression for the mass within r_{cool} , and differentiating, we obtain the rate at which gas can cool:

$$\frac{dm_{\text{cool}}}{dt} = \frac{1}{2} m_{\text{hot}} \frac{r_{\text{cool}}}{r_{\text{vir}}} \frac{1}{t_{\text{cool}}}. \quad (2)$$

There are various possible choices for the cooling time t_{cool} . Some early works used the Hubble time, $t_{\text{cool}} = t_{\text{H}}$ (e.g. Kauffmann et al.

1993). In our earlier models (e.g. SP99), we used the time since the last halo major merger t_{mrg} , defined as a merger in which the halo grows in mass by at least a factor of 2. Here, we follow Springel et al. (2001) and Croton et al. (2006) and assume that the cooling time is equal to the halo dynamical time, $t_{\text{cool}} = t_{\text{dyn}} = r_{\text{vir}}/V_{\text{vir}}$. Note that because in general $t_{\text{dyn}} < t_{\text{mrg}} < t_{\text{H}}$, and the cooling rate $dm_{\text{cool}}/dt \propto t_{\text{cool}}^{-1/2}$, the choice $t_{\text{cool}} = t_{\text{dyn}}$ results in higher cooling rates than assuming $t_{\text{cool}} = t_{\text{H}}$, while using $t_{\text{cool}} = t_{\text{mrg}}$ produces intermediate results.

It can occur that $r_{\text{cool}} > r_{\text{vir}}$, indicating that the cooling time is shorter than the dynamical time. In this case, we assume that the cooling rate is given by the rate at which gas can fall into the halo, which is governed by the mass accretion history.

We note that there are several rather arbitrary choices that must be made in any semi-analytic cooling model – for example, the profile of the hot gas and whether it is ‘reset’, and the time to which the cooling time is compared (see above) – and different groups tend to make slightly different choices for these ingredients. Changing these ingredients in reasonable ways leads to overall variations in the cooling rates (changes in the redshift and halo mass dependence tend to be small) of at most a factor of 2–3. These differences are then typically compensated by adjusting the SN feedback and/or AGN feedback parameters.

We have adopted choices similar to those of Croton et al. (2006), in part to facilitate comparison with their results, and also because it has been shown that this recipe produces good agreement with the cooling rates and accumulation of gas in fully 3D hydrodynamic simulations (Yoshida et al. 2002) without star formation, SN feedback, or chemical enrichment. We have also compared our results with the cooling rates presented by Kereš et al. (2005), and find good agreement.

Recently, studies based on 1D and 3D hydrodynamic simulations (Birnboim & Dekel 2003; Kereš et al. 2005) have highlighted a distinction between gas which is accreted in a ‘cold flow’ mode, in which the gas particles are never heated much above $\sim 10^4$ K, and a ‘hot flow’ mode in which gas is first shock heated to close to the virial temperature of the halo, forming a quasi-hydrostatic halo, and then cools in a manner similar to a classical cooling flow. The possible importance of distinguishing between gas flows occurring in the regime $t_{\text{cool}} < t_{\text{ff}}$ versus $t_{\text{cool}} > t_{\text{ff}}$ (where t_{ff} is the free-fall time) has been highlighted many times in the literature (Binney 1977; Rees & Ostriker 1977; Silk 1977; White & Frenk 1991). Although other criteria have been proposed (see Croton et al. 2006), we will identify gas cooling which occurs in time-steps in which $r_{\text{cool}} > r_{\text{vir}}$ as ‘cold mode’ and the reverse ($r_{\text{cool}} < r_{\text{vir}}$) as ‘hot mode’. This distinction will be relevant later, when we begin to consider the impact of heating by AGN-driven radio jets.

As in most semi-analytic models, we assume that all new cold gas is accreted by the central galaxy in the halo. Because of this, satellite galaxies tend to consume their gas and become red, non-star-forming and gas-poor. Realistically, satellite galaxies can probably retain their hot gas haloes, and thus receive new cold gas, for some time after they merge with another halo. This aspect of the modelling should be improved; however, for the moment, we simply keep this problem in mind, and in some cases restrict our analysis to central galaxies.

2.3 Photoionization squelching

Photoionization heating may ‘squelch’ or suppress the collapse of gas into small-mass haloes (Efstathiou 1992; Quinn et al. 1996; Thoul & Weinberg 1996). This mechanism may play an important role in reconciling the (large) number of small-mass satellite haloes predicted by CDM with the observed number of satellite galaxies in the Local Group (Benson et al. 2002; Somerville 2002). Gnedin (2000, hereafter G00) showed that the fraction of baryons that can collapse into haloes of a given mass in the presence of a photoionizing background can be described in terms of the ‘filtering mass’ M_{F} . Haloes less massive than M_{F} contain fewer baryons than the universal average. G00 parametrized the collapsed baryon fraction as a function of redshift and halo mass with the expression

$$f_{\text{b,coll}}(z, M_{\text{vir}}) = \frac{f_{\text{b}}}{[1 + 0.26M_{\text{F}}(z)/M_{\text{vir}}]^3}, \quad (3)$$

where f_{b} is the universal baryon fraction and M_{vir} is the halo virial mass.

The filtering mass is a function of redshift, and this function depends on the re-ionization history of the Universe. Kravtsov, Gnedin & Klypin (2004b) provide fitting formulae for the filtering mass in the simulations of G00, parametrized according to the redshift at which the first H II regions begin to overlap (z_{overlap}) and the redshift at which most of the medium is re-ionized (z_{reion}). In the simulations of G00, re-ionization occurs fairly late ($z_{\text{overlap}} = 8$, $z_{\text{reion}} = 7$). Recent results from the *WMAP* satellite, however, suggest an earlier epoch of re-ionization, $z_{\text{reion}} \gtrsim 10$ (Spergel et al. 2007). We make use

of the fitting functions (B2) and (B3) from appendix B of Kravtsov et al. (2004b) to compute the initial fraction of baryons that can collapse as a function of halo mass and redshift, with $z_{\text{overlap}} = 11$ and $z_{\text{reion}} = 10$.

As shown by Somerville (2002) using a similar treatment of photoionization squelching, we find that our model reproduces the luminosity function of satellite galaxies in the Local Group (Macciò et al., in preparation).

2.4 Disc sizes

When gas cools, it is assumed to initially settle into a thin exponential disc, supported by its angular momentum. We assume that the gas has acquired angular momentum before its collapse, along with the DM, via tidal torques (Peebles 1969). Given the halo’s concentration parameter c_{NFW} , spin parameter λ and the fraction of baryons in the disc f_{disc} , we can use angular momentum conservation arguments to compute the scale radius of the exponential disc after collapse. We include the ‘adiabatic contraction’ of the halo due to the gravitational force of the collapsing baryons. Our approach is based on work by Blumenthal et al. (1986), Flores et al. (1993) and Mo, Mao & White (1998, hereafter MMW98), and is described in detail in Somerville et al. (2008, hereafter S08). In S08, we showed that this model produces good agreement with the observed radial sizes of discs as a function of stellar mass both locally and out to $z \sim 2$.

2.5 Star formation

2.5.1 Quiescent star formation

During the ‘quiescent’ phase of galaxy evolution (i.e. in undisturbed discs) we adopt a star formation recipe based on the empirical Schmidt–Kennicutt law (Kennicutt 1989, 1998). The SFR density (per unit area) is given by

$$\dot{\Sigma}_{\text{SFR}} = A_{\text{Kenn}} \Sigma_{\text{gas}}^{N_{\text{K}}}, \quad (4)$$

where $A_{\text{Kenn}} = 1.67 \times 10^{-4}$, $N_{\text{K}} = 1.4$, Σ_{gas} is the surface density of cold gas in the disc (in units of $M_{\odot} \text{pc}^{-2}$), and Σ_{SFR} has units of $M_{\odot} \text{yr}^{-1} \text{kpc}^{-2}$. The normalization quoted above is appropriate for a Chabrier IMF, and has been converted from the value given in Kennicutt (1998), which was based on a Salpeter IMF.

We assume that the gas profile is also an exponential disc, with a scale length proportional to the scalelength of the stellar disc: $r_{\text{gas}} = \chi_{\text{gas}} r_{\text{disc}}$, where the stellar scalelength r_{disc} is determined as described in Section 2.4. We adopt $\chi_{\text{gas}} = 1.5$, which yields average gas surface densities in the range $\sim 4\text{--}60 M_{\odot} \text{pc}^{-2}$, consistent with the observations of Kennicutt (1998). This value is also consistent with observations of the radial extent of H I gas in spiral galaxies (Broeils & Rhee 1997).

We further adopt a critical surface density threshold Σ_{crit} , and assume that only gas lying at surface densities above this value is available for star formation. We can then compute the radius within which the gas density exceeds the critical value:

$$r_{\text{crit}} = -\ln \left[\frac{\Sigma_{\text{crit}}}{\Sigma_0} \right] r_{\text{gas}}, \quad (5)$$

where $\Sigma_0 \equiv m_{\text{cold}}/(2\pi r_{\text{gas}}^2)$. The fraction of the total gas supply that is eligible for star formation is then

$$f_{\text{gas}}(r < r_{\text{crit}}) = 1 - (1 + r_{\text{crit}}/r_{\text{gas}}) \frac{\Sigma_{\text{crit}}}{\Sigma_0} \quad (6)$$

and the total SFR is

$$\begin{aligned} \dot{m}_* &= \int_0^{r_{\text{crit}}} \dot{\Sigma}_{\text{SFR}} 2\pi r \, dr \\ &= \frac{2\pi A_{\text{K}} \Sigma_0 N_{\text{K}} r_{\text{gas}}^2}{N_{\text{K}}^2} \\ &\quad \times \left[1 - \left(1 + \frac{N_{\text{K}} r_{\text{crit}}}{r_{\text{gas}}} \right) \exp(-N_{\text{K}} r_{\text{crit}}/r_{\text{gas}}) \right]. \end{aligned}$$

Schaye (2004) investigated star formation thresholds in models of isolated self-gravitating discs embedded in DM haloes, containing metals and dust, and exposed to an ultraviolet (UV) background. They found that the gas was able to form a cold interstellar phase only above a critical surface density threshold of $\Sigma_{\text{crit}} \sim 3\text{--}10 M_{\odot} \text{pc}^{-2}$, which is consistent with observations of SF thresholds in spiral galaxies (Martin & Kennicutt 2001). We find that adopting a value of $\Sigma_{\text{crit}} = 6 M_{\odot} \text{pc}^{-2}$ produces good agreement with the observations of global SFR versus gas density of Kennicutt (1998), and also with observed gas fractions as a function of stellar mass.

We allow the normalization of the star formation law A_{Kenn} to be adjusted as a free parameter. We find that using a value of $A_{\text{Kenn}} = 8.33 \times 10^{-5}$, a factor of 2 lower than the one measured by Kennicutt (1998) gives good agreement with observed SFRs and gas fractions as a function of stellar mass. We adopt the observed value for the slope of the SFR law, $N_{\text{K}} = 1.4$.

We account for the mass-loss from stars (recycled gas) using the instantaneous recycling approximation. Thus, for an instantaneous SFR \dot{m} , we form a mass $dm_* = (1 - R)\dot{m} dt$ of long-lived stars in a time-step dt . We adopt a recycled fraction $R = 0.43$, appropriate for a Chabrier IMF (Bruzual & Charlot 2003).

2.5.2 Merger-driven starbursts

As in SPF01, we parametrize the efficiency of star formation in a merger-triggered ‘burst’ mode as a function of the mass ratio of the merging pair. This is supported both by observations of star formation enhancement in galaxy pairs (Woods & Geller 2007) and by numerical simulations of galaxy mergers (Cox et al. 2008). However, first, an important question arises: which quantity should we use for the mass ratio? Many previous works have either used the ratio of the virial masses of the two DM haloes, or else the baryonic masses of the two galaxies. Because the ratio of baryons to DM can vary by several orders of magnitude across haloes of different masses, and moreover is a systematic function of halo mass (see Section 3.1), we find that our results can depend quite sensitively on this choice. Moreover, the baryonic mass ratio is sensitive to the modelling of SN and AGN feedback.

When we consider that the simulation results clearly indicate that the efficiency of the starburst is mainly determined by the strength of the torques during the later stages of the merger, it is clear that what should be relevant is the *total* mass (baryons and DM) in the *central* parts of the galaxies. Therefore we define $m_{\text{core}} = M_{\text{DM}}(r < 2r_s)$, i.e. the DM mass within twice the characteristic NFW scale radius $r_s \equiv r_{\text{vir}}/c_{\text{NFW}}$, assuming that the DM follows an NFW profile. For a Milky Way sized halo ($M_{\text{vir}} \sim 2 \times 10^{12} M_{\odot}$), $r_s \sim 27$ kpc and so $2 r_s$ corresponds to about 60 kpc, close to the scale that we expect to be relevant. We then define the mass ratio $\mu \equiv (m_{\text{core},1} + m_{\text{bar},1})/(m_{\text{core},2} + m_{\text{bar},2})$, i.e. as the ratio of the DM ‘core’ plus the total baryonic mass (stars plus cold gas) of the smaller to the larger galaxy.

Now defining e_{burst} as the fraction of the total cold gas reservoir in the galaxy that is consumed by the burst, we parametrize the burst

efficiency via

$$e_{\text{burst}} = e_{\text{burst},0} \mu^{\gamma_{\text{burst}}}. \quad (7)$$

This functional form has been shown to describe well the scaling of burst efficiency with merger mass ratio in hydrodynamic simulations of galaxy mergers (Cox et al. 2008, hereafter C08). Again based on C08, we assume that mergers with mass ratios below 1:10 do not produce bursts, i.e. $e_{\text{burst}} = 0$ for $\mu < 0.1$.

Numerical studies (Mihos & Hernquist 1994; C08) have furthermore shown that the burst efficiency in *minor* mergers ($\mu \lesssim 0.25$) depends on the bulge-to-total ratio of the progenitor galaxies, because the presence of a bulge stabilizes the galaxy and reduces the efficiency of the burst. To reflect the joint dependence on merger mass ratio and bulge fraction, we adopt the results of C08:

$$\gamma_{\text{burst}} = \begin{cases} 0.61 & B/T \leq 0.085, \\ 0.74 & 0.085 < B/T \leq 0.25, \\ 1.02 & B/T > 0.25, \end{cases} \quad (8)$$

where B/T is the ratio of the stellar mass in the spheroidal component to the total stellar mass (disc plus spheroid) in the larger progenitor galaxy at the beginning of the merger.

We have studied the burst efficiency $e_{\text{burst},0}$ and burst time-scale τ_{burst} in equal mass mergers in a large suite of numerical simulations containing stellar feedback as well as feedback from energy released by accretion on to a central black hole (Robertson et al. 2006b). We find that the burst efficiency can be fitted by

$$e_{\text{burst},0} = 0.60 [V_{\text{vir}}/(\text{km s}^{-1})]^{0.07} (1 + q_{\text{EOS}})^{-0.17} \times (1 + f_g)^{0.07} (1 + z)^{0.04} \quad (9)$$

with a scatter of 4.9 per cent, and the burst time-scale (assuming a double exponential form for the SFR peak) is fitted by

$$\tau_{\text{burst}} = 191 \text{ Gyr} [V_{\text{vir}}/(\text{km s}^{-1})]^{-1.88} (1 + q_{\text{EOS}})^{2.58} (1 + f_g)^{-0.74} (1 + z)^{-0.16} \quad (10)$$

with a logarithmic scatter of 0.36. Here, V_{vir} is the virial velocity of the progenitor galaxies, q_{EOS} is the effective equation of state of the gas (see Robertson et al. 2006b), $f_g \equiv m_{\text{cold}}/(m_{\text{cold}} + m_{\text{star}})$ is the cold gas fraction in the disc, and z is the redshift for which the progenitor disc models were constructed. It is important to note that the simulations used to obtain these fitting functions span the range $V_{\text{vir}} = 60\text{--}500 \text{ km s}^{-1}$, $q_{\text{EOS}} = 0.25\text{--}1$, $f_g = 0.01\text{--}0.8$, and $z = 0\text{--}6$. The results of the fitting formulae should be used with caution outside of this range of values for the input parameters.

The parameter q_{EOS} can be thought of as parametrizing the multiphase nature of the ISM, such that $q_{\text{EOS}} = 0$ corresponds to an isothermal gas, and $q_{\text{EOS}} = 1$ corresponds to the fully pressurized multiphase ISM. Increasing q_{EOS} (or adopting a ‘stiffer’ equation of state) increases the dynamical stability of the gas, and suppresses the starburst. Thus larger values of q_{EOS} give smaller values of $e_{\text{burst},0}$ and larger values of τ_{burst} (because the burst is more extended). In this paper we adopt a value corresponding to a stiff equation of state, $q_{\text{EOS}} = 1$.

For the burst efficiency, we can see that the only significant dependence on these parameters is on the equation of state q_{EOS} . The burst time-scale is more sensitive to other parameters (as also found by C08), and in particular has quite a strong dependence on V_{vir} . For our adopted fiducial value of $q_{\text{EOS}} = 1$, the typical value of the burst efficiency is $e_{\text{burst}} \sim 0.8$, and the burst time-scale (exponential decline time) for a Milky Way sized galaxy ($V_{\text{vir}} \sim 130 \text{ km s}^{-1}$) is $\tau_{\text{burst}} \sim 100 \text{ Myr}$.

We now parametrize the ‘burst’ mode of star formation as $\dot{m}_* = m_{\text{burst}}/\tau_{\text{burst}}$. At the beginning of the merger, we allocate a reservoir of ‘burst fuel’ $m_{\text{burst}} = e_{\text{burst}} m_{\text{cold}}$, where m_{cold} is the combined cold gas from both of the progenitor galaxies. The burst continues until this fuel is exhausted, and in the absence of new sources of fuel, the burst SFR will decline exponentially, with exponential decline time τ_{burst} . However, particularly in the early Universe, it can frequently happen that a new merger occurs while a burst from an earlier merger is still going on. In this case, we add the new burst fuel to the reservoir, and assign a new burst time-scale based on the updated galaxy properties. Note that the ‘quiescent’ mode of star formation still goes on as before (the burst efficiencies computed from the simulations have the quiescent star formation subtracted out).

