Mon. Not. R. Astron. Soc. **365,** 11–28 (2006) doi:10.1111/j.1365-2966.2005.09675.x

The many lives of active galactic nuclei: cooling flows, black holes and the luminosities and colours of galaxies

Darren J. Croton,[1⋆] Volker Springel,[1] Simon D. M. White,[1] G. De Lucia,[1] C. S. Frenk,[2] L. Gao,[1] A. Jenkins,[2] G. Kauffmann,[1] J. F. Navarro[3] and N. Yoshida[4]

[1]*Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Str. 1, D-85740 Garching, Germany*
[2]*Institute for Computational Cosmology, Physics Department, Durham University, South Road, Durham DH1 3LE*
[3]*Department of Physics and Astronomy, University of Victoria, PO Box 3055 STN CSC, Victoria, BC, V8W 3P6, Canada*
[4]*Department of Physics, Nagoya University, Chikusa-ku, Nagoya 464-8602, Japan*

Accepted 2005 September 30. Received 2005 September 30; in original form 2005 August 12

ABSTRACT

We simulate the growth of galaxies and their central supermassive black holes by implementing a suite of semi-analytic models on the output of the Millennium Run, a very large simulation of the concordance Λ cold dark matter cosmogony. Our procedures follow the detailed assembly history of each object and are able to track the evolution of all galaxies more massive than the Small Magellanic Cloud throughout a volume comparable to that of large modern redshift surveys. In this first paper we supplement previous treatments of the growth and activity of central black holes with a new model for 'radio' feedback from those active galactic nuclei that lie at the centre of a quasi-static X-ray-emitting atmosphere in a galaxy group or cluster. We show that for energetically and observationally plausible parameters such a model can simultaneously explain: (i) the low observed mass drop-out rate in cooling flows; (ii) the exponential cut-off at the bright end of the galaxy luminosity function; and (iii) the fact that the most massive galaxies tend to be bulge-dominated systems in clusters and to contain systematically older stars than lower mass galaxies. This success occurs because static hot atmospheres form only in the most massive structures, and radio feedback (in contrast, for example, to supernova or starburst feedback) can suppress further cooling and star formation without itself requiring star formation. We discuss possible physical models that might explain the accretion rate scalings required for our phenomenological 'radio mode' model to be successful.

Key words: black hole physics – galaxies: active – cooling flows – galaxies: evolution – galaxies: formation – cosmology: theory.

1 INTRODUCTION

The remarkable agreement between recent measurements of cosmic structure over a wide range of length- and time-scales has established a standard paradigm for structure formation, the Λ cold dark matter (ΛCDM) cosmogony. This model can simultaneously match the microwave background fluctuations seen at $z \sim 1000$ (e.g. Spergel et al. 2003), the power spectrum of the low-redshift galaxy distribution (e.g. Percival et al. 2002; Tegmark et al. 2004), the non-linear mass distribution at low redshift as characterized by cosmic shear (e.g. Van Waerbeke et al. 2002) and the structure seen in the $z = 3$ Lyα forest (e.g. Mandelbaum et al. 2003). It also reproduces the present acceleration of the cosmic expansion inferred from supernova observations (Riess et al. 1998; Perlmutter et al. 1999), and it is consistent with the mass budget inferred for the present Universe from the dynamics of large-scale structure (Peacock et al. 2001), the baryon fraction in rich clusters (White et al. 1993) and the theory of big bang nucleosynthesis (Olive, Steigman & Walker 2000). A working model for the growth of all structure thus appears well established.