2.6 Merger remnants and morphology

2.6.1 Spheroid formation

Numerical simulations of mergers of galaxy discs have also shown that major mergers $\mu > 0.25$ leave behind a spheroidal remnant, while smaller mass ratio minor mergers ($\mu < 0.25$) tend to just thicken the disc, perhaps driving minor growth of a spheroid via bar instabilities. Most previous semi-analytic models have assumed a sharp threshold in merger mass ratio (e.g. $\mu > f_{\text{ellip}} \simeq 0.25\text{--}0.3$) for determining whether the stars after a merger are placed in a ‘spheroidal’ component or not. However, in reality there will be a continuum, whereby larger mass ratios result in more heating and a transfer of more material to a dynamically hot spheroidal component. To represent this continuum, we define the function

$$f_{\text{sph}} = 1 - \left[1 + \left(\frac{\mu}{f_{\text{ellip}}} \right)^8 \right]^{-1} \quad (11)$$

which determines the fraction of the disc stars that is transferred to the ‘spheroid’ or bulge component following a merger (as with bursts, we assume that mergers with mass ratio $\mu < 0.1$ have no effect).

Thus, consider a merger of two galaxies with bulge masses B_1 and B_2 , and disc masses D_1 and D_2 . The mass of the new bulge will be $B_{\text{new}} = B_1 + B_2 + f_{\text{sph}}(D_1 + D_2)$, and the mass of the surviving disc will be $D_{\text{new}} = (1 - f_{\text{sph}})(D_1 + D_2)$. All new stars formed in the burst mode are also deposited in the spheroid component.

In this paper, we assume that all spheroid growth is connected with mergers. That is, we do not consider formation of spheroids via disc instabilities. We have experimented with including spheroid formation via disc instabilities, and find that it has only a minor impact on the results presented here.

2.6.2 Formation of diffuse stellar haloes

There is now considerable observational evidence for spatially extended stellar components surrounding brightest group and cluster galaxies (e.g. Gonzalez, Zabludoff & Zaritsky 2005; Zibetti et al. 2005). It is now thought that these ‘diffuse stellar haloes’ (DSH) originated from tidally disrupted merging satellites and/or scattering of stars during major mergers (Murante et al. 2004; Monaco et al. 2006; Conroy, Wechsler & Kravtsov 2007; Murante et al. 2007; Purcell, Bullock & Zentner 2007). In our models, the stars from all satellites that are deemed to be tidally destroyed before they merge (according to the criteria described in Section 2.1), are deposited in a DSH component. In addition, when two galaxies merge, we

assume that a fraction f_{scatter} of the stars from the satellite may be scattered into the DSH (thus the galaxy’s mass following a merger increases by $(1 - f_{\text{scatter}})m_{\text{star,sat}}$, where $m_{\text{star,sat}}$ is the stellar mass of the merging satellite).

2.7 Supernova feedback

Cold gas may be ejected from the galaxy by winds driven by SN feedback. The rate of reheating of cold gas is given by

$$\dot{m}_{\text{rh}} = \epsilon_0^{\text{SN}} \left(\frac{200 \text{ km s}^{-1}}{V_{\text{disc}}} \right)^{\alpha_{\text{rh}}} \dot{m}_*, \quad (12)$$

where ϵ_0^{SN} and α_{rh} are free parameters (we expect $\alpha_{\text{rh}} \simeq 2$ for ‘energy-driven’ winds; see e.g. Kauffmann et al. 1993). We take the circular velocity of the disc V_{disc} to be equal to the maximum rotation velocity of the DM halo, V_{max} .

The heated gas is either trapped within the potential well of the DM halo, so deposited in the ‘hot gas’ reservoir, or is ejected from the halo into the ‘diffuse’ intergalactic medium (IGM). The fraction of reheated gas that is ejected from the halo is given by

$$f_{\text{eject}}(V_{\text{vir}}) = [1.0 + (V_{\text{vir}}/V_{\text{eject}})^{\alpha_{\text{eject}}}]^{-1}, \quad (13)$$

where $\alpha_{\text{eject}} = 6$ and V_{eject} is a free parameter in the range $\simeq 100\text{--}150 \text{ km s}^{-1}$.

We keep track of this ejected gas in a ‘diffuse gas reservoir’, which recollapses into the halo in later time-steps and once again becomes available for cooling. Following Springel et al. (2001) and De Lucia, Kauffmann & White (2004), we model the rate of re-infall of ejected gas by

$$\dot{m}_{\text{re-infall}} = \chi_{\text{re-infall}} \left(\frac{m_{\text{eject}}}{t_{\text{dyn}}} \right), \quad (14)$$

where $\chi_{\text{re-infall}}$ is a free parameter, m_{eject} is the mass of ejected gas in the ‘diffuse reservoir’, and $t_{\text{dyn}} = r_{\text{vir}}/V_{\text{vir}}$ is the dynamical time of the halo.

Varying $\chi_{\text{re-infall}}$ is degenerate with variations in the other SN feedback parameters, ϵ_0^{SN} , α_{rh} and V_{eject} . For larger values of $\chi_{\text{re-infall}}$, the ejected gas is reincorporated in the halo more quickly and tends to cool rapidly, so the SN feedback must be made more efficient in order to retain good agreement with the abundance of low-mass galaxies. On the other hand, if $\chi_{\text{re-infall}} = 0$ (ejected gas is never re-accreted), then the baryon fractions in clusters are too low. We have chosen to adopt the minimal value of $\chi_{\text{re-infall}}$ that allows us to fit the cluster baryon fractions and the mass function of low-mass galaxies simultaneously.

2.8 Chemical evolution

We track the production of metals using a simple approach that is commonly adopted in semi-analytic models (see e.g. Somerville & Primack 1999; Cole et al. 2000; De Lucia et al. 2004). In a given time-step, where we create a parcel of new stars dm_* , we also create a mass of metals $dM_Z = y dm_*$, which we assume to be instantaneously mixed with the cold gas in the disc. The yield y is assumed to be constant, and is treated as a free parameter in our model. We track the mean metallicity of the cold gas Z_{cold} , and when we create a new parcel of stars they are assumed to have the same metallicity as the mean metallicity of the cold gas in that time-step. SN feedback ejects metals from the disc, along with cold gas. These metals are either mixed with the hot gas in the halo, or ejected from the halo into the ‘diffuse’ IGM, in the same proportion as the reheated cold gas. The ejected metals in the ‘diffuse gas’

reservoir are also re-accreted into the halo in the same manner as the gas (see Section 2.7).

Throughout this paper, the yield y and all metallicities are given in solar units, which we take to be $Z_{\odot} = 0.02$. Although this formally represents the total metallicity, we note that as we track only the enrichment associated with Type II SNe, our metallicity estimates probably correspond more closely with α -type elements.

2.9 The growth of supermassive black holes

We assume that every top-level halo in our merger tree contains a seed black hole with mass M_{seed} . Typically, we assume $M_{\text{seed}} \simeq 100 M_{\odot}$, however, we have checked that the results presented here are not sensitive to this choice for a range of values $M_{\text{seed}} \sim 100\text{--}10^4 M_{\odot}$. Black holes of approximately this mass could be left behind as remnants of massive Population III stars (e.g. Abel, Bryan & Norman 2002), or could form via direct core collapse. In our models, all ‘bright mode’ accretion on to SMBHs is triggered by galaxy–galaxy mergers, and we assume that this mode of BH accretion is regulated, and eventually halted, by feedback from the BH itself. Our treatment of BH growth and AGN activity is closely based on an analysis of a large suite of numerical hydrodynamic simulations including BH growth and feedback (Cox et al. 2006b; Robertson et al. 2006a,b,c; Hopkins et al. 2007a), which utilize the methodology developed in Di Matteo et al. (2005) and Springel et al. (2005b). We now briefly summarize the results of those simulations and the manner in which we implement them in our semi-analytic model.

In the merger simulations, as the galaxies near their final coalescence, the accretion on to the BH rises to approximately the Eddington rate. This rapid accretion continues until the energy being deposited into the ISM in the central region of the galaxy is sufficient to significantly offset and eventually halt accretion via a pressure-driven outflow. Di Matteo et al. (2005) and Robertson et al. (2006c) found that the merger simulations naturally produced black holes and spheroidal remnants that obeyed the observed BH mass versus spheroid mass relationship. The normalization of the relationship depends on the fraction of the AGN’s energy that is coupled to the ISM, and was chosen to reproduce the normalization of the observed relation. Based on further analysis of these simulations, Hopkins et al. (2007a) suggested that the BH mass is largely determined by the depth of the potential well in the central regions of the galaxy. As shown by Robertson et al. (2006b) and Cox et al. (2006b), mergers of progenitor galaxies with higher gas fractions suffer more dissipation, and produce more compact remnants than those with less gas. Therefore, mergers with high gas fractions will produce a remnant with a deeper potential well and a larger BH mass to spheroid mass ratio than gas-poor mergers. This picture predicts that there should be a ‘Black Hole Fundamental Plane’, whereby galaxies with smaller effective radius for their mass host larger mass black holes; there is observational evidence for the existence of such a BH fundamental plane in nearby dormant BH host galaxies (Marconi & Hunt 2003; Hopkins et al. 2007b).

Hopkins et al. (2007a) find that the relationship between progenitor gas fraction and the final BH mass to spheroid stellar mass ratio at the end of the merger obtained in their simulations can be parametrized as

$$\log(M_{\text{BH}}/M_{\text{sph}}) = -3.27 + 0.36 \operatorname{erf}[(f_{\text{gas}} - 0.4)/0.28] \quad (15)$$

with a scatter around this relationship of $\sim 0.2\text{--}0.3$ dex, corresponding to the expected range of orbital parameters.

In our semi-analytic model, at the beginning of each merger above a critical mass ratio ($\mu_{\text{crit}} \sim 0.1$), we compute the expected mass of the spheroid that will be left behind at the end of the merger, where we assume that all of the new stars formed in the burst mode will end up in the spheroid, along with the ‘heated’ disc stars specified by equation (11). We then use equation (15), above, to compute $m_{\text{BH,final}}$, the BH mass at the end of the merger, based on the initial ‘effective’ gas fraction $f_{\text{gas,eff}} = (m_{\text{cold},1} + m_{\text{cold},2})/(m_{\text{bar},1} + m_{\text{bar},2})$ (i.e. the sum of the cold gas masses in both galaxies, divided by the sum of their baryonic masses). We allow the value of $m_{\text{BH,final}}$ given by equation (15) to be scaled by an adjustable free parameter $f_{\text{BH,final}}$.

We assume that the BH in the two progenitor galaxies merge rapidly to form a new BH, and that mass is conserved in the BH merger. We allow the BH to grow at the Eddington rate until it reaches a mass $M_{\text{BH,crit}}$, whereupon it enters the ‘blow-out’ phase and begins a power-law decline in the accretion rate, according to the family of light curves defined by Hopkins et al. (2006b). From the simulations, $M_{\text{BH,crit}} = f_{\text{BH,crit}} 1.07 (M_{\text{BH,final}}/10^9 M_{\odot})^{1.1}$, where we introduce the adjustable parameter $f_{\text{BH,crit}}$, which determines how much of the BH growth occurs in the Eddington-limited versus power-law decline (‘blow-out’) phases. When the BH reaches the mass $M_{\text{BH,final}}$, ‘bright mode’ accretion is switched off. If the pre-existing BH is more massive than $M_{\text{BH,crit}}$, it goes straight into the ‘blow-out’ mode until it reaches $M_{\text{BH,final}}$. If the pre-existing BH is more massive than $M_{\text{BH,final}}$, the BH does not grow at all and there is no AGN activity.

2.10 AGN-driven galactic scale winds

In the numerical merger simulations, the energy being released during the rapid growth of the BH also drives powerful galactic-scale winds (Di Matteo et al. 2005; Springel et al. 2005a). We again make use of the simulations to parametrize this process in our semi-analytic model. We start by equating the momentum associated with the radiative energy from the accreting BH with the momentum of the outflowing wind:

$$\frac{\epsilon_{\text{wind}} E_{\text{BH}}}{c} = M_{\text{outflow}} V_{\text{esc}}, \quad (16)$$

where ϵ_{wind} is the effective coupling efficiency, $E_{\text{BH}} = \eta_{\text{rad}} m_{\text{acc}} c^2$, M_{outflow} is the mass of the ejected gas and V_{esc} is the escape velocity of the galaxy. We then obtain the following expression for the mass outflow rate due to the AGN-driven wind:

$$\frac{dM_{\text{out}}}{dt} = \epsilon_{\text{wind}} \eta_{\text{rad}} \frac{c}{V_{\text{esc}}} \dot{m}_{\text{acc}}. \quad (17)$$

We find that this simple formula provides quite a good description of the outflow rates in the simulations, as shown in Fig. 1. We see that equation (17) provides a much better description of the simulation results than simply assuming $M_{\text{outflow}} \propto \dot{m}_{\text{BH}}$.

2.11 Radio mode feedback

In addition to the rapid growth of BH in the merger-fuelled, radiatively efficient ‘bright mode’, we assume that BH also experience a low-Eddington-ratio, radiatively inefficient mode of growth associated with efficient production of radio jets that can heat gas in a quasi-hydrostatic hot halo. We base our fiducial model on the assumption that the ‘radio mode’ is fuelled by Bondi–Hoyle accretion (Bondi 1952):

$$\dot{m}_{\text{Bondi}} = \pi (GM_{\text{BH}})^2 \rho_0 c_s^{-3}, \quad (18)$$

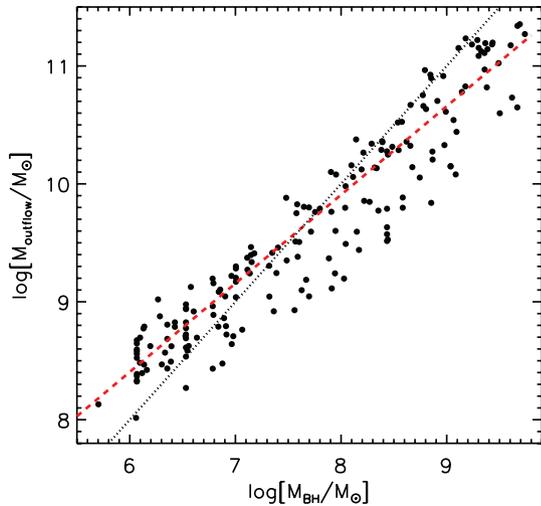


Figure 1. Mass ejected in an outflow as a function of the final BH mass, from the numerical merger simulations. The dashed line shows the scaling predicted by the momentum conservation argument (equation 17), with $\epsilon_{\text{wind}} = 0.5$. The dotted line shows a simple scaling of ejected mass with BH mass, which does not fit the simulation results as well.

where $\rho_0 \equiv \rho(r_A)$ is the density of the gas at the accretion radius r_A , c_s is the sound speed of the gas, and we have assumed an adiabatic index of $\gamma_1 = 5/3$ for the gas. We adopt the isothermal cooling flow solution of NF00, in which thermal instabilities act to maintain the density such that the sound crossing time is of the order of the local cooling time:

$$\frac{r_A}{c_s} = K \frac{3}{2} \frac{\mu m_p kT}{\rho(r_A) \Lambda(T, Z_h)}. \quad (19)$$

Here, $r_A \equiv 2GM_{\text{BH}}/c_s^2$ is the Bondi accretion radius, K is a dimensionless constant which depends on the details of the flow, kT is the temperature of the gas and $\Lambda(T, Z_h)$ is the cooling function. Solving for the density ρ_0 and substituting into equation (18), we obtain

$$\dot{m}_{\text{radio}} = \kappa_{\text{radio}} \left[\frac{kT}{\Lambda(T, Z_h)} \right] \left(\frac{M_{\text{BH}}}{10^8 M_{\odot}} \right), \quad (20)$$

where we have subsumed all constants into the factor κ_{radio} . A similar model has also been considered by Churazov et al. (2005) and Croton et al. (2006).

We can test the validity of this model using recent observations of the central density and temperature of hot X-ray emitting gas in nine nearby elliptical galaxies by Allen et al. (2006, hereafter A06). Deep Chandra observations allowed A06 to obtain measurements or reliable extrapolations of the gas properties within one order of magnitude of the Bondi radius for eight of the systems. Each system also has a measured velocity dispersion, which allows an estimate of the BH mass using the relation of Tremaine et al. (2002). In Fig. 2, we compare the Bondi accretion rates with the quantity $\zeta \equiv (T_7/\Lambda_{23})(m_{\text{BH},8})$, where we define $T_7 \equiv T/10^7$ K, $\Lambda_{23} \equiv \Lambda(T)/(10^{-23} \text{ erg cm}^3 \text{ s}^{-1})$, and $m_{\text{BH},8} \equiv M_{\text{BH}}/10^8 M_{\odot}$. We use the published values of gas density and temperature and the BH mass estimates from A06, and assume that the hot gas has a metallicity of one-third solar. A formal fit gives a slope of 1.23 in ζ , but we see that the NF00 isothermal cooling flow model is quite consistent with the data. In terms of the scaled quantities T_7 , Λ_{23} and $m_{\text{BH},8}$, $\kappa_{\text{radio}} = 2.25 \times 10^{-3}$ provides the best fit to the data.

In our fiducial semi-analytic model, we assume that whenever ‘hot mode’ gas is present in the halo, the central BH accretes at the

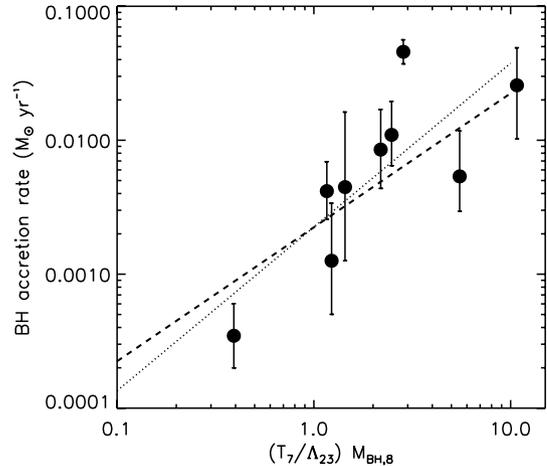


Figure 2. Black hole accretion rate as a function of $\zeta \equiv (T_7/\Lambda_{23})(m_{\text{BH},8})$, from the observational analysis of A06 (filled circles). The dashed line shows the scaling predicted by the NF00 ‘isothermal cooling flow’ model (see text), while the dotted line shows a linear fit to the data points.

rate given by equation (18), but we allow κ_{radio} to be adjusted as a free parameter. We then assume that the energy that effectively couples to and heats the hot gas is given by $L_{\text{heat}} = \kappa_{\text{heat}} \eta_{\text{rad}} \dot{m}_{\text{radio}} c^2$. Assuming that all the hot gas is at the virial temperature of the halo T_{vir} , the mass of gas that can be heated per unit time is then

$$\dot{m}_{\text{heat}} = \frac{L_{\text{heat}}}{(3/2)kT/(\mu m_p)} = \frac{L_{\text{heat}}}{(3/4)V_{\text{vir}}^2}, \quad (21)$$

using $kT/(\mu m_p) = (1/2)V_{\text{vir}}^2$. The net cooling rate is then the usual cooling rate \dot{m}_{cool} minus this heating rate \dot{m}_{heat} . If the heating rate exceeds the cooling rate, the cooling rate is set to zero.