In this cosmogony, galaxies form when gas condenses at the centres of a hierarchically merging population of dark haloes, as originally proposed by White & Rees (1978). Attempts to understand this process in detail have consistently run into problems stemming from a mismatch in shape between the predicted distribution of dark halo masses and the observed distribution of galaxy luminosities. Most stars are in galaxies of Milky Way brightness; the galaxy abundance declines exponentially at brighter luminosities and increases sufficiently slowly at fainter luminosities that relatively few stars are

⋆E-mail: darren@astro.berkeley.edu

in dwarfs. In contrast, the theory predicts a much broader halo mass distribution – a constant mass-to-light ratio would produce more high- and low-luminosity galaxies than are observed while underpredicting the number of galaxies like the Milky Way. Attempts to solve these problems initially invoked cooling inefficiencies to reduce gas condensation in massive systems, and supernova feedback to reduce star formation efficiency in low-mass systems (White & Rees 1978; White & Frenk 1991). Formation of dwarfs may also be suppressed by photoionization heating (Efstathiou 1992). As Thoul & Weinberg (1995) emphasized, cooling effects alone are too weak to produce the bright-end cut-off of the luminosity function, and recent attempts to fit observed luminosity functions have been forced to include additional feedback processes in massive systems (e.g. Benson et al. 2003). In this paper we argue that radio sources may provide the required feedback while at the same time providing a solution to two other long-standing puzzles.

An important unanswered question is why the gas at the centre of most galaxy clusters is apparently not condensing and turning into stars when the observed X-ray emission implies a cooling time which is much shorter than the age of the system. This cooling flow puzzle was noted as soon as the first X-ray maps of clusters became available (Cowie & Binney 1977; Fabian & Nulsen 1977), and it was made more acute when X-ray spectroscopy demonstrated that very little gas is cooling through temperatures even a factor of 3 below that of the bulk of the gas (Peterson et al. 2001; Tamura et al. 2001). A clue to the solution may come from the observation (Burns, Gregory & Holman 1981) that every cluster with a strong cooling flow also contains a massive and active central radio galaxy. Tabor & Binney (1993) suggested that radio galaxies might regulate cooling flows, and this idea has gained considerable recent support from X-ray maps which show direct evidence for an interaction between radio lobes and the intracluster gas (McNamara et al. 2000, 2005; Fabian et al. 2003). A number of authors have suggested ways in which the radio source might replace the thermal energy lost to X-ray emission (Binney & Tabor 1995; Churazov et al. 2002; Brüggen & Kaiser 2002; Ruszkowski & Begelman 2002; Kaiser & Binney 2003; Omma et al. 2004). We do not focus on this aspect of the problem here, but rather demonstrate that if the scaling properties of radio source feedback are set so that they can plausibly solve the cooling flow problem, they induce a cut-off at the bright end of the galaxy luminosity function which agrees well with observation.

Another puzzling aspect of the galaxy population is the fact that the most massive galaxies, typically ellipticals in clusters, are made of the oldest stars and so finished their star formation earlier than lower mass galaxies (Bender & Saglia 1999). Confirming evidence for this comes from look-back studies which show that both star +formation and active galactic nucleus (AGN) activity take place more vigorously and in higher mass objects at redshifts of 1 to 2 than in the present Universe (e.g. Shaver et al. 1996; Madau et al. 1996). Cowie et al. (1996) termed this phenomenon 'down-sizing', and prima facie it conflicts with hierarchical growth of structure in a ΛCDM cosmogony where massive dark haloes assemble at lower redshift than lower mass haloes (e.g. Lacey & Cole 1993). This puzzle is related to the previous two; the late-forming high-mass haloes in ΛCDM correspond to groups and clusters of galaxies, and simple theories predict that at late times their central galaxies should grow to masses larger than those observed through accretion from cooling flows. In the model that we present below, radio galaxies prevent significant accretion, thus limiting the mass of the central galaxies and preventing them from forming stars at late times when their mass and morphology can still change through mergers. The result is a galaxy luminosity function with a sharper high-mass cut-off in which the most massive systems are red, dead and elliptical.

To make quantitative predictions for the galaxy population in a ΛCDM universe, it is necessary to carry out simulations. Present numerical capabilities allow reliable simulation of the coupled nonlinear evolution of dark matter and diffuse gas, at least on the scales that determine the global properties of galaxies. Once gas cools and condenses into halo cores, however, both its structure and the rates at which it turns into stars and feeds black holes are determined by small-scale 'interstellar medium' processes which are not resolved. These are usually treated through semi-analytic recipes, parametrized formulae which encapsulate 'subgrid' physics in terms of star formation thresholds, Schmidt 'laws' for star formation, Bondi models for black hole feeding, etc. The form and the parameters of these recipes are chosen to reproduce the observed systematics of star formation and AGN activity in galaxies (e.g. Kennicutt 1998). With a well-constructed scheme it is possible to produce stable and numerically converged simulations which mimic real star-forming galaxies remarkably well (Springel & Hernquist 2003a). In strongly star-forming galaxies, massive stars and supernovae produce winds which redistribute energy, mass and heavy elements over large regions (Heckman, Armus & Miley 1990; Martin 1999). Even stronger feedback is possible, in principle, from AGN (Begelman, de Kool & Sikora 1991). In both cases the determining processes occur on very small scales and so have to be included in simulations through parametrized semi-analytic models. Unfortunately, the properties of simulated galaxies turn out to depend strongly on how these unresolved star formation and feedback processes are treated.

Since the diffuse gas distribution and its cooling on to galaxies are strongly affected by the description adopted for the subgrid physics, every modification of a semi-analytic model (or of its parameters) requires a simulation to be repeated. This makes parameter studies or tests of, say, the effects of different AGN feedback models into a very expensive computing exercise. A cost-effective alternative is to represent the behaviour of the diffuse gas also by a semi-analytic recipe. Since the dark matter couples to the baryons only through gravity, its distribution on scales of galaxy haloes and above is only weakly affected by the details of galaxy formation. Its evolution can therefore be simulated once, and the evolution of the baryonic component can be included in post-processing by applying semi-analytic models to the stored histories of all dark matter objects. Since the second step is computationally cheap, available resources can be used to carry out the best possible dark matter simulation, and then many parameter studies or tests of alternative models can be carried out in post-processing. This simulation approach was first introduced by Kauffmann et al. (1999), and it is the approach that we apply here to the Millennium Run, the largest calculation to date of the evolution of structure in the concordance ΛCDM cosmogony (Springel et al. 2005b).

This paper is organized as follows. In Section 2 we describe the Millennium Run and the post-processing that we carried out to construct merging history trees for all the dark haloes within it. Section 3 presents the model for the formation and evolution of galaxies and their central supermassive black holes that we implement on these merging trees. Section 4 describes the main results of our modelling, concentrating on the influence of 'radio mode' feedback on the properties of the massive galaxy population. In Section 5 we discuss physical models for black hole accretion which may explain the phenomenology required for our model to be successful. Finally, Section 6 summarizes our conclusions and suggests some possible directions for future investigation.

2 THE DARK MATTER SKELETON: THE MILLENNIUM RUN

Our model for the formation and evolution of galaxies and their central supermassive black holes is implemented on top of the Millennium Run, a very large dark matter simulation of the concordance ΛCDM cosmology with $2160^3 \simeq 1.0078 \times 10^{10}$ particles in a periodic box of $500\,h^{-1}$ Mpc on a side. A full description is given in Springel et al. (2005b); here we summarize the main simulation characteristics and the definition and construction of the dark matter merging history trees that we use in our galaxy formation modelling. The dark matter distribution is illustrated in the top panel of Fig. 1 for a $330 \times 280 \times 15\,h^{-1}$ Mpc slice cut from the full volume. The projection is colour-coded by density and local velocity dispersion, and illustrates the richness of dark matter structure for comparison with structure in the light distribution to which we will come later. Dark matter plots on a wider range of scales may be found in Springel et al. (2005b).

2.1 Simulation characteristics

We adopt cosmological parameter values consistent with a combined analysis of the 2-degree Field Galaxy Redshift Survey (2dFGRS) (Colless et al. 2001) and first-year *Wilkinson Microwave Anisotropy Probe* (*WMAP*) data (Spergel et al. 2003; Seljak et al. 2005). They are $\Omega_{\rm m} = \Omega_{\rm dm} + \Omega_{\rm b} = 0.25$, $\Omega_{\rm b} = 0.045$, $h = 0.73$, $\Omega_\Lambda = 0.75$, $n = 1$ and $\sigma_8 = 0.9$. Here $\Omega_{\rm m}$ denotes the total matter density in units of the critical density for closure, $\rho_{\rm crit} = 3H_0^2/(8\pi G)$. Similarly, $\Omega_{\rm b}$ and Ω_Λ denote the densities of baryons and dark energy at the present day. The Hubble constant is given as $H_0 = 100\,h\,{\rm km\,s^{-1}\,Mpc^{-1}}$, while σ_8 is the rms linear mass fluctuation within a sphere of radius $8\,h^{-1}$ Mpc extrapolated to $z = 0$.