We apply this heating term *only* for time-steps in which the haloes are cooling in the ‘hot mode’ (see Section 2.2). That is, if $r_{\text{cool}} > r_{\text{vir}}$ in a given time-step, we assume that the gas is not susceptible to the heating by radio jets, so it cools at the normal rate.

3 RESULTS

3.1 Properties of nearby galaxies

If we adopt a specific set of values for the cosmological parameters, it is relatively straightforward to answer the following question: how must DM halo mass and galaxy mass (or luminosity) be related in order to reconcile CDM with observations? Numerical N -body simulations can now accurately predict the multiplicity function of DM haloes and subhaloes (i.e. the number density of haloes of a given mass), and it is then straightforward to adopt a parametric or non-parametric model relating halo properties to galaxy properties, and to adjust the model to fit the observed stellar mass function or luminosity function of galaxies. This exercise has been carried out in terms of luminosity by Kravtsov et al. (2004a), and in terms of stellar mass by Wang et al. (2006) and Moster et al. (in preparation). It has been shown that if galaxies and haloes are related in this way, one then also reproduces the observed correlation functions of galaxies as a function of stellar mass (Wang et al. 2006; Moster et al., in preparation). We show the function $f_{\text{star}}(M_{\text{halo}})$ derived by Moster et al. (in preparation) in Fig. 3. This quantity is defined as $f_{\text{star}}(M_{\text{halo}}) \equiv m_{\text{star}}/(f_b M_{\text{halo}})$, or the galaxy’s stellar mass divided by the universal baryon fraction times the halo mass. For central galaxies, M_{halo} is the virial mass of the halo. For non-central galaxies,

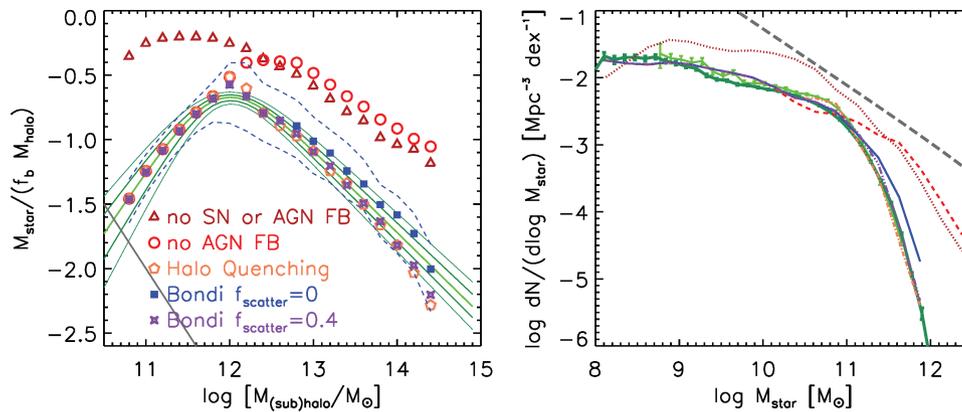


Figure 3. Left-hand panel: Fraction of baryons in the form of stars as a function of halo mass (for central galaxies) or subhalo mass (for satellite galaxies). The solid green lines show the empirical relation (with 1 and 2σ errors) obtained by Moster et al. (2008; see text). Triangles (brown) show the models with no SN or AGN FB; open dots (red) show the model without AGN feedback; pentagons (orange) show the HQ model; solid (blue) squares show the fiducial (isothermal Bondi) model with $f_{\text{scatter}} = 0$, and crosses (purple) show $f_{\text{scatter}} = 0.4$. The dashed lines show the 16th and 84th percentiles for the fiducial model. The diagonal grey line in the bottom left-hand corner shows the stellar mass corresponding to the smallest galaxies that we can accurately resolve, $\sim 10^9 M_\odot$. Right-hand panel: Galaxy stellar mass functions for the same models [dotted (brown) no SN or AGN FB; short dashed (red) no-AGN FB; dot-dashed (orange) HQ; solid (blue) fiducial isothermal Bondi ($f_{\text{scatter}} = 0$); triple-dot-dashed (purple) ($f_{\text{scatter}} = 0.4$)]. Green lines with error bars show the observed galaxy stellar mass functions derived from SDSS by Bell et al. (2003b, light green) and Panter et al. (2007, dark green). The long-dashed grey line shows the DM halo mass function with the masses shifted by a factor equal to the universal baryon fraction.

M_{halo} is the virial mass of the halo just before it became subsumed in a larger halo.

Just by comparing the halo mass function with the observed stellar mass function (see Fig. 3, right-hand panel), we see that in order to reconcile the DM halo mass function predicted by CDM with the observed galaxy stellar mass function, star formation must not only be inefficient overall ($f_{\text{star}} \sim 0.2\text{--}0.3$ at its peak), but the function must be a strong function of halo mass. Apparently, the conversion of baryons into stars is highly inefficient both in small-mass haloes and in large ones, and this efficiency peaks in haloes with mass $\sim 10^{12} M_\odot$. The interesting question that then arises, of course, is which physical processes are responsible for shaping this highly variable efficiency, and for setting the characteristic halo mass scale $\sim 10^{12} M_\odot$? The semi-analytic models can give us some insights into this question. We note that our adopted halo mass resolution ($10^{11} M_\odot$ for host haloes, $10^{10} M_\odot$ for subhaloes) means that our simulations should be reliable and complete for galaxies with stellar masses greater than $\sim 10^9 M_\odot$. Below this mass, we cannot accurately resolve a galaxy’s formation history. If we switch off both AGN feedback and SN feedback,¹ f_{star} is far too high and too flat below $10^{12} M_\odot$ (the mild decline at the low-mass end is due to photoionization squelching). We adjust the parameters of our model for SN-driven winds in order to match the empirical values of f_{star} below $M_{\text{halo}} \sim 10^{12} M_\odot$, and find that we require $\epsilon_{\text{SN}}^0 \sim 1.3$, $\alpha \sim 2$, and $V_{\text{eject}} \sim 120 \text{ km s}^{-1}$. Consulting equation (12), we see that this implies that in large galaxies ($V_{\text{disc}} \sim 200 \text{ km s}^{-1}$), the SN-driven mass outflow rate is comparable to the SFR, and the outflow rate increases fairly strongly with decreasing disc circular velocity V_{disc} . This normalization is in good agreement with the observational results of e.g. Martin (1999), although it is unclear that the strong scaling with circular velocity is supported by these observations. Also, these winds can escape the potential well of the DM halo in

haloes with $V_{\text{vir}} \lesssim 120 \text{ km s}^{-1}$, which is again consistent with the observations of Martin (1999).

In models with no feedback from AGN, we can see from Fig. 3 that f_{star} does turn over at large halo masses: this is because large-mass objects have formed more recently, and have had less time to cool. In some of the earliest explorations of galaxy formation in the CDM paradigm, it was suggested that this cooling time argument could explain the characteristic mass scale of galaxies (White & Rees 1978; Blumenthal et al. 1984). However, one can see that the turnover occurs at too high a mass, and too much gas cools and forms stars in large-mass haloes. This result is obtained not only in semi-analytic models by many different groups (e.g. Benson et al. 2003; Cattaneo et al. 2006; Croton et al. 2006), but also in numerical hydrodynamic simulations (Balogh et al. 2001; Borgani et al. 2006; Cattaneo et al. 2007).

For comparison, Fig. 3 also shows the predictions for $f_{\text{star}}(M_{\text{halo}})$ in a very simple implementation of the concept of AGN heating, what we shall call the ‘halo quenching’ (HQ) model. In this model, we simply shut off cooling flows when the host halo exceeds a mass of $M_Q = 1.3 \times 10^{12} M_\odot$. This model is based on the idea that haloes around $10^{12} M_\odot$ lie near the transition between the ‘cold flow mode’ (gas cooling more rapidly than the free-fall time) and the formation of quasi-hydrostatic hot haloes (‘hot flow mode’), and that radio jets from SMBH can easily keep gas hot if it is in a quasi-hydrostatic hot halo, but not if it is cooling in the cold flow mode (Binney 2004; Dekel & Birnboim 2006). This idea has been previously implemented in a full semi-analytic model by Cattaneo et al. (2006), and was found to very successfully reproduce both the luminosity function and magnitude dependent colour distributions of galaxies. We also find that this model reproduces $f_{\text{star}}(M_{\text{halo}})$, and hence the galaxy stellar mass function, extremely well.

We can then compare this with the prediction of our fiducial model, in which accretion on to a central SMBH is modelled assuming Bondi accretion and the isothermal cooling flow model of NF00 (see Section 2.1.1). We adjust the scaling factor κ_{radio} in order to reproduce $f_{\text{star}}(M_{\text{halo}})$ as well as possible over its whole range. Of course, in this model, the scaling of the heating rate as a function of

¹ In this model, we fix the metallicity of the hot gas to be one-third of solar for purposes of computing the cooling rates.

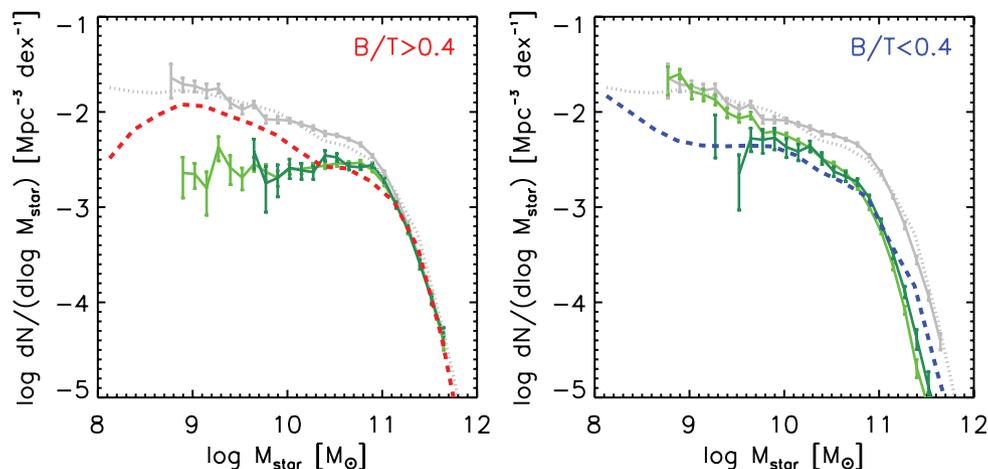


Figure 4. The galaxy stellar mass function divided by morphological type. Solid lines with error bars show observational estimates from SDSS (light green) and 2MASS (dark green; Bell et al. 2003) for early-type galaxies (left-hand panel) and late-type galaxies (right-hand panel). Light (grey) lines show the observed mass function for galaxies of all morphological types. Dashed lines show the model predictions for bulge dominated ($B/T > 0.4$; left-hand panel) and disc-dominated ($B/T < 0.4$; right-hand panel) galaxies, for the fiducial isothermal Bondi model with $f_{\text{scatter}} = 0.4$.

halo mass, and hence the shape of $f_{\text{star}}(M_{\text{halo}})$, is determined by the isothermal Bondi accretion model. Indeed, we can see that our fiducial model (with $f_{\text{scatter}} = 0$) slightly overpredicts f_{star} for large-mass haloes $M_{\text{halo}} \gtrsim 10^{13} M_{\odot}$. This results in a small excess of large-mass galaxies ($M_{\text{star}} \gtrsim 10^{11} M_{\odot}$) in the predicted stellar mass function (Fig. 3, right-hand panel). Note that at halo masses $M_{\text{halo}} \lesssim 10^{12}$, the results for f_{star} are the same as in the model without AGN feedback. The radio heating mode is ineffective in small-mass haloes for multiple reasons: (1) low-mass haloes cool mainly in the ‘cold flow’ mode, and we have assumed that the ‘radio mode’ is fuelled by hot gas and that radio jets can only heat gas that is in a quasi-hydrostatic hot halo; (2) low-mass haloes tend to host disc-dominated galaxies, which do not contain massive black holes.

How concerned should we be about the galaxies in high-mass haloes being too heavy in our fiducial model? The discrepancy amounts to about 0.15–0.2 dex in stellar mass.² A number of recent observational studies have suggested that the luminosities and therefore stellar masses of the central galaxies in clusters may be underestimated by a significant factor (as much as 1.5 mag) in surveys such as SDSS and Two Micron All Sky Survey (2MASS) (Desroches et al. 2007; Lauer et al. 2007; von der Linden et al. 2007), upon which our local galaxy stellar mass function and luminosity function estimates are based. Other studies have shown that a significant fraction of the stars in these galaxies are distributed in a very extended ‘halo’ or envelope (Gonzalez et al. 2005; Zibetti et al. 2005). One possible origin of this extended DSH is stars that are scattered to large radii in mergers (Murante et al. 2004, 2007). In order to explore this idea, we run a model in which a fraction f_{scatter} of the stars in merged satellite galaxies is added to such a diffuse component, which is tracked separately from the main stellar body of the galaxy. We show the results of such a model with $f_{\text{scatter}} = 0.4$ (probably an upper limit on the physically plausible value of this parameter) in Fig. 3, and find that in this model, the stellar mass function is reproduced extremely accurately.

² Although we do not present any results in terms of luminosity in this paper, we see a similar discrepancy in the predicted luminosity functions in all bands. Therefore we probably cannot ascribe the problem to the stellar mass estimates.

As discussed in Section 2.6.1, in each galaxy, we track separately the stars that have survived in an undisturbed disc and stars that have been ‘heated’ by mergers to form a spheroid. Thus we can assign a crude morphological type based on the ratio of the mass in the ‘bulge’ to that in the ‘disc’, B/T . In Fig. 4, we show the stellar mass functions divided into spheroid- and disc-dominated galaxies, compared with observational estimates similarly divided in terms of morphological type (Bell et al. 2003b), for the fiducial model with $f_{\text{scatter}} = 0.4$. There is a small excess of massive disc-dominated galaxies, which may indicate that there is still a small degree of overcooling in our most massive haloes. There is also quite a large excess of low-mass spheroid-dominated galaxies, and a deficit of low-mass disc-dominated galaxies. These problems persist even if we exclude satellite galaxies from our analysis, and cannot be eliminated by simply adjusting the parameter f_{sph} without ruining the agreement for massive galaxies.

Another of the important free parameters in our model is the normalization of the star formation recipe A_{Kenn} . The strongest constraint on this parameter is the ratio of cold gas to stars in galactic discs. Increasing A_{Kenn} causes gas to be converted into stars more rapidly and leads to lower gas fractions. We compare the predicted cold gas fractions $f_{\text{gas}} \equiv m_{\text{cold}}/(m_{\text{cold}} + m_{\text{star}})$ as a function of stellar mass in our fiducial model with observational estimates in Fig. 5. For the models, we consider disc-dominated galaxies ($B/T < 0.4$) which are the central galaxies in their halo (we suspect that the cold gas fractions of satellite galaxies may be too low because we assume that all cooling gas is accreted on to the central galaxy). We compare with the observational estimates of Bell et al. (2003a) for morphologically late-type galaxies and with galaxies on the blue sequence from Kannappan (2004). The agreement is good at stellar masses greater than $\sim 10^{9.5}$, but the gas fractions are a bit low for lower mass galaxies. Note that if we had not adopted a critical density in our star formation law (see Section 2.5), i.e. if we tune $\Sigma_{\text{crit}} \rightarrow 0$, then the gas fractions in low-mass galaxies come out far too low and we do not reproduce the observed trend that gas fractions are higher in low-mass galaxies. A further check on the cold gas content of our galaxies comes from observations of the H I and H₂ mass function. We show the prediction of our fiducial model compared with these observations in Fig. 5, and find reasonable agreement, but with a

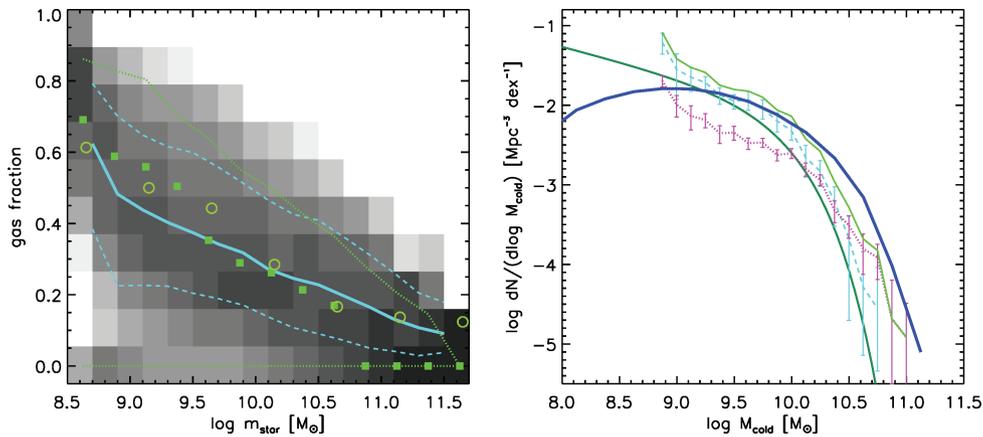


Figure 5. Left-hand panel: The cold gas fraction as a function of stellar mass. The (green) squares show observational estimates for morphologically late-type galaxies derived from the data of Bell et al. (2003a), and open circles show the observational estimates for blue galaxies from Kannappan (2004). The shaded area shows the conditional probability distribution $P(f_{\text{gas}}|m_{\text{star}})$ for central disc-dominated galaxies predicted by our fiducial (isothermal Bondi) model. The (light blue) solid line shows the median of this distribution, and dashed (light blue) lines show the 16th and 84th percentiles. Right-hand panel: The galactic cold gas mass function. The thick (dark blue) line shows the prediction of our fiducial model. The thick solid (dark green) curve shows the observed H I mass function of Zwaan et al. (2005), the dashed (light blue) line shows the H I mass function of Rosenberg & Schneider (2002), the dotted (magenta) line shows the H₂ mass function of Keres, Yun & Young (2003) and the solid (light green) line shows the sum of the Rosenberg & Schneider (2002) H I and H₂ mass functions.

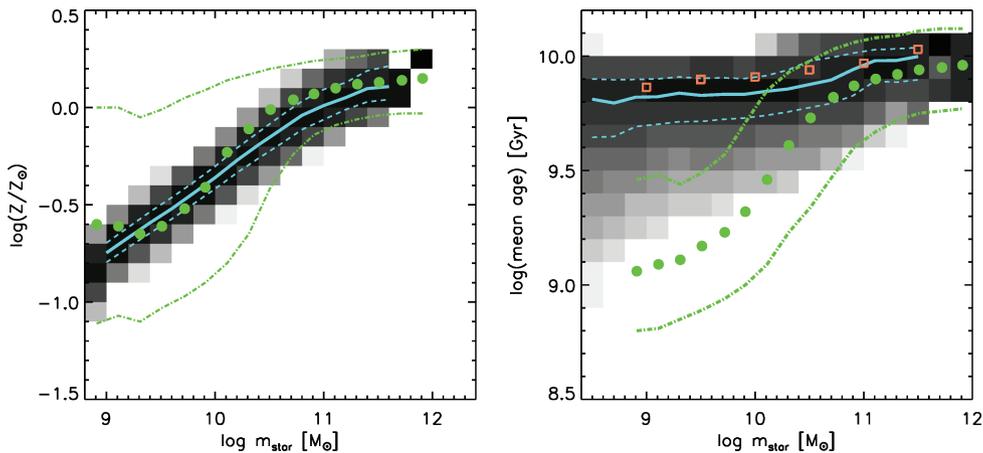


Figure 6. Left-hand panel: Stellar mass versus stellar metallicity. Grey shading shows the conditional probability $P(Z_*|m_*)$ for our fiducial model ($f_{\text{scatter}} = 0.4$), and the (light blue) solid and dashed lines show the 50th, 16th and 84th percentiles. The (green) filled circles and dot-dashed lines show the observational estimates from Gallazzi et al. (2005). Right-hand panel: Stellar mass versus stellar mass-weighted mean stellar age. Shading, lines and symbols are as in the left-hand panel. Open square symbols show the predictions of the semi-analytic model of Croton et al. (2006) for comparison.

hint of a deficit of low-gas-mass galaxies, and a small excess on the high-gas-mass end.

A complementary tracer of star formation is the heavy elements locked up in stars. We show the stellar mass weighted mean stellar metallicity as a function of stellar mass for galaxies in our fiducial model ($f_{\text{scatter}} = 0.4$) in Fig. 6. Our predictions may be compared with observational estimates based on SDSS spectra from Gallazzi et al. (2005). It is worth noting that the estimates of Gallazzi et al. (2005) effectively measure a combination of α -process elements and Fe, while our modelling includes only enrichment due to Type II SN, and therefore our ‘metallicities’ correspond more closely to α -type elements. Also, the Gallazzi et al. (2005) estimates are effectively luminosity weighted, not stellar mass weighted, and may be systematically biased towards higher values for supersolar metallicities (see the discussion in Gallazzi et al. 2005). Considering these potential biases, and the relatively crude nature of our chemical evolution

model, we find fairly good agreement with the observed stellar mass versus metallicity relation. Note that although the normalization of this relation can be adjusted by tuning the value of the stellar yield, y (the results shown here adopt $y = 1.5$ in solar units), the shape of this relation is a fairly complex product of various ingredients of the model. For example, the low-mass slope is primarily determined by the mass-dependent star formation efficiency caused by our star formation threshold and the strongly mass-dependent SN feedback efficiency that we have assumed. The turnover on the high-mass end is shaped by the quenching of star formation by AGN feedback and gas-poor mergers.

In Fig. 6 we also show our model predictions for the stellar mass weighted mean stellar age of galaxies as a function of stellar mass, compared with the observational estimates of Gallazzi et al. (2005). The models predict a weak trend of older ages in more massive galaxies, with the ages of massive galaxies slightly older than the

observational estimates. However, the predicted trend in our models is much weaker than the observed trend found by Gallazzi et al. (2005), and low-mass galaxies in our models are much older than the observations indicate. Croton et al. (2006) showed that models without AGN feedback predicted that massive galaxies have ages as young as low-mass galaxies, and that introducing radio mode AGN feedback produced an age-mass trend with the correct sense (more massive galaxies are older). However, they did not compare directly with observational estimates. We reproduce the predictions from the AGN feedback model shown in fig. 10 of Croton et al. (2006), and see that their results are very similar to ours (in fact low-mass galaxies are slightly older in their models than in ours). We should keep in mind that if a galaxy has a significant-by-mass older stellar population with a small ‘frosting’ of young stars, the ages derived from stellar absorption lines (mainly Balmer lines) as in the Gallazzi et al. (2005) approach will be biased towards young ages. This discrepancy is worth examining in more detail, but this is beyond the scope of this paper.

The relationship between galaxy mass and BH mass is clearly a key result that our model should reproduce. Recall that in our model, this relationship is set by the depth of the potential well of the galaxy at the time when the BH forms, which in turn is determined by the gas fraction of the progenitor galaxies of the last merger (see Section 2.9). More gas-rich progenitors suffer more dissipation when they merge, and produce more compact remnants with deeper potential wells. A deeper potential well requires more energy, and therefore a more massive BH in order to halt further accretion and growth. Although we have seen that the predicted gas fractions of discs at the present day agree reasonably well with observations, the gas fractions of the progenitors of black hole hosts depend on many factors, such as the masses of those progenitors at the time when the BH is formed, the epoch of formation of BH of a given mass, and the details of the star formation and feedback modelling. It is therefore a non-trivial success of our model that we reproduce the observed slope and scatter of the $M_{\text{BH}}-M_{\text{sph}}$ (black hole mass versus spheroid mass) relationship, as seen in Fig. 7.

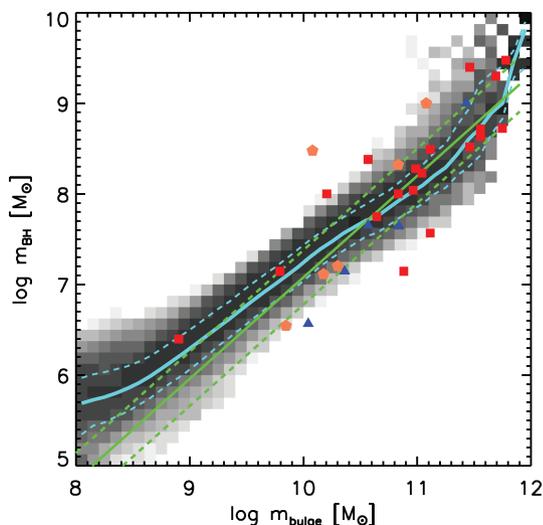


Figure 7. Predicted relationship between bulge mass and black hole mass [grey shading indicates the conditional probability $P(m_{\text{bh}}|m_{\text{bulge}})$; light blue solid and dashed line shows the median and 16th and 84th percentiles] compared with the observed relation from Häring & Rix (2004, green lines). Symbols show the measurements for individual galaxies from Häring & Rix (2004).

It is interesting that our model predicts a small upward curvature at the high-mass end, which Wyithe (2006) argue is present in the observed relation. Our predicted relation also has a somewhat flatter slope at low BH masses than the extrapolation of the Häring & Rix (2004) results; however, there are currently very few robust BH mass estimates at such low masses.

3.2 Group and cluster properties

We have so far focused on the properties of individual galaxies. We now consider predictions for a few properties of groups and clusters. We have selected these quantities because they help to constrain some of the free parameters or uncertain ingredients in our models. In Fig. 8, we show the hot gas fraction ($f_{\text{hot}} \equiv m_{\text{hot}}/M_{\text{vir}}$), i.e. the mass of hot gas contained in the DM halo divided by the total virial mass of the halo. The hot gas fraction in our models has a sharp ‘step’ at $M_{\text{vir}} \sim 10^{12} M_{\odot}$, because of the rapid transition between haloes in which the gas cools rapidly compared with the dynamical time, so there is typically little hot gas present in the halo, and haloes in which the cooling time is longer compared with the dynamical time, so the halo can build up a reservoir of hot gas. In our fiducial model, this hot gas is then maintained by AGN ‘radio mode’ heating. Our results are in reasonable agreement with the hot gas fractions in clusters estimated from observations of their X-ray emitting gas by Vikhlinin et al. (2006). These observations are somewhat uncertain, because the X-ray emission typically cannot be detected all the way out to the cluster virial radius, so it must be extrapolated. However, these observations provide an important constraint on the modelling of re-infall of gas that has been ejected by SNe (see Section 2.7). If we do not allow this ejected gas to be re-accreted at all, then the baryon fractions in clusters are predicted to be significantly smaller than the universal value, in conflict with observations.

In Fig. 8 we also show the predicted metallicity of the hot gas in haloes. In our model, the hot gas is enriched by the ejection of metals from the cold gas in galactic discs by SN-driven winds. We tuned the chemical yield Y to reproduce the metallicities of stars in galaxies, so the metallicity of the hot cluster gas is a cross-check on our chemical evolution and SN feedback modelling. We find that the hot gas in cluster-mass haloes is enriched to about 0.25 of the solar value, and is nearly constant above about $M_{\text{vir}} \sim 10^{13} M_{\odot}$. This is close to the value of $\sim 0.3 Z_{\odot}$ measured for hot gas in clusters (Arnaud et al. 1992). These measurements of hot gas metallicity are primarily sensitive to iron, while as we discussed earlier in this section, our chemical evolution modelling traces only the metals produced by Type II SNe, so we do not expect perfect agreement. Also, some additional metals may be driven out of the galaxies by strong shocks during mergers (Cox et al. 2006a).

We have argued that a significant fraction of stars may be scattered into a DSH by mergers. It is important to check whether the predicted mass of stars in these DSH is in agreement with direct observational measurements. In Fig. 9 we show the mass of the DSH divided by the total mass of the DSH plus the main galaxy [$f_{\text{DSH}} \equiv m_{\text{DSH}}/(m_{\text{DSH}} + m_{\text{BCG}})$], as a function of the virial mass of the halo. The model predictions are consistent with the range of observational estimates from Gonzalez et al. (2005), which we have adopted from the results presented in Conroy et al. (2007). In agreement with previous studies by Monaco et al. (2006) and Conroy et al. (2007), we find that the model with $f_{\text{scatter}} = 0.4$ is able to reproduce the observed stellar mass function of

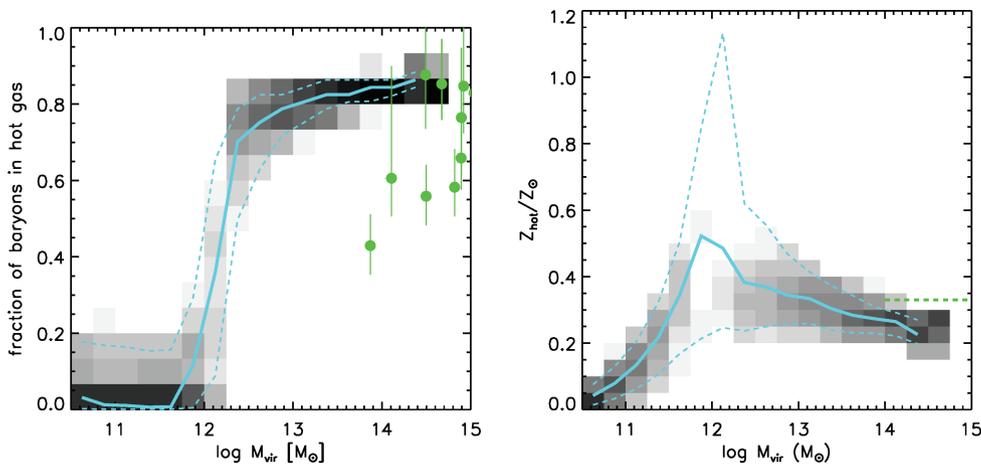


Figure 8. Left-hand panel: The fraction of halo baryons in the form of hot gas as a function of halo virial mass. The grey-shaded area shows the conditional probability distribution $P(f_{\text{hot}}|M_{\text{vir}})$ for our fiducial model, with the 16th, 50th and 84th percentiles shown with (light blue) curves. The sharp discontinuity at $M_{\text{vir}} \simeq 10^{12} M_{\odot}$ represents the transition from rapid cooling (cold flows) to the formation of a hot halo. The (green) solid circles show the observational estimates of hot gas fraction from Vikhlinin et al. (2006). Right-hand panel: The metallicity of hot cluster gas as a function of halo mass in our fiducial model. The green dashed line shows the observed value in clusters, approximately one-third of the solar value (Arnaud et al. 1992).

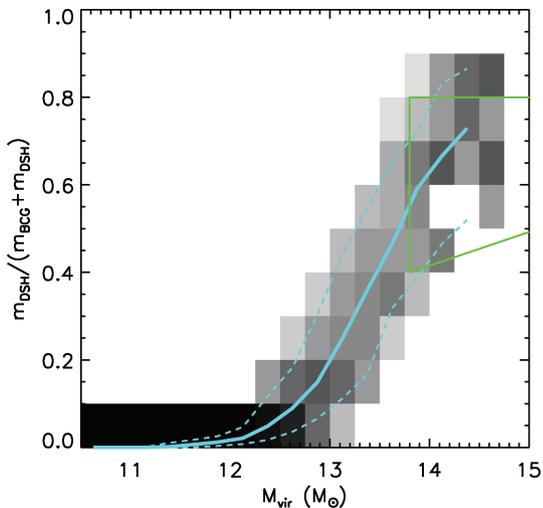


Figure 9. Mass contained in a ‘DSH’ relative to the mass of the central galaxy plus the DSH mass. The shading and (light blue) curves show the predictions of our fiducial model; the green box shows the approximate locus of observational estimates from Gonzalez et al. (2005).

galaxies at $z \sim 0$ as well as the fraction of stellar mass in the DSH component.³

3.3 Radio mode heating

Several other groups have implemented heating by radio jets from AGN in semi-analytic models, and shown that in this way they can

³ We note that the published values of $f_{\text{DSH}} \simeq 0.33$ from Zibetti et al. (2005) are significantly lower than the values we have adopted here. While several factors may play a role, such as the different extent of the photometry and sample selection, much of the discrepancy with the results of Gonzalez et al. (2005) is apparently due to the details of the way the BCG and DSH (or intracluster light) are defined (A. Gonzalez and S. Zibetti, private communication; see also Zibetti (2008), Section 5.1). When Zibetti adopts the same decomposition method as Gonzalez, he finds much more consistent values $f_{\text{DSH}} \simeq 0.67$ (Zibetti 2008).

solve the overcooling problem and other related problems (Bower et al. 2006; Croton et al. 2006; Monaco et al. 2007). However, these papers have not addressed whether the *amount of energy required* or the *scalings as a function of halo mass* adopted in these models are consistent with constraints from observations of radio galaxies and cooling flow clusters. We turn now to this question. Fig. 10 shows the predicted cooling rate as a function of halo mass in a model without AGN heating. The shaded area shows the cooling rates predicted by the full semi-analytic model, while the smooth lines show the cooling rate given by equation (2), assuming that $m_{\text{hot}} = f_b M_{\text{vir}}$ and $Z_{\text{hot}}/Z_{\odot} = 0.33$ (we refer to this as the ‘static halo’ cooling model). The lower of these lines is for redshift $z = 0$, and the higher is for $z = 2$. The divide between ‘cold mode’ and ‘hot mode’ haloes at $\sim 10^{12} M_{\odot}$ is indicated by the colour of the lines (with blue indicating cold mode, red indicating hot mode). Clearly, the static halo model prediction can differ from the cooling rate in the full SAM because, as we saw in Fig. 8, haloes of a given mass have a range of values of hot gas fraction and metallicity due to their different formation histories. Here we see that, in the absence of AGN heating, cluster mass haloes would be expected to have cooling flows of hundreds up to one thousand solar masses per year, which we know to be in conflict with X-ray observations. In the middle panel we show the rate at which gas is heated by the ‘radio jets’ in our fiducial (isothermal Bondi) model. Note that although we show non-zero heating rates below $\sim 10^{12} M_{\odot}$, actually, most of these haloes are cooling in the ‘cold flow’ mode and so their cold gas accretion rates are unaffected by the AGN heating. We note that the heating and cooling rates cross near the ‘magic’ halo mass of $\sim 10^{12} M_{\odot}$, and that the heating rate is a steeper function of halo mass than the cooling rate at large masses, so that there is a lot of ‘excess’ energy being deposited in the hot gas. At the moment, in our simple modelling, this excess energy is not accounted for. The final panel in this plot shows the net cooling rate including the AGN heating. Cooling flows are quenched entirely in haloes more massive than $M_{\text{vir}} \sim 10^{13} M_{\odot}$. However, this is not a sharp cut-off. There is a transition region in the range $10^{12} \lesssim M_{\text{vir}} \lesssim 10^{13}$ where some haloes have had their cooling flows quenched and some have only had them reduced.