The chosen simulation volume is a periodic box of size $500\,h^{-1}$ Mpc, which implies a particle mass of $8.6 \times 10^8\,h^{-1}\,{\rm M}_\odot$. This volume is large enough to include interesting objects of low space density, such as quasars or rich galaxy clusters, the largest of which contain about 3 million simulation particles at $z = 0$. At the same time, the mass resolution is sufficient that haloes that host galaxies as faint as $0.1\,L_\star$ are typically resolved with at least $\sim$100 particles. Note that although discreteness noise significantly affects the merger histories of such low-mass objects, the galaxies that reside in haloes with $\lesssim$100 particles are usually sufficiently far down the luminosity function that any uncertainty in their properties has little impact on our results or conclusions.

The initial conditions at $z = 127$ are created by displacing particles from a homogeneous, 'glass-like' distribution (White 1996) using a Gaussian random field with a ΛCDM linear power spectrum as given by the Boltzmann code CMBFAST (Seljak & Zaldarriaga 1996). The simulation is then evolved to the present epoch using a leapfrog integration scheme with individual and adaptive time-steps, with up to 11 000 time-steps for individual particles.

The simulation itself is carried out with a special version of the GADGET-2 code (Springel, Yoshida & White 2001b; Springel 2005) optimized for very low memory consumption so that it would fit into the nearly 1 TB of physically distributed memory available on the parallel IBM[1] p690 computer[2] used for the calculation. The computational algorithm uses the 'TreePM' method (Xu 1995; Bode, Ostriker & Xu 2000; Bagla 2002) to evaluate gravitational forces, combining a hierarchical multipole expansion, or 'tree' algorithm (Barnes & Hut 1986), and a classical, Fourier transform particle–mesh method (Hockney & Eastwood 1981). An explicit force-split in Fourier space produces a very nearly isotropic force law with negligible force errors at the force matching scale. The short-range gravitational force law is softened on comoving scale $5\,h^{-1}$ kpc (Plummer-equivalent) which may be taken as the spatial resolution limit of the calculation, thus achieving a dynamic range of 10^5 in 3D. The calculation, performed in parallel on 512 processors, requires slightly less than 350 000 processor hours of CPU time, or 28 days of wall-clock time.

[1] IBM Corporation, White Plains, USA.

[2] This computer is operated by the Computing Centre of the Max Planck Society in Garching, Germany.

2.2 Haloes, substructure, and merger tree construction

Our primary application of the Millennium Run in this paper uses finely resolved hierarchical merging trees which encode the full formation history of tens of millions of haloes and the subhaloes that survive within them. These merging history trees are the backbone of the model that we implement in post-processing in order to simulate the wide range of baryonic processes that are important during the formation and evolution of galaxies and their central supermassive black holes.

We store the full particle data between $z = 20$ and 0 at 60 output times spaced in expansion factor according to the formula

$$\log(1 + z_n) = n(n + 35)/4200. \quad (1)$$

This spacing is 'locally' logarithmic but becomes smoothly finer at lower redshift, with a temporal resolution by redshift zero of approximately 300 Myr. Additional outputs are added at $z = 30$, 50, 80 and 127 to produce a total of 64 snapshots in all. We note that each snapshot has a total size in excess of 300 GB, giving a raw data volume of nearly 20 TB.