We can compare the heating rates needed in our model in order to solve the overcooling problem and the galaxy mass problem with

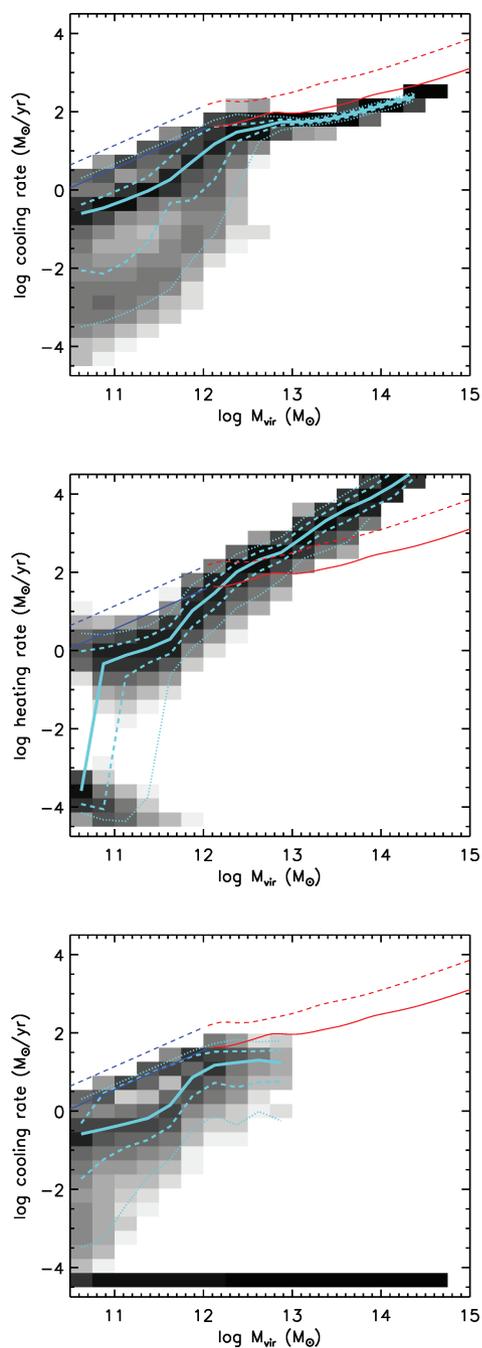


Figure 10. Top: Cooling rate as function of halo mass, for model with no AGN feedback. Grey shading indicates the conditional probability $P(\dot{m}_{\text{cool}}|M_h)$. Light blue lines indicate the 2.2th, 16th, 50th, 84th and 98th percentiles. The smooth red and blue solid and dashed lines indicate the expected cooling rates in a ‘static halo’ model (see text) for $z = 0$ and 2, respectively; the dark blue section of these lines indicates the approximate regime for ‘cold mode’ infall, and the red lines for ‘hot mode’ (see text). Middle: Heating rate by ‘radio jets’ in our fiducial (isothermal Bondi) model. Bottom: Net cooling rate in fiducial model with radio mode feedback. Haloes with cooling rates below $10^{-4.5} M_{\odot} \text{ yr}^{-1}$ are plotted in the bottom-most bin.

observations of the hot bubbles associated with radio jets, seen in the X-ray gas in groups and clusters. By estimating the amount of energy required to inflate the bubbles, these systems can be used as ‘calorimeters’, giving an estimate of the power being injected by the jets. A06 find a tight correlation between the Bondi accretion

rate, and the jet power. For systems that also have a black hole mass estimate or its proxy, for example from a measured velocity dispersion or bulge mass, we can then assess the fraction of the black hole’s rest mass that is being extracted as kinetic energy that can heat the gas. In Fig. 11, we show observational estimates of the rate of energy injection, or jet power, as a function of black hole mass, from observations of elliptical galaxies with associated hot gas bubbles by A06 and Rafferty et al. (2006). The Rafferty et al. (2006) estimates of jet power overlap those of A06, but extend to considerably higher values for a given BH mass. This may be because the Rafferty et al. sample, which extends to higher redshift, contains more massive clusters than the very nearby sample of A06. These massive clusters may have higher gas densities at the relevant radii, thus allowing more efficient coupling of the radio jet with the ICM (S. Allen, private communication).

We also show the time-averaged heating rate as a function of BH mass estimated by Best et al. (2006), from observations of the radio-loud fraction of SDSS galaxies. The observed scaling of jet power with BH mass is a bit steeper than it would be if the jet power were proportional to the Eddington luminosity: the jet power scales approximately as $M_{\text{BH}}^{1.6}$, or as $L_{\text{Edd}} M_{\text{BH}}^{0.6}$.

These results may be compared directly with the ‘jet power’ as a function of BH mass incorporated in our fiducial model. Recall that in our model, we assumed that the central density and temperature of the gas was set by the isothermal cooling flow model of NF00 (see Section 2.1.1) and the accretion rate on to the BH was then set by the Bondi accretion rate. We allowed an overall scaling factor κ_{radio} , which we adjusted to the *minimum* value that produced a good fit to the empirical constraint on galaxy stellar mass as a function of host halo mass discussed in Section 3.1. We find that the resulting accretion rates are about an order of magnitude higher than the Bondi rates estimated by A06, and the jet power at a given BH mass is also higher than the A06 estimates. However, the jet powers are consistent with the higher values in the Rafferty et al. (2006) sample. It is also important to remember that these observational estimates are lower limits. They include only the energy associated with inflating the bubbles, while significant energy can also be dissipated through sound waves, viscosity and weak shocks (McNamara et al. 2005; Nulsen et al. 2005a,b; Fabian et al. 2006; Forman et al. 2007). Binney, Bibi & Omma (2007) analysed 3D adaptive grid simulations of heating of cooling flows, and found that the bubbles reflected only ~ 10 per cent of the total injected energy.

However, it is also possible that we are overestimating the energy required, because we are insisting that AGN heating does the whole job, while as we have discussed, there may be several other processes that help to reduce the efficiency of cooling in group and cluster-mass haloes. Furthermore, the semi-analytic cooling model is probably only accurate to a factor of 2–3 at best, and may be overestimating the cooling rates. We intend to test and better calibrate our models by comparing with the results of numerical simulations, but this is not entirely straightforward, because SN feedback probably plays an important role in altering the equation of state and metallicity of the gas, which affects the cooling rates. Given these uncertainties, we conclude that the heating due to AGN predicted by our simple model is not only very successful at solving the over-cooling problem, it is also reasonably consistent with the direct observational constraints.

Fig. 11 (right-hand panel) also shows the predicted ‘specific jet energy’ (jet energy divided by BH mass) as a function of the halo virial velocity. One can see that the isothermal Bondi model predicts a fairly strong dependence of jet power on halo virial velocity (or

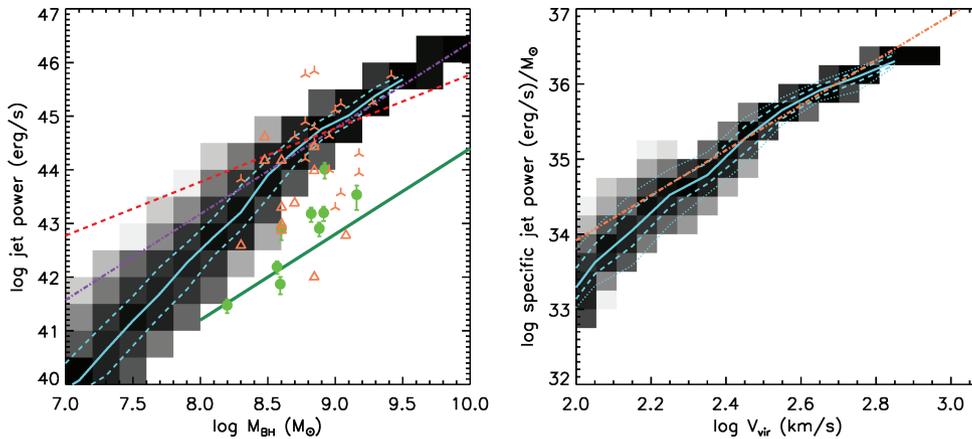


Figure 11. Left-hand panel: Rate of energy input by the ‘radio mode’ (jet power) as a function of BH mass. Large, solid green circles show the observational estimates from A06. Open and skeletal triangles (orange) show the observational estimates from Rafferty et al. (2006), for systems at redshift $z < 0.05$ and $z > 0.05$, respectively. The thick (dark green) solid line shows the time-averaged heating rate derived from observations by Best et al. (2006). The grey-shaded area shows the conditional probability distribution $P(P_{\text{jet}} | m_{\text{BH}})$, and the light blue curves show the median and 16th and 84th percentiles of this distribution. The (red) dashed line shows the heating rate that would result if all BH accreted at a fixed fraction ($f_{\text{R}} = 3.5 \times 10^{-3}$ is shown here) of their Eddington rate, and the (purple) dot-dashed line shows $P_{\text{jet}} \propto m_{\text{BH}}^{1.6}$. Right-hand panel: The jet power divided by the black hole mass (specific jet power) as a function of halo virial velocity ($\propto T_{\text{vir}}^{1/2}$). The grey-shaded region shows the conditional probability distribution, and light blue lines show the 2.2th, 16th, 50th, 84th and 98th percentiles (from bottom to top), for our fiducial isothermal Bondi model. The dot-dashed (orange) line shows the scaling for the fiducial model of Croton et al. (2006).

temperature), and in fact this scaling is important in fitting the shape of the $f_{\text{star}}(M_{\text{halo}})$ function at intermediate masses – we find that if we assume that the accretion rate (and hence the jet power) is just a function of BH mass, we obtain qualitatively similar results, but we do not get as good a match to the $f_{\text{star}}(M_{\text{halo}})$ function and hence the galaxy stellar mass function. For comparison, we also show the empirical scaling adopted in the fiducial Munich SAM (e.g. Croton et al. 2006). As also noted by Croton et al. (2006), their adopted scaling $\dot{m}_{\text{acc}}/m_{\text{BH}} \propto V_{\text{vir}}^3$ is very similar to that predicted by the isothermal Bondi model, and it yields very similar results when we adopt it in our SAM.

3.4 Star formation quenching

As we noted in Section 1, there are two puzzles in galaxy formation that we wish to address in this paper. One is that real galaxies do not grow as massive as we would have predicted in the absence of AGN feedback. The other is that star formation in most massive galaxies has been quenched, while most low-mass galaxies continue to form stars. These two problems seem very likely to be interconnected, but it is not obvious that a specific model which solves one problem will necessarily solve the other. In order to assess the quenching problem, it is common practice to compare model predictions with observed optical or optical–NIR colour–magnitude distributions. However, these kinds of predictions are quite sensitive to metallicity and dust, which complicates the interpretation. Instead, we make use of the physical properties, specific SFR (SSFR $\equiv \dot{m}_{\text{star}}/m_{\text{star}}$) and stellar mass, derived from *GALEX* UV photometry plus SDSS five-band optical photometry (Salim et al. 2007; Schiminovich et al. 2007). Yi et al. (2005) and Kaviraj et al. (2007) have shown that the NUV–optical colours are a highly effective way to probe small amounts of recent star formation in galaxies with an underlying old stellar population.

Fig. 12 shows the conditional probability distribution of SSFR as a function of stellar mass $P(\text{SSFR} | m_{\text{star}})$. The top left-hand panel shows the observed distribution derived from the *GALEX* + SDSS

data by Schiminovich et al. (2007). The star-forming sequence, quenched sequence, and the dividing line (sometimes called the ‘green valley’) derived from *GALEX* + SDSS by Salim et al. (2007) are also shown, and are repeated on every panel. It is important to remember that the *GALEX*–SDSS survey is incomplete in the bottom left-hand region of the plot. We also show the same distribution, as predicted by the no-AGN FB model (top right-hand panel), the HQ model (bottom left-hand panel) and the fiducial isothermal Bondi model (bottom right-hand panel). The SFRs shown for the models have been averaged over the past 10^8 yr.

All of the models shown have the same problem with low-mass galaxies. The star-forming sequence is nearly flat, rather than being tilted such that less massive galaxies have higher SSFR, as in the observations, and the SSFRs are too low. Though we have tried extensive experiments with parameter variation, we have not succeeded in solving this problem. The problem also seems to be quite robust to the star formation recipe that we adopt. If we remove the SF threshold, thereby effectively increasing the star formation efficiency in low-mass galaxies, then the galaxies consume more gas and have lower gas fractions at the present day. Their SSFR are still low, because they have very little fuel for star formation. If we increase the star formation threshold, which makes star formation even more inefficient in low-mass galaxies, the galaxies have higher gas fractions at the present day, but most of that gas is not allowed to make stars, so the SSFR are again almost the same as before.

In the model without AGN feedback, nearly all massive galaxies have high SSFR and would also have blue colours, in drastic conflict with the observations. In the HQ model, which did as well or better than our fiducial model at matching $f_{\text{star}}(M_{\text{halo}})$ and the galaxy stellar mass function, essentially *all* massive galaxies are completely quenched. This is not surprising, as we know that there is a fairly tight relationship between stellar mass and halo mass, so a sharp cut-off in halo mass produces a fairly sharp cut-off in stellar mass. This model cannot account for the population of massive galaxies with small but detectable amounts of recent star formation, seen in the *GALEX* observations (Yi et al. 2005 argue that the NUV light in

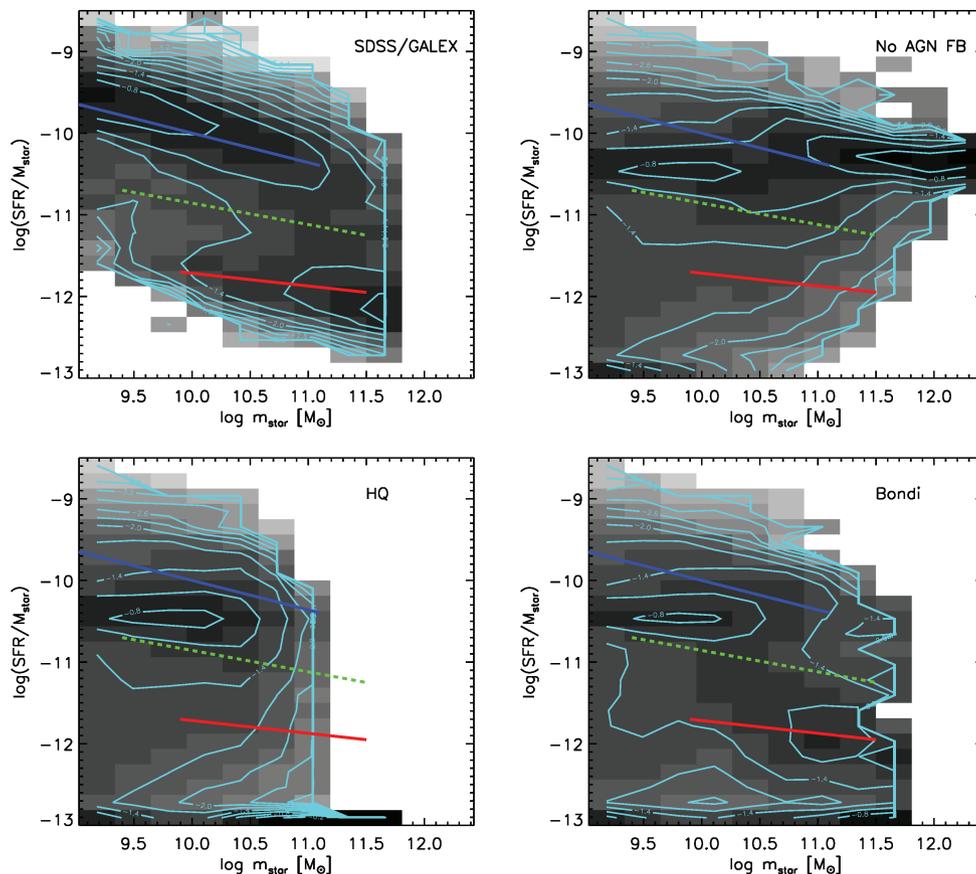


Figure 12. SSFR (SFR divided by stellar mass) versus stellar mass. Grey shading and contours indicate the conditional probability $P(\text{SSFR}|m_{\text{star}})$. The diagonal (dark blue) solid line in the upper left-hand part of the plot and the (red) line in the lower right-hand part of the plot indicate the ‘star-forming sequence’ and ‘quenched sequence’ from the observational results of Salim et al. (2007). The middle, dashed (green) line indicates the dividing line between the star-forming or active galaxies and quenched galaxies (sometimes called the ‘green valley’). These active, valley and quenched sequences based on the observed distributions from *GALEX* are repeated on all four panels. Top left-hand panel: Observed SSFR versus mass distribution from *GALEX* (Schiminovich et al. 2007). Top right-hand panel: Predicted distribution from the model with no AGN feedback. Bottom left-hand panel: Predicted distribution from the HQ model. Bottom right-hand panel: Predicted distribution from the fiducial (isothermal Bondi) model.

these galaxies is indeed due to star formation and not a UV upturn or AGN). However, our fiducial isothermal Bondi model does produce such a population of massive galaxies whose star formation has been substantially, but not completely, quenched. The SSFR versus m_{star} distribution predicted by this model looks qualitatively quite similar to the observations (for massive galaxies).

We analyse the distribution of SSFR versus stellar mass in more detail in Fig. 13. Here, we show histograms of SSFR in stellar mass bins, for the observations and two of the models: the no-AGN FB model and the fiducial model. From this comparison we can see that our fiducial model is not producing quite enough massive, actively star-forming galaxies ($m_{\text{star}} \gtrsim 10.7$). It seems that star formation is actually being quenched a bit too effectively in massive galaxies. However, overall the agreement is quite good.

3.5 Global formation histories of stars, cold gas and metals

In this paper, we have focused mainly on predictions of the present-day ($z \sim 0$) properties of galaxies. We plan to explore the predictions for the properties of high-redshift galaxies in detail in future papers. However, in this section we present predictions for the global histories of several important (and observationally accessible) components of galaxies: the SFR, stars, cold gas and metals.

3.5.1 Dependence on cosmology: the WMAP3 model

In this section we introduce a new model, which has the same recipes for galaxy and BH formation as our fiducial isothermal Bondi model, but adopts the cosmological parameters (see Table 1) from the 3-yr analysis of the *WMAP* data reported in Spergel et al. (2007). We will refer to this as the *WMAP3* model.⁴ For our purposes, the most important differences between the Concordance Λ CDM (C- Λ CDM) model that we have been using so far and the *WMAP3* model is that *WMAP3* has a lower value of σ_8 , the normalization of the primordial power spectrum, and the primordial power spectrum also has a ‘tilt’ ($n_s = 0.96$) while C- Λ CDM has a scale-free initial power spectrum ($n_s = 1$). This results in less power on small scales, and hence later structure formation in *WMAP3* relative to C- Λ CDM. In order to reproduce the $z = 0$ observations as

⁴ While this paper was in the referring process, new results for the cosmological parameters derived from the five year *WMAP* data, combined with distance estimates from Type Ia SNe and baryon acoustic oscillations, were posted on astro-ph (Komatsu et al. 2008). The new estimates of the parameters most relevant to our results, $\sigma_8 = 0.817 \pm 0.026$ and $n_s = 0.960^{+0.14}_{-0.013}$, are intermediate between the values adopted in our C- Λ CDM and *WMAP3* models, though somewhat closer to the *WMAP3* values.

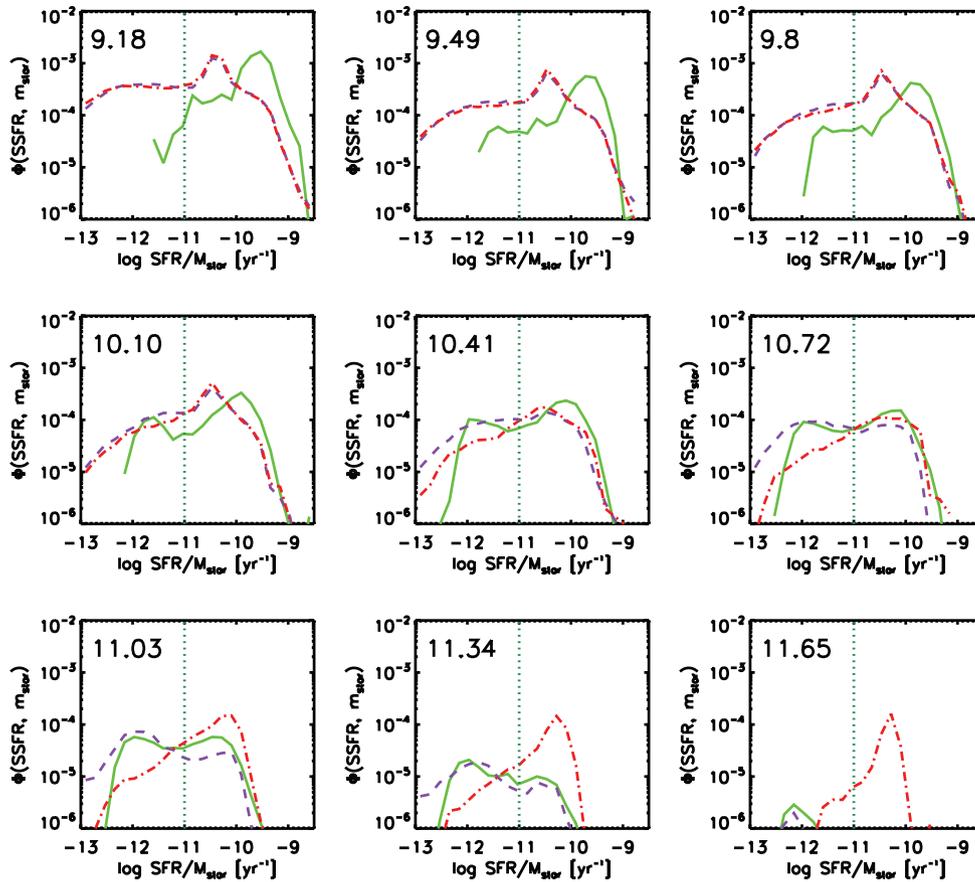


Figure 13. Distribution of SSFRs in stellar mass bins (as indicated on the panels). Solid (green) lines show the observational results from *GALEX* (Schiminovich et al. 2007), dot-dashed (red) lines show the no-AGN FB model and dashed (purple) lines show the fiducial (isothermal Bondi) model. Vertical dotted lines show the rough location of the division between the ‘active’ and ‘quenched’ populations.

before, we retuned the star formation efficiency, radio mode heating efficiency, and scattering parameters (we used $A_{\text{Kenn}} = 1.67 \times 10^{-4}$, $\kappa_{\text{radio}} = 6.0 \times 10^{-3}$ and $f_{\text{scatter}} = 0.2$), but left the other parameters the same. After this retuning, the *WMAP3* model produces nearly indistinguishable results from the Λ -CDM model for all of the quantities that we have shown so far (see also Wang et al. 2007).

3.5.2 The global star formation and mass assembly history

Fig. 14 (top left-hand panel) shows the global SFR density of all galaxies predicted in the three Λ -CDM models: the model with no-AGN FB, the HQ model and the fiducial isothermal Bondi model. A compilation of observational estimates is also shown. The no-AGN FB model overpredicts the amount of star formation at low redshift, and the SFR does not decline as rapidly as the data indicate, while both the HQ model and the isothermal Bondi models produce very good agreement with the observations at $z \lesssim 2$. We can see that the effect of the AGN heating starts to become significant only at $z \lesssim 4$. At higher redshifts, the AGN heating has little or no impact on the global SFR, because most star formation is taking place in relatively small-mass haloes, which are not affected by the ‘radio mode’ heating, as we have already discussed. It is also interesting that the simple HQ model produces such similar results to the fiducial isothermal Bondi model, even over a large range in redshift. At $z \gtrsim 4$, the Λ -CDM models predict more star formation

than the observational compilation of Hopkins & Beacom (2006), but agree with the higher estimates in the literature (e.g. Steidel et al. 1999; Giavalisco et al. 2004). The *WMAP3* model predicts a much more rapidly declining SFR at $z \gtrsim 2$, somewhat lower than the observations but well within the observational errors.

The lower set of lines shows the star formation contributed by merger-triggered bursts. The contribution due to bursts is much lower than in our previous models (SPF01), for several reasons. (1) Our new treatment of dynamical friction and tidal stripping and destruction means that many low-mass satellites take longer than a Hubble time to merge, or are tidally destroyed before they can merge. Therefore the number of minor mergers is lower. Also, we do not include satellite–satellite mergers here, which were included in our previous models. (2) Our newly calibrated burst efficiencies in minor mergers, which are based on a greatly expanded and improved set of hydrodynamic simulations, are lower than in our previous models. (3) Our Kennicutt star formation law causes the star formation efficiencies in quiescent discs to increase with increasing redshift. As we already showed in SPF01, this leaves less cold gas fuel for bursts, leading to a decreased contribution to the SFR from bursts.

In the top right-hand panel of Fig. 14 we show the complementary quantity ρ_{star} , the integrated cosmic stellar mass density. All of the Λ -CDM models predict a significantly *earlier* assembly of stars in galaxies than observations of high-redshift galaxies indicate. The mass density of long-lived stars in both our fiducial and the HQ

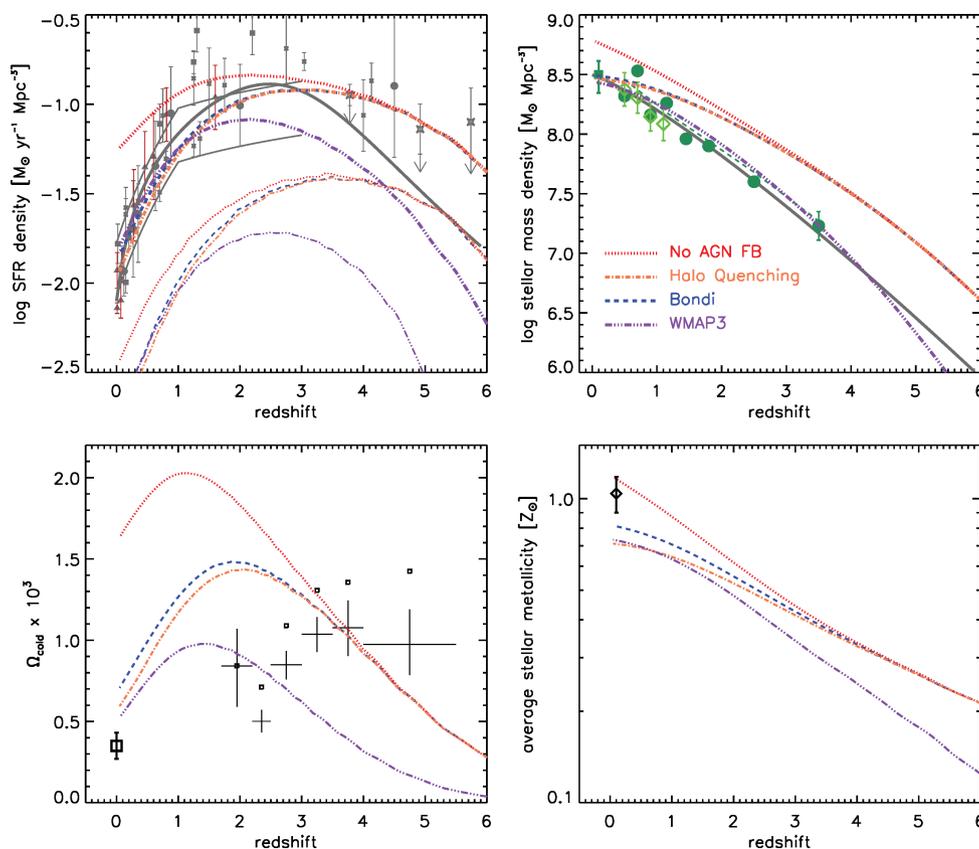


Figure 14. In all panels, dotted (red), dot-dashed (orange) and dashed (blue) lines show predictions from the Λ -CDM no-AGN FB model, the HQ model, and the fiducial (isothermal Bondi) model, respectively. The triple-dot-dashed (purple) line shows the *WMAP3* model. Top left-hand panel: SFR density as a function of redshift. The upper set of thicker lines shows the total SFR in the models, and the lower set of thin lines shows the SFR due to bursts. Symbols show the compilation of observational results of Hopkins (2004), converted to a Chabrier IMF. The thick solid (grey) curve is the fit to the observational compilation presented by Hopkins & Beacom (2006). The thin (grey) solid broken curve shows observational estimates from *GALEX* (Schiminovich et al. 2005). Top right-hand panel: The integrated global stellar mass density as a function of redshift. Symbols show observational estimates from Bell et al. (2003b, $z \sim 0$), Borch et al. (2006, $z \sim 0.2-1$; diamonds) and Fontana et al. (2006, $z \sim 0.2-3.5$; filled circles). The thick (grey) curve shows the fit to the observational compilation presented in Wilkins et al. (2008). Bottom left-hand panel: The mass density of cold gas in units of the critical density $\Omega_{\text{cold}} \equiv \rho_{\text{cold}}/\rho_{\text{crit}}$. Symbols show the observational estimates at $z \sim 0$ from the blind H I survey of Zwaan et al. (2005), and from damped Lyman α systems (crosses) at high redshift (Prochaska et al. 2005). Small open squares also include the contribution from lower column density absorption systems (Lyman-limit systems). Bottom right-hand panel: The mass-weighted average metallicity of stars $\langle Z_{\text{star}} \rangle \equiv \rho_Z/\rho_{\text{star}}$ as a function of redshift. The diamond symbol at $z \sim 0.1$ shows the observational estimate from SDSS galaxies from Gallazzi et al. (2008).

model is a factor of ~ 3 higher at $z \sim 2$ and a factor of ~ 2 higher at $z \sim 1$ than the observations. However, the *WMAP3* model produces excellent agreement with the stellar mass density as a function of redshift. We note here that this tension in the model results (i.e. that the Λ -CDM model fits the SFR history data better, while the *WMAP3* model provides a better fit to the stellar mass density) is connected with a possible inconsistency between the two data sets that has been noted recently in several papers (e.g. Hopkins & Beacom 2006; Fardal et al. 2007; Davé 2008; Wilkins, Trentham & Hopkins 2008). One possible resolution of this tension can be obtained if the stellar IMF has changed with time, and was more top-heavy at high redshift (we return to this issue in Section 4).

3.5.3 Evolution of cold gas and metals

In the bottom left-hand panel of Fig. 14 we investigate another complementary quantity, the mass density of gas that has cooled but not yet formed stars. In our models, all of this cold gas is assumed to reside in galactic discs. One can compare the model predictions with the total mass density of H I gas from blind H I surveys at

$z \sim 0$ (Zwaan et al. 2005), as well as with estimates of cold gas at high redshift from damped Lyman α systems (e.g. Prochaska, Herbert-Fort & Wolfe 2005, and references therein). Note that all the observational estimates shown here are for atomic gas only, and do not include the contribution from molecular gas. Therefore, these observations are lower limits for the model predictions of cold gas. The Λ -CDM models are consistent with the observations to $z \sim 4$, but are somewhat low at $z \sim 5$ (note however that Prochaska et al. (2005) ‘caution the reader that the results at $z > 4$ should be confirmed by higher resolution observations’). However, Ω_{cold} in the *WMAP3* model is much lower than the observations at $z \gtrsim 3$, by about a factor of 2 at $z = 3.5$ and an order of magnitude at $z \sim 5$.

In the last (bottom right-hand) panel of Fig. 14 we show the stellar mass weighted, globally averaged metallicity as a function of redshift predicted by the no-AGN FB model, the HQ model and the isothermal Bondi models. We can compare this with the results of the observational analysis of SDSS galaxies by Gallazzi et al. (2008), who find a mass-weighted average stellar metallicity at $z \sim 0$ of about solar. The no-AGN FB model overshoots this value, while the models with AGN FB slightly underestimate it. Gallazzi

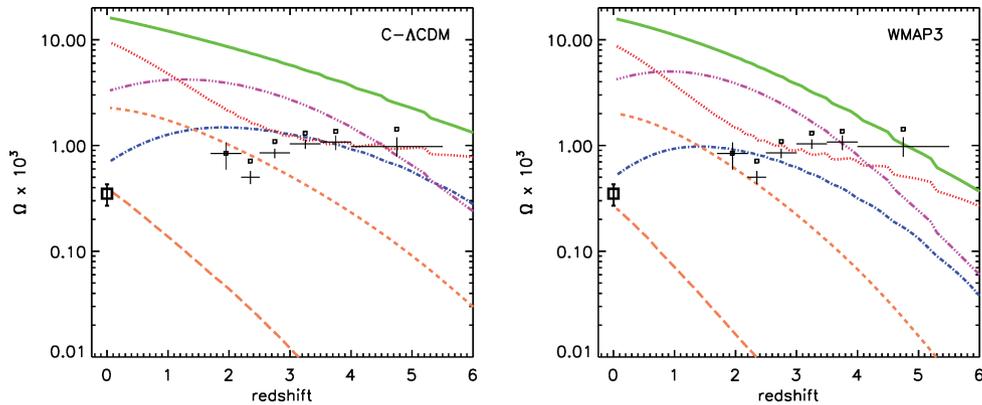


Figure 15. The global density in units of the critical density of various components as a function of redshift predicted by the isothermal Bondi model (left-hand panel: C- Λ CDM with $f_{\text{scatter}} = 0.4$; right-hand panel: *WMAP3* with $f_{\text{scatter}} = 0.2$). From the highest to the lowest curve at $z = 0$, we show the universal baryon fraction times the mass contained in virialized haloes above our mass resolution ($10^{10} M_{\odot}$; solid green); shock-heated hot gas in haloes (dotted red); gas in the ‘IGM’ (see text; triple-dot-dashed magenta); stars in galaxies (short-dashed orange); cold gas in galaxies (dot-dashed blue) and stars in DSHs around galaxies (long-dashed orange). We repeat the observational estimates of Ω_{cold} from Fig. 14 (squares and crosses).

et al. (2008) showed that the Millennium Simulations, based on the semi-analytic model of Croton et al. (2006), give very similar results to our models – they underproduce the stellar metallicity at $z = 0$ by about 20–30 per cent. The models predict early enrichment, with the mean stellar mass at $z \sim 6$ about 25 per cent of solar, and 50 per cent of solar at $z \sim 2.5$. Because of the difficulty of obtaining an unbiased stellar mass weighted global mean metallicity at high redshift, we do not attempt a quantitative comparison with observations, but qualitatively, these results seem consistent with the relatively high metallicities detected in high-redshift galaxies (e.g. Erb et al. 2006). As expected, the *WMAP3* model predicts somewhat lower metallicity at high redshift.

3.5.4 The baryon budget and its evolution

In Fig. 15, we provide the answer to the question ‘where are the baryons’ in our fiducial isothermal Bondi model, in both the C- Λ CDM and *WMAP3* cosmologies. All quantities in Fig. 15 are shown in units of the critical density. We show the mass density in collapsed DM haloes above our mass resolution of $M_{\text{vir}} = 10^{10} M_{\odot}$ multiplied by the universal baryon fraction, which represents all of the baryons that are available for cooling at a given redshift in our model. We also show the mass density of hot gas in DM haloes, and of warm/hot diffuse gas that has either been prevented from collapsing into haloes by the photoionizing background or has been ejected by SN feedback. These two components (warm/hot diffuse gas and hot gas in haloes) dominate the baryon budget at all redshifts, in agreement with observational constraints at low redshift (e.g. Fukugita & Peebles 2004) and the predictions of numerical hydrodynamic simulations (Bertone, Schaye & Dolag 2008, and references therein). Finally, we show the baryons in cold gas in galactic discs, in stars in the discs and bulges of galaxies and in DSHs around galaxies. Cold gas dominates over stars until $z \sim 2$, where stars begin to dominate. Our model predicts that stars in DSH comprise about 15 per cent of the mass of stars in galaxies at $z \sim 0$.

From this plot we can see that, in the *WMAP3* model, there is just enough baryonic material in collapsed haloes that *can* cool at $z \gtrsim 3$ to account for the DLAS data. If some of the gas that is ejected from galaxies in our model instead remains in galactic discs in the form of cold gas, perhaps this model could be reconciled with the DLAS data. Note, however, that there is no room for more cold gas

in galaxies at $z = 0$, so in order to solve the problem, the gas needs to be retained at high redshift but ejected or prevented from cooling at low redshift. One can think of plausible physical reasons that SN-driven winds might have more difficulty escaping galaxies at high redshift, for example, the winds might stall against the higher density IGM that is presumably present at these early epochs.

4 DISCUSSION AND CONCLUSIONS

We have presented a new semi-analytic model for the self-consistent evolution of galaxies, black holes and AGN in the framework of the CDM model of structure formation. Our models are built on those described in SP99, SPF2001 and subsequent papers, but we present several important new ingredients here.