Together with each particle coordinate dump, the simulation code directly produces a friends-of-friends (FOF) group catalogue on the fly and in parallel. FOF groups are defined as equivalence classes where any pair of two particles is placed into the same group if their mutual separation is less than 0.2 of the mean particle separation (Davis et al. 1985). This criterion combines particles into groups with a mean overdensity of about 200, corresponding approximately to that expected for a virialized group. The group catalogue saved to disk for each output only keeps groups with at least 20 particles.

High-resolution simulations like the present one exhibit a rich substructure of gravitationally bound dark matter subhaloes orbiting within larger virialized haloes (e.g. Ghigna et al. 1998). The FOF group-finder built into our simulation code identifies the haloes but not their subhaloes. Since we wish to follow the fate of infalling galaxies and their haloes, and these are typically identifiable for a substantial time as a dark matter subhalo within a FOF halo, we apply in post-processing an improved and extended version of the SUBFIND algorithm of Springel et al. (2001a). This computes an adaptively smoothed dark matter density field within each halo using a kernel-interpolation technique, and then exploits the topological connectivity of excursion sets above a density threshold to identify substructure candidates. Each substructure candidate is subjected to a gravitational unbinding procedure. If the remaining bound part has more than 20 particles, the subhalo is kept for further analysis and some of its basic physical properties are determined (angular momentum, maximum of its rotation curve, velocity dispersion, etc.). After all subhaloes are identified they are extracted from the FOF halo so that the remaining featureless 'background' halo can also be subjected to the unbinding procedure. This technique, however, neglects the fact that substructures embedded within a halo help to