- (i) Improved modelling of tidal stripping and destruction of orbiting subhaloes and of dynamical friction and satellite merger time-scales.
- (ii) Improved modelling of disc sizes, including realistic DM halo profiles and the effect of ‘adiabatic contraction’.
- (iii) A more realistic recipe for star formation in quiescent discs, based on the empirical Kennicutt law and including a surface density threshold for star formation.
- (iv) Updated modelling of the efficiency and time-scale of starbursts, based on an extensive suite of numerical hydrodynamic simulations of galaxy mergers.
- (v) Tracking of a ‘DSH’ component, which is built up of tidally destroyed satellites and stars scattered in mergers.
- (vi) A self-regulated model for black hole growth and ‘bright mode’ accretion ‘light curves’ based on numerical merger simulations with AGN feedback.
- (vii) Galaxy-scale AGN-driven winds, based on numerical merger simulations.
- (viii) Fuelling of black holes with hot gas via Bondi accretion, regulated according to the isothermal cooling flow model of NF00.
- (ix) Heating by radio jets, calibrated against observations of X-ray cavities in cooling flow clusters.

We explored the predictions of our new models for a broad range of physical properties of galaxies, groups and clusters at $z \sim 0$. One key result is an explanation of the origin of the characteristic shape of the galaxy stellar mass or luminosity function. We cast this in

terms of the ‘star formation efficiency function’ (i.e. the fraction of baryons in stars as a function of host halo mass), which can be derived empirically by mapping between (sub)halo mass and stellar mass such that the observed stellar mass function is reproduced (Wang et al. 2006; Moster et al., in preparation). We found that, in our models, the shape of this function arises from the fact that SN-driven winds can more efficiently heat and expel gas in lower mass haloes, while radio mode heating by AGN is more efficient in higher mass haloes, because these haloes have larger mass black holes and these black holes can accrete hot gas more efficiently. The peak in this function at $M_{\text{vir}} \sim 10^{12} M_{\odot}$ occurs because these haloes are too massive for SN-driven winds to escape easily, and do not have massive enough black holes or a large enough virial temperature to fuel efficient radio mode heating.

We showed that our model also reproduces the stellar mass function as a function of galaxy morphology (at least for galaxies more massive than $\sim 10^{10} M_{\odot}$), the cold gas fractions of disc galaxies as a function of stellar mass and the cold gas mass function, the stellar mass–metallicity relation for galaxies, and the BH mass versus spheroid mass relation.

Our model for heating by radio jets is similar in concept, but different in detailed implementation, to other models that have been presented in the literature. We tested the basic assumption of our model for fuelling of BH by hot gas, the isothermal cooling flow model proposed by NF00, using observations of central temperatures and densities in hot X-ray emitting gas in nine systems by A06, and found consistency. We further compared the jet power required to solve the overcooling/massive galaxy problem in our models with direct measurements of jet power from the energetics of bubbles detected in X-ray gas around elliptical galaxies (A06; Rafferty et al. 2006). We found that our required jet powers lie above the observations of A06 but agree with the higher values obtained by Rafferty et al. (2006).

We compared the results of our fiducial model with a very simple implementation of AGN heating, in which we simply switched off cooling in haloes more massive than $\sim 10^{12} M_{\odot}$ (the HQ model). We found that this model produced very similar results to our fiducial model for the $z = 0$ stellar mass fraction as a function of halo mass [$f_{\text{star}}(M_{\text{halo}})$], the stellar mass function, the cold gas fractions and cold gas mass function, and the global star formation and stellar mass assembly histories.

We investigated whether the same models which provided a good match to the stellar mass function also reproduced the distribution of (specific) SFRs as a function of stellar mass. We found that our fiducial model qualitatively reproduces the main features of the observed distribution of SSFR versus stellar mass: the models produce a star-forming sequence, which dominates at low stellar masses ($\lesssim 2\text{--}3 \times 10^{10} M_{\odot}$) and a quenched population at high stellar masses. The transition mass is naturally predicted by the model, and again corresponds to the halo mass scale where AGN heating becomes effective. This is in marked contrast to models without AGN feedback, in which all massive galaxies are actively star forming.

However, a more detailed comparison with the observations indicates that the low-mass star-forming sequence in the models occurs at too low an SSFR, and does not have the right slope (our star-forming sequence is flat in SSFR versus stellar mass space, rather than tilted such that low-mass galaxies have high SSFR, as the observations indicate). This problem is present in all of our models, and is independent of AGN feedback and robust to the star formation recipe and the values of our free parameters. It may be a symptom of the same malady that is responsible for producing low-

mass galaxies with older ages than those estimated from absorption lines in the spectra of nearby galaxies.

In the HQ model, the quenching is clearly too sharp a function of stellar mass. Essentially all galaxies more massive than $m_{\text{star}} \sim 10^{11} M_{\odot}$ are completely quenched, while the observations indicate that massive galaxies have a wide range of SSFRs, from small to moderate amounts of recent star formation to activity levels that place them on the star-forming sequence.

We investigated the cosmic histories of the major baryonic components of the Universe as predicted in our models: star formation and stellar mass, cold gas, warm/hot gas and metals. We found that our fiducial Λ -CDM model produces very good agreement with the global SFR density at $z \lesssim 2$, but predicts a higher and flatter SF history at $2 \lesssim z \lesssim 6$ than most observations indicate. We found that our prediction of the global SFR density is apparently not affected by AGN feedback at redshifts above $z \gtrsim 3$, because most star formation is taking place in small-mass haloes.

All of the Λ -CDM models predict a significantly larger amount of mass in long-lived stars in galaxies at redshifts $z \gtrsim 0.5$, by about a factor of 2 at $z \sim 1$ and a factor of 3 at $z \sim 2$ for the fiducial and HQ models. The fact that we reproduce the stellar mass density at $z \sim 0$ (by construction) then implies that there is not enough evolution in the galaxy population between $z \sim 1$ and 0 in our models with AGN feedback. This problem does not seem to be specific to our models, but is common to all the recently published models in which AGN feedback is implemented in a similar way (Bower et al. 2006; Cattaneo et al. 2006; Croton et al. 2006). Indeed, one can see from the predictions of the global SFR history presented in these works (e.g. fig. 5 of Croton et al. 2006, fig. 8 of Bower et al. 2006) that these models all produce very similar predictions for the global SF history and for the integrated stellar mass density.

The *WMAP3* model, in which structure formation occurs later due to the reduced power on small scales, has a much lower and more steeply declining global SFR at $z \gtrsim 2$ (about an order of magnitude lower than Λ -CDM at $z = 6$). The SFR is lower than the observational compilation of Hopkins & Beacom (2006) at these redshifts by about 0.2–0.3 dex, but still well within the observational errors. In contrast to the Λ -CDM model, the stellar mass density predicted by the *WMAP3* model is in excellent agreement with observational estimates from $z \sim 4\text{--}0$. The delayed SF and stellar mass assembly history in *WMAP3* relative to Λ -CDM has also been illustrated by Wang et al. (2007).

We also compared the predicted global mass density of cold gas in galactic discs Ω_{cold} as a function of redshift in our models with estimates of H I gas mass density at $2 \lesssim z \lesssim 4.5$ from quasar absorption systems (damped Lyman α systems and Lyman-limit systems). We found that our Λ -CDM models had no difficulty producing enough cold gas in discs up to $z \sim 4$, where the observations are the most secure, but the *WMAP3* model did not fare so well. The predicted values of Ω_{cold} in this model were too low by a factor of 2–10 at $z \gtrsim 3$.

We analyse the redshift-dependent breakdown of all the baryons in our fiducial model into each of the various components that we track: hot gas in haloes, warm/hot diffuse gas in the IGM, cold gas in galactic discs, stars in galaxies and stars in DSHs. We find that hot gas in haloes and warm/hot gas in the IGM dominate at all redshifts, in agreement with the predictions of numerical cosmological simulations (Bertone et al. 2008, and references therein) and with the observational baryon census at low redshift (Fukugita & Peebles 2004). Therefore, in order to produce more cold gas at high redshift, we either require more efficient cooling of hot gas or less efficient reheating and ejection of cold gas by SN feedback. Since our models

are already overproducing stars at high redshift, the latter is probably a more promising solution. Our models predict efficient early metal enrichment, with the average stellar mass-weighted stellar metallicity reaching 25 per cent of solar at $z \sim 6$ and 50 per cent of solar at $z \sim 2.5$.

The picture of the cosmic build-up of stars that we see in our models is interesting in the context of a problem pointed out in several recent papers (e.g. Hopkins & Beacom 2006; Fardal et al. 2007; Davé 2008; Wilkins et al. 2008): when the best available observational estimate of the dust-corrected and incompleteness-corrected SFR density is integrated, accounting for gas recycling under the assumption of a universal stellar IMF, the mass of long-lived stars is overestimated by about a factor of 2–3. These authors suggest that a possible resolution to this problem is a non-universal IMF, which was more top-heavy at high redshift. In the context of the two models we have considered, the Λ -CDM models predict early structure formation, accompanied by a lot of early star formation. We might be able to reconcile these models with all the data if in fact the IMF was top-heavy at early times, so that most of the star formation that we see does not produce long-lived stars. On the other hand, the *WMAP3* model implies later structure formation, and hence less star formation at high redshift. This provides an alternate means of reconciling observations of the SF history and stellar mass assembly history, which requires only that the current observational estimates of the SFR at $z \gtrsim 2$ are too high by about a factor of 2–3. While the *WMAP3* picture seems more attractive in many respects, it is a concern that the *WMAP3* models do not seem to be able to account for DLAS at $z \gtrsim 3$, again as a consequence of the reduced small-scale power. Better constraints on the amount of cold gas at high redshift from new facilities such as ALMA could help to resolve this question.

In agreement with previous work, we have shown that the inclusion of AGN feedback in semi-analytic models can plausibly solve the overcooling problem, the massive galaxy problem, and the star formation quenching problem in the local universe – a huge step forward. However, we have also shown that several potentially serious discrepancies still remain, and we have argued that these discrepancies are not peculiar to our implementation, but are common to all of the CDM-based semi-analytic models currently on the market. These discrepancies are connected with low-mass galaxies ($m_* \lesssim \text{few} \times 10^{10} M_\odot$), however, in the picture that we are developing, small galaxies grow into massive galaxies, and the growth of galaxies and AGN are intimately interconnected, so these problems on small-scales may indicate or cause more pervasive problems. It is likely that these problems are connected to the modelling of cooling, star formation and/or SN feedback, or possibly to the CDM power spectrum on small scales. *Direct* AGN feedback⁵ is probably not the solution.

In this paper, we have focused on predictions of the physical properties of galaxies at $z \sim 0$ and the global evolution of the major baryonic components of the Universe over time. In a planned series of papers, we will investigate the predictions of the models we have presented here for multiwavelength, observable properties (e.g. UV through FIR luminosities and colours) of galaxies at low and high redshifts, and examine in more detail the distribution functions and scaling relations of intrinsic galaxy properties (e.g. stellar mass, SFR, metallicity, etc.) at high redshift. In addition, we will explore the predictions of our models for AGN properties as a function of

redshift and environment, and the relationship between AGN and their host galaxies.

ACKNOWLEDGMENTS

We would like to thank E. Bell, B. Panter and D. Schiminovich for providing us with their data in electronic form. We warmly thank B. Allgood, E. Bell, J. Bromley, D. Croton, G. de Lucia, A. Dekel, M. Elvis, S. Faber, A. Fabian, S. Jester, A. Kravtsov, C. Martin, L. Moustakas, P. Natarajan, H.-W. Rix, S. Trager, R. Wechsler and A. Walen for discussions that contributed to this work, and S. Allen, A. Gonzalez and S. Zibetti for help interpreting their observational results. We also thank E. Bell and S. Trager for careful readings of an earlier draft of the manuscript, and B. Moster for providing us with his results in advance of publication. RSS thanks the ITC at the CfA for hospitality. BER gratefully acknowledges support from a Spitzer Fellowship through a NASA grant administered by the Spitzer Science Center. This work was supported in part by a grant from the W. M. Keck Foundation.

REFERENCES

- Abel T., Bryan G. L., Norman M. L., 2002, *Sci*, 295, 93
 Allen S. W., Schmidt R. W., Ebeling H., Fabian A. C., van Speybroeck L., 2004, *MNRAS*, 353, 457
 Allen S. W., Dunn R. J. H., Fabian A. C., Taylor G. B., Reynolds C. S., 2006, *MNRAS*, 372, 21 (A06)
 Arnaud M., Rothenflug R., Boulade O., Vigroux L., Vangioni-Flam E., 1992, *A&A*, 254, 49
 Bacon D. J. et al., 2005, *MNRAS*, 363, 723
 Baldry I. K., Glazebrook K., Brinkmann J., Ivezić Ž., Lupton R. H., Nichol R. C., Szalay A. S., 2004, *ApJ*, 600, 681
 Balogh M. L., Pearce F. R., Bower R. G., Kay S. T., 2001, *MNRAS*, 326, 1228
 Barnes J. E., 1988, *ApJ*, 331, 699
 Barnes J. E., 1992, *ApJ*, 393, 484
 Barnes J., Hernquist L., 1996, *ApJ*, 471, 115
 Barton E. J., Geller M. J., Kenyon S. J., 2000, *ApJ*, 530, 660
 Barton E. J., Arnold J. A., Zentner A. R., Bullock J. S., Wechsler R. H., 2007, *ApJ*, 671, 1538
 Baugh C. M., 2006, *Rep. Prog. Phys.*, 69, 3101
 Bell E. F., McIntosh D. H., Katz N., Weinberg M. D., 2003a, *ApJ*, 585, L117
 Bell E. F., McIntosh D. H., Katz N., Weinberg M. D., 2003b, *ApJS*, 149, 289
 Benson A. J., Lacey C. G., Baugh C. M., Cole S., Frenk C. S., 2002, *MNRAS*, 333, 156
 Benson A. J., Bower R. G., Frenk C. S., Lacey C. G., Baugh C. M., Cole S., 2003, *ApJ*, 599, 38
 Bertone S., Schaye J., Dolag K., 2008, *Space Science Reviews*, 134, 295
 Best P. N., Kauffmann G., Heckman T. M., Brinchmann J., Charlot S., Ivezić Ž., White S. D. M., 2005, *MNRAS*, 362, 25
 Best P. N., Kaiser C. R., Heckman T. M., Kauffmann G., 2006, *MNRAS*, 368, L67
 Best P. N., von der Linden A., Kauffmann G., Heckman T. M., Kaiser C. R., 2007, *MNRAS*, 379, 894
 Binney J., 1977, *ApJ*, 215, 483
 Binney J., 2004, *MNRAS*, 347, 1093
 Binney J., Tabor G., 1995, *MNRAS*, 276, 663
 Binney J., Bibi F. A., Omma H., 2007, *MNRAS*, 377, 142
 Birnboim Y., Dekel A., 2003, *MNRAS*, 345, 349
 Bîrzan L., Rafferty D. A., McNamara B. R., Wise M. W., Nulsen P. E. J., 2004, *ApJ*, 607, 800
 Blandford R. D., Begelman M. C., 1999, *MNRAS*, 303, L1
 Blumenthal G. R., Faber S. M., Primack J. R., Rees M., 1984, *Nat*, 311, 517
 Blumenthal G., Faber S. M., Flores R., Primack J. R., 1986, *ApJ*, 301, 27
 Bondi H., 1952, *MNRAS*, 112, 195

⁵ By this we mean feedback by an AGN within the galaxy itself. Indirect feedback from AGN in external galaxies, e.g. via pre-heating, may be a promising solution (e.g. Scannapieco & Oh 2004).