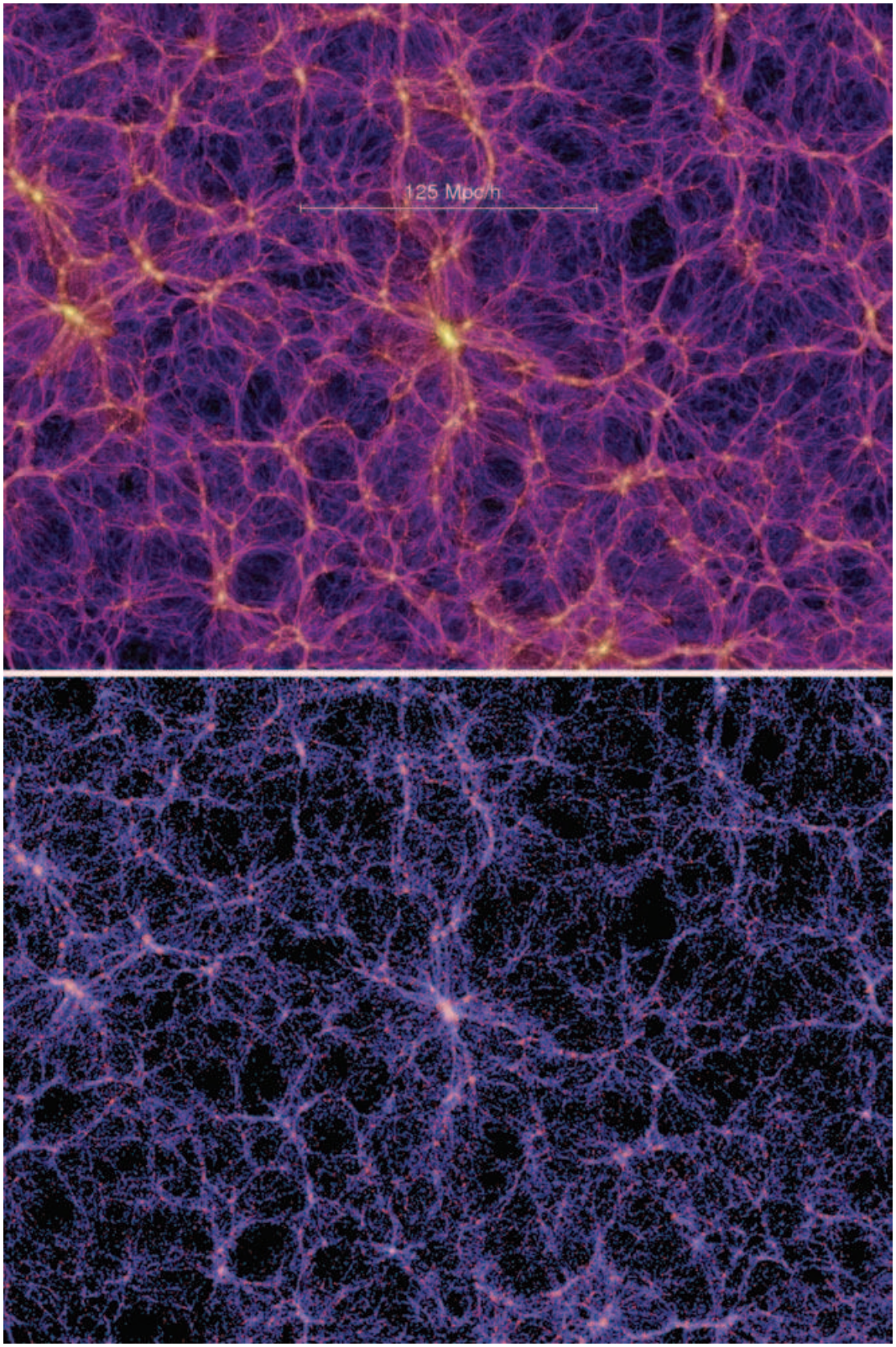

Figure 1. The redshift zero distribution of dark matter (top) and of galaxy light (bottom) for a slice of thickness $15\,h^{-1}$ Mpc, cut from the Millennium Run. For the dark matter distribution, intensity encodes surface density and colour encodes local velocity dispersion. For the light distribution, intensity encodes surface brightness and colour encodes mean $B - V$ colour. The linear scale is shown by the bar in the top panel.

bind its material, and thus removing them can, in principal, unbind some of the FOF halo that may otherwise still be loosely bound. We accept this small effect for technical reasons related to the robustness of our halo definition procedures and the need for unambiguous particle–subhalo assignments in our data structures. We note that the total mass in substructures is typically below 10 per cent and often substantially smaller. Thus any bias in the bound mass of the parent halo due to additional unbinding is always very small.

To compute a virial mass estimate for each FOF halo we use the spherical-overdensity approach, where the centre is determined using the minimum of the gravitational potential within the group and we define the boundary at the radius that encloses a mean overdensity of 200 times the critical value. The virial mass, radius and circular velocity of a halo at redshift z are then related by

$$M_{\rm vir} = \frac{100}{G} H^2(z) R_{\rm vir}^3 = \frac{V_{\rm vir}^3}{10GH(z)}, \quad (2)$$

where $H(z)$ is the Hubble constant at redshift z.

At $z = 0$ our procedures identify 17.7×10^6 FOF groups, down from a maximum of 19.8×10^6 at $z = 1.4$ when groups were more abundant but of lower mass on average. The $z = 0$ groups contain a total of 18.2×10^6 subhaloes, with the largest FOF group containing 2328 of them. (Note that with our definitions, all FOF groups contain at least one subhalo, the *main* subhalo which is left over after removal of any substructure and the unbound component. We require this main subhalo to contain at least 20 particles.)

Having found all haloes and subhaloes at all output snapshots, we then characterize structural evolution by building merging trees that describe in detail how haloes grow as the universe evolves. Because structures merge hierarchically in CDM universes, there can be several progenitors for any given halo, but in general there is only one descendant. (Typically the cores of virialized dark matter structures do not split into two or more objects.) We therefore construct merger trees based on defining a unique descendant for each halo and subhalo. This is, in fact, sufficient to define the entire merger tree, since the progenitor information then follows implicitly. Further details can be found in Springel et al. (2005b).

We store the resulting merging histories tree by tree. Since each tree contains the full formation history of some $z = 0$ halo, the physical model for galaxy formation can be computed sequentially tree by tree. It is thus unnecessary to load all the history information on the simulation into computer memory at the same time. Actually, this would be currently impossible, since the combined trees contain a total of around 800 million subhaloes for each of which a number of attributes are stored. Note that although evolving the galaxy population sequentially in this way allows us to consider, in principle, all interactions between galaxies that end up in the same present-day FOF halo, it does not allow us to model longer range interactions which might take place between galaxies that end up in different FOF haloes at $z = 0$.