- Borch A. et al., 2006, *A&A*, 453, 869
- Borgani S. et al., 2001, *ApJ*, 561, 13
- Borgani S. et al., 2006, *MNRAS*, 367, 1641
- Bower R. G., Benson A. J., Malbon R., Helly J. C., Frenk C. S., Baugh C. M., Cole S., Lacey C. G., 2006, *MNRAS*, 370, 645
- Boylan-Kolchin M., Ma C.-P., Quataert E., 2008, *MNRAS*, 383, 93
- Brinchmann J., Charlot S., White S. D. M., Tremonti C., Kauffmann G., Heckman T., Brinkmann J., 2004, *MNRAS*, 351, 1151
- Broeils A. H., Rhee M.-H., 1997, *A&A*, 324, 877
- Bromley J. M., Somerville R. S., Fabian A. C., 2004, *MNRAS*, 350, 456
- Brotherton M. S. et al., 1999, *ApJ*, 520, L87
- Bruzual G., Charlot S., 2003, *MNRAS*, 344, 1000
- Bullock J. S., Dekel A., Kolatt T. S., Kravtsov A. V., Klypin A. A., Porciani C., Primack J. R., 2001a, *ApJ*, 555, 240
- Bullock J. S. et al., 2001b, *MNRAS*, 321, 559
- Canalizo G., Stockton A., 2001, *ApJ*, 555, 719
- Cattaneo A., Dekel A., Devriendt J., Guiderdoni B., Blaizot J., 2006, *MNRAS*, 370, 1651
- Cattaneo A. et al., 2007, *MNRAS*, 377, 63
- Chabrier G., 2003, *PASP*, 115, 763
- Chartas G., Brandt W. N., Gallagher S. C., 2003, *ApJ*, 595, 85
- Chartas G., Brandt W. N., Gallagher S. C., Proga D., 2007, *AJ*, 133, 1849
- Churazov E., Sunyaev R., Forman W., Böhringer H., 2002, *MNRAS*, 332, 729
- Churazov E., Sazonov S., Sunyaev R., Forman W., Jones C., Böhringer H., 2005, *MNRAS*, 363, L91
- Cole S., Aragón-Salamanca A., Frenk C. S., Navarro J. F., Zepf S. E., 1994, *MNRAS*, 271, 781
- Cole S., Lacey C. G., Baugh C. M., Frenk C. S., 2000, *MNRAS*, 319, 168
- Colina L. et al., 2001, *ApJ*, 563, 546
- Conroy C., Wechsler R. H., Kravtsov A. V., 2006, *ApJ*, 647, 201
- Conroy C., Wechsler R. H., Kravtsov A. V., 2007, *ApJ*, 668, 826
- Cowie L. L., Binney J., 1977, *ApJ*, 215, 723
- Cox T. J., Di Matteo T., Hernquist L., Hopkins P. F., Robertson B., Springel V., 2006a, *ApJ*, 643, 692
- Cox T. J., Dutta S. N., Di Matteo T., Hernquist L., Hopkins P. F., Robertson B., Springel V., 2006b, *ApJ*, 650, 791
- Cox T. J., Jonsson P., Somerville R. S., Primack J. R., Dekel A., 2008, *MNRAS*, 384, 386 (C08)
- Croton D. J. et al., 2006, *MNRAS*, 365, 11
- Davé R., 2008, *MNRAS*, 385, 147
- de Kool M., Arav N., Becker R. H., Gregg M. D., White R. L., Laurent-Muehleisen S. A., Price T., Korista K. T., 2001, *ApJ*, 548, 609
- De Lucia G., Kauffmann G., White S. D. M., 2004, *MNRAS*, 349, 1101
- Dekel A., Birnboim Y., 2006, *MNRAS*, 368, 2
- Dekel A., Birnboim Y., 2008, *MNRAS*, 383, 119
- Dekel A., Silk J., 1986, *ApJ*, 303, 39
- Desjacques V., Nusser A., 2005, *MNRAS*, 361, 1257
- Desroches L.-B., Quataert E., Ma C.-P., West A. A., 2007, *MNRAS*, 377, 402
- Di Matteo T., Springel V., Hernquist L., 2005, *Nat*, 433, 604
- Dunn R. J. H., Fabian A. C., 2006, *MNRAS*, 373, 959
- Dunn R. J. H., Fabian A. C., 2008, *MNRAS*, 385, 757
- Efstathiou G., 1992, *MNRAS*, 256, 43p
- Efstathiou G., Rees M. J., 1988, *MNRAS*, 230, 5p
- Eisenstein D. J. et al., 2005, *ApJ*, 633, 560
- Erb D. K., Shapley A. E., Pettini M., Steidel C. C., Reddy N. A., Adelberger K. L., 2006, *ApJ*, 644, 813
- Fabian A. C., 1994, *ARA&A*, 32, 277
- Fabian A. C., Nulsen P. E. J., 1977, *MNRAS*, 180, 479
- Fabian A. C., Sanders J. S., Allen S. W., Crawford C. S., Iwasawa K., Johnstone R. M., Schmidt R. W., Taylor G. B., 2003, *MNRAS*, 344, L43
- Fabian A. C., Sanders J. S., Taylor G. B., Allen S. W., Crawford C. S., Johnstone R. M., Iwasawa K., 2006, *MNRAS*, 366, 417
- Fardal M. A., Katz N., Weinberg D. H., Davé R., 2007, *MNRAS*, 379, 985
- Farrar D. et al., 2001, *MNRAS*, 326, 1333
- Fender R. P., Belloni T. M., Gallo E., 2004, *MNRAS*, 355, 1105
- Ferrarese L., Merritt D., 2000, *ApJ*, 539, L9
- Flores R., Primack J. R., Blumenthal G., Faber S. M., 1993, *ApJ*, 412, 443
- Fontana A. et al., 2006, *A&A*, 459, 745
- Forman W. et al., 2007, *ApJ*, 665, 1057
- Fukugita M., Peebles P. J. E., 2004, *ApJ*, 616, 643
- Gallazzi A., Charlot S., Brinchmann J., White S. D. M., Tremonti C. A., 2005, *MNRAS*, 362, 41
- Gallazzi A., Brinchmann J., Charlot S., White S. D. M., 2008, *MNRAS*, 383, 1439
- Ganguly R., Brotherton M. S., 2008, *ApJ*, 672, 102
- Gebhardt K. et al., 2000, *ApJ*, 539, L13
- Giavalisco M. et al., 2004, *ApJ*, 600, L103
- Gnedin N. Y., 2000, *ApJ*, 542, 535 (G00)
- Gonzalez A. H., Zabludoff A. I., Zaritsky D., 2005, *ApJ*, 618, 195
- Häring N., Rix H.-W., 2004, *ApJ*, 604, L89
- Hernquist L., 1989, *Nat*, 340, 687
- Hernquist L., 1992, *ApJ*, 400, 460
- Hernquist L., 1993, *ApJ*, 409, 548
- Heymans C. et al., 2005, *MNRAS*, 361, 160
- Hoekstra H. et al., 2006, *ApJ*, 647, 116
- Hopkins A. M., 2004, *ApJ*, 615, 209
- Hopkins A. M., Beacom J. F., 2006, *ApJ*, 651, 142
- Hopkins P. F., Hernquist L., Cox T. J., Di Matteo T., Martini P., Robertson B., Springel V., 2005a, *ApJ*, 630, 705
- Hopkins P. F., Hernquist L., Cox T. J., Di Matteo T., Robertson B., Springel V., 2005b, *ApJ*, 630, 716
- Hopkins P. F., Hernquist L., Cox T. J., Di Matteo T., Robertson B., Springel V., 2005c, *ApJ*, 632, 81
- Hopkins P. F., Hernquist L., Martini P., Cox T. J., Robertson B., Di Matteo T., Springel V., 2005d, *ApJ*, 625, L71
- Hopkins P. F., Hernquist L., Cox T. J., Di Matteo T., Robertson B., Springel V., 2006a, *ApJS*, 163, 1
- Hopkins P. F., Hernquist L., Cox T. J., Robertson B., Di Matteo T., Springel V., 2006b, *ApJ*, 639, 700
- Hopkins P. F., Hernquist L., Cox T. J., Robertson B., Springel V., 2006c, *ApJS*, 163, 50
- Hopkins P. F., Somerville R. S., Hernquist L., Cox T. J., Robertson B., Li Y., 2006d, *ApJ*, 652, 864
- Hopkins P. F., Hernquist L., Cox T. J., Robertson B., Krause E., 2007a, *ApJ*, 669, 45
- Hopkins P. F., Hernquist L., Cox T. J., Robertson B., Krause E., 2007b, *ApJ*, 669, 67
- Hopkins P. F., Hernquist L., Cox T. J., Kereš D., 2008, *ApJS*, 175, 356
- Jahnke K. et al., 2004, *ApJ*, 614, 568
- Jena T. et al., 2005, *MNRAS*, 361, 70
- Jester S., 2005, *ApJ*, 625, 667
- Kang X., Jing Y. P., Silk J., 2006, *ApJ*, 648, 820
- Kannappan S. J., 2004, *ApJ*, 611, L89
- Kauffmann G., Haehnelt M., 2000, *MNRAS*, 311, 576
- Kauffmann G., White S. D. M., Guiderdoni B., 1993, *MNRAS*, 264, 201
- Kauffmann G., Colberg J. M., Diaferio A., White S. D. M., 1998, *MNRAS*, 303, 188
- Kauffmann G. et al., 2003a, *MNRAS*, 341, 54
- Kauffmann G. et al., 2003b, *MNRAS*, 346, 1055
- Kauffmann G., Heckman T. M., Best P. N., 2008, *MNRAS*, 384, 953
- Kaviraj S. et al., 2007, *ApJS*, 173, 619
- Kennicutt R. C., 1989, *ApJ*, 344, 685
- Kennicutt R. C., 1998, *ApJ*, 498, 181
- Keres D., Yun M. S., Young J. S., 2003, *ApJ*, 582, 659
- Kereš D., Katz N., Weinberg D. H., Davé R., 2005, *MNRAS*, 363, 2
- Khochfar S., Ostriker J. P., 2007, *ApJ*, 680, 54
- Kollmeier J. A. et al., 2006, *ApJ*, 648, 128
- Komatsu E. et al., 2008, *ApJS*, in press (arXiv:0803.0547)
- Körding E. G., Jester S., Fender R., 2006, *MNRAS*, 372, 1366
- Kormendy J., Richstone D., 1995, *ARA&A*, 33, 581
- Kravtsov A. V., Berlind A. A., Wechsler R. H., Klypin A. A., Gottlöber S., Allgood B., Primack J. R., 2004a, *ApJ*, 609, 35
- Kravtsov A. V., Gnedin O. Y., Klypin A. A., 2004b, *ApJ*, 609, 482

- Krongold Y., Nicastro F., Elvis M., Brickhouse N., Binette L., Mathur S., Jiménez-Bailón E., 2007, *ApJ*, 659, 1022
- Lauer T. R. et al., 2007, *ApJ*, 662, 808
- Li C., Kauffmann G., Heckman T., Jing Y. P., White S. D. M., 2007, *MNRAS*, 385, 1903
- Lin L. et al., 2007, *ApJ*, 660, L51
- Maccarone T. J., Gallo E., Fender R., 2003, *MNRAS*, 345, L19
- Macciò A. V., Dutton A. A., van den Bosch F. C., Moore B., Potter D., Stadel J., 2007, *MNRAS*, 378, 55
- Magorrian J. et al., 1998, *AJ*, 115, 2285
- Maller A. H., Bullock J. S., 2004, *MNRAS*, 355, 694
- Marconi A., Hunt L. K., 2003, *ApJ*, 589, L21
- Martin C. L., 1999, *ApJ*, 513, 156
- Martin C. L., Kennicutt R. C., 2001, *ApJ*, 555, 301
- Martini P., Schneider D. P., 2003, *ApJ*, 597, L109
- Martini P., Weinberg D. H., 2001, *ApJ*, 547, 12
- Mathews W. G., Bregman J. N., 1978, *ApJ*, 224, 308
- McNamara B. R., Nulsen P. E. J., 2007, *ARA&A*, 45, 117
- McNamara B. R., Nulsen P. E. J., Wise M. W., Rafferty D. A., Carilli C., Sarazin C. L., Blanton E. L., 2005, *Nat*, 433, 45
- McNamara B. R. et al., 2006, *ApJ*, 648, 164
- Menci N., Fontana A., Giallongo E., Grazian A., Salimbeni S., 2006, *ApJ*, 647, 753
- Mihos J. C., Hernquist L., 1994, *ApJ*, 425, 13
- Mihos J. C., Hernquist L., 1996, *ApJ*, 464, 641
- Mo H. J., Mao S., White S. D. M., 1998, *MNRAS*, 295, 319 (MMW98)
- Monaco P., Murante G., Borgani S., Fontanot F., 2006, *ApJ*, 652, L89
- Monaco P., Fontanot F., Taffoni G., 2007, *MNRAS*, 375, 1189
- Murante G. et al., 2004, *ApJ*, 607, L83
- Murante G., Giovali M., Gerhard O., Arnaboldi M., Borgani S., Dolag K., 2007, *MNRAS*, 377, 2
- Naab T., Johansson P. H., Ostriker J. P., Efstathiou G., 2007, *ApJ*, 658, 710
- Narayan R., Yi I., 1994, *ApJ*, 428, L13
- Navarro J. F., Frenk C. S., White S. D. M., 1997, *ApJ*, 490, 493
- Nulsen P. E. J., Fabian A. C., 2000, *MNRAS*, 311, 346 (NF00)
- Nulsen P. E. J., Hambrick D. C., McNamara B. R., Rafferty D., Birzan L., Wise M. W., David L. P., 2005a, *ApJ*, 625, L9
- Nulsen P. E. J., McNamara B. R., Wise M. W., David L. P., 2005b, *ApJ*, 628, 629
- Omnia H., Binney J., Bryan G., Slyz A., 2004, *MNRAS*, 348, 1105
- Panter B., Jimenez R., Heavens A. F., Charlot S., 2007, *MNRAS*, 378, 1550
- Peebles P. J. E., 1969, *ApJ*, 155, 393
- Percival W. J. et al., 2002, *MNRAS*, 337, 1068
- Peterson J. R., Fabian A. C., 2006, *Phys. Rep.*, 427, 1
- Peterson J. R., Kahn S. M., Paerels F. B. S., Kaastra J. S., Tamura T., Bleeker J. A. M., Ferrigno C., Jernigan J. G., 2003, *ApJ*, 590, 207
- Pounds K. A., Page K. L., 2006, *MNRAS*, 372, 1275
- Pounds K. A., King A. R., Page K. L., O'Brien P. T., 2003, *MNRAS*, 346, 1025
- Prochaska J. X., Herbert-Fort S., Wolfe A. M., 2005, *ApJ*, 635, 123
- Purcell C. W., Bullock J. S., Zentner A. R., 2007, *ApJ*, 666, 20
- Quinn T., Katz N., Efstathiou G., 1996, *MNRAS*, 278, L49
- Rafferty D. A., McNamara B. R., Nulsen P. E. J., Wise M. W., 2006, *ApJ*, 652, 216
- Rees M. J., Ostriker J. P., 1977, *MNRAS*, 179, 541
- Roberts M. S., Haynes M. P., 1994, *ARA&A*, 32, 115
- Robertson B., Bullock J. S., Cox T. J., Di Matteo T., Hernquist L., Springel V., Yoshida N., 2006a, *ApJ*, 645, 986
- Robertson B., Cox T. J., Hernquist L., Franx M., Hopkins P. F., Martini P., Springel V., 2006b, *ApJ*, 641, 21
- Robertson B., Hernquist L., Cox T. J., Di Matteo T., Hopkins P. F., Martini P., Springel V., 2006c, *ApJ*, 641, 90
- Rosenberg J. L., Schneider S. E., 2002, *ApJ*, 567, 247
- Salim S. et al., 2007, *ApJS*, 173, 267
- Sánchez S. F. et al., 2004, *ApJ*, 614, 586
- Sanders D. B., Mirabel I. F., 1996, *ARA&A*, 34, 749
- Scannapieco E., Oh S. P., 2004, *ApJ*, 608, 62
- Schawinski K. et al., 2006, *Nat*, 442, 888
- Schaye J., 2004, *ApJ*, 609, 667
- Schiminovich D. et al., 2005, *ApJ*, 619, L47
- Schiminovich D. et al., 2007, *ApJS*, 173, 315
- Shakura N. I., Syunyaev R. A., 1973, *A&A*, 24, 337
- Sheth R. K., Tormen G., 1999, *MNRAS*, 308, 119
- Sijacki D., Springel V., di Matteo T., Hernquist L., 2007, *MNRAS*, 380, 877
- Silk J., 1977, *ApJ*, 211, 638
- Silk J., Rees M. J., 1998, *A&A*, 331, L1
- Somerville R. S., 2002, *ApJ*, 572, L23
- Somerville R. S., Kolatt T. S., 1999, *MNRAS*, 305, 1 (SK99)
- Somerville R. S., Primack J. R., 1999, *MNRAS*, 310, 1087
- Somerville R. S., Primack J. R., Faber S. M., 2001, *MNRAS*, 320, 504 (SPF01)
- Somerville R. S. et al., 2008, *ApJ*, 672, 776 (S08)
- Spergel D. N. et al., 2003, *ApJS*, 148, 175
- Spergel D. N. et al., 2007, *ApJS*, 170, 377
- Springel V., White S. D. M., Tormen G., Kauffmann G., 2001, *MNRAS*, 328, 726
- Springel V., Di Matteo T., Hernquist L., 2005a, *ApJ*, 620, L79
- Springel V., Di Matteo T., Hernquist L., 2005b, *MNRAS*, 361, 776
- Steenbrugge K. C. et al., 2005, *A&A*, 434, 569
- Steidel C., Adelberger K., Giavalisco M., Dickinson M., Pettini M., 1999, *ApJ*, 519, 1
- Sutherland R., Dopita M. A., 1993, *ApJS*, 88, 253
- Taylor J. E., Babul A., 2004, *MNRAS*, 348, 811
- Tegmark M. et al., 2004, *ApJ*, 606, 702
- Thoul A. A., Weinberg D. H., 1996, *ApJ*, 465, 608
- Toomre A., Toomre J., 1972, *ApJ*, 178, 623
- Tremaine S. et al., 2002, *ApJ*, 574, 740
- Vanden Berk D. E. et al., 2006, *AJ*, 131, 84
- Vestergaard M., 2004, *ApJ*, 601, 676
- Vikhlinin A., Kravtsov A., Forman W., Jones C., Markevitch M., Murray S. S., Van Speybroeck L., 2006, *ApJ*, 640, 691
- Voigt L. M., Fabian A. C., 2004, *MNRAS*, 347, 1130
- Volonteri M., Rees M. J., 2005, *ApJ*, 633, 624
- Volonteri M., Haardt F., Madau P., 2003, *ApJ*, 582, 559
- von der Linden A., Best P. N., Kauffmann G., White S. D. M., 2007, *MNRAS*, 379, 867
- Wang J., De Lucia G., Kitzbichler M. G., White S. D. M., 2007, *MNRAS*, 384, 1301
- Wang L., Li C., Kauffmann G., De Lucia G., 2006, *MNRAS*, 371, 537
- Wechsler R. H., Bullock J. S., Primack J. R., Kravtsov A. V., Dekel A., 2002, *ApJ*, 568, 52
- White S. D. M., Frenk C. S., 1991, *ApJ*, 379, 52
- White S. D. M., Rees M. J., 1978, *MNRAS*, 183, 341
- White S. D. M., Navarro J. F., Evrard A. E., Frenk C. S., 1993, *Nat*, 366, 429
- Wilkins S. M., Trentham N., Hopkins A. M., 2008, *MNRAS*, 385, 687
- Woods D. F., Geller M. J., 2007, *AJ*, 134, 527
- Woods D. F., Geller M. J., Barton E. J., 2006, *AJ*, 132, 197
- Wyithe J. S. B., 2006, *MNRAS*, 365, 1082
- Wyithe J. S. B., Loeb A., 2002, *ApJ*, 581, 886
- Yi S. K. et al., 2005, *ApJ*, 619, L111
- Yoshida N., Stoehr F., Springel V., White S. D. M., 2002, *MNRAS*, 335, 762
- Zentner A. R., Bullock J. S., 2003, *ApJ*, 598, 49
- Zibetti S., 2008, in Davies J., Disney M., eds, *IAU Symp. 244, Statistical Properties of the IntraCluster Light from SDSS Image Stacking*. Cambridge Univ. Press, Cambridge, p. 176
- Zibetti S., White S. D. M., Schneider D. P., Brinkmann J., 2005, *MNRAS*, 358, 949
- Zwaan M. A., Meyer M. J., Staveley-Smith L., Webster R. L., 2005, *MNRAS*, 359, L30

This paper has been typeset from a $\text{\TeX}/\text{\LaTeX}$ file prepared by the author.