3 BUILDING GALAXIES: THE SEMI-ANALYTIC MODEL

3.1 Overview

In the following subsections we describe the baryonic physics of one particular model for the formation and evolution of galaxies and of their central supermassive black holes. A major advantage of our simulation strategy is that the effects of parameter variations within this model (or indeed alternative assumptions for some of the processes) can be explored at relatively little computational expense since the model operates on the stored data base of merger trees; the simulation itself and the earlier stages of post-processing do not need to be repeated. We have, in fact, explored a wide model and parameter space to identify our current best model. We summarize the main parameters of this model, their 'best' values and their plausible ranges in Table 1. These choices produce a galaxy population which matches quite closely many observed quantities. In this paper we discuss the field galaxy luminosity–colour distribution; the mean stellar mass–stellar age relation; the Tully–Fisher relation, cold gas fractions and gas-phase metallicities of Sb/c spirals; the colour–magnitude relation of ellipticals; the bulge mass–black hole mass relation; and the volume-averaged cosmic star formation and black hole accretion histories. In Springel et al. (2005b) we also presented results for galaxy correlations as a function of absolute magnitude and colour, for the baryonic 'wiggles' in the large-scale power spectrum of galaxies, and for the abundance, origin and fate of high-redshift supermassive black holes which might correspond to the $z \sim 6$ quasars discovered by the Sloan Digital Sky Survey (SDSS) (Fan et al. 2001).

In our model we aim to motivate each aspect of the physics of galaxy formation using the best available observations and simulations. Our parameters have been chosen to reproduce local galaxy properties and are stable in the sense that none of our results is critically dependent on any single parameter choice; plausible changes in one parameter or recipe can usually be accommodated through adjustment of the remaining parameters within their own plausible range. The particular model that we present is thus not unique. Importantly, our model for radio galaxy heating in cooling flows,

Table 1. A summary of our main model parameters and their best values and plausible ranges, as described in the text. Once set, these values are kept fixed for all results presented in this paper, in particular for models in which AGN feedback is switched off.

Parameter	Description	Best value	Plausible range
$f_{\rm b}$	Cosmic baryon fraction (Section 3.3)	0.17	fixed
z_0, $z_{\rm r}$	Redshift of reionization (Section 3.3)	8, 7	fixed
$f_{\rm BH}$	Merger cold gas BH accretion fraction (Section 3.4.1)	0.03	002–004
$\kappa_{\rm AGN}$	Quiescent hot gas BH accretion rate ($\rm M_\odot\ yr^{-1}$) (Section 3.4.2)	6×10^{-6}	$(4–8) \times 10^{-6}$
$\alpha_{\rm SF}$	Star formation efficiency (Section 3.5)	0.07	005–015
$\epsilon_{\rm disc}$	SN feedback disc reheating efficiency (Section 3.6)	3.5	1–5
$\epsilon_{\rm halo}$	SN feedback halo ejection efficiency (Section 3.6)	0.35	01–05
$\gamma_{\rm ej}$	Ejected gas reincorporation efficiency (Section 3.6)	0.5	01–10
$T_{\rm merger}$	Major merger mass ratio threshold (Section 3.7)	0.3	02–04
R	Instantaneous recycled fraction of SF to the cold disc (Section 3.9)	0.3	02–04
Y	Yield of metals produced per unit SF (Section 3.9)	0.03	002–004

which is the main focus of this paper, is only weakly influenced by the remaining galaxy formation and black hole growth physics. This is because our radio mode feedback is active only in massive objects and at late times, and it has no effect during the principal growth phase of most galaxies and AGN.

The distribution of galaxy light in our 'best' model is shown in the bottom panel of Fig. 1 for comparison with the mass distribution in the top panel. For both the volume is a projected $330 \times 280 \times 15\,h^{-1}$ Mpc slice cut from the full $0.125\,h^{-3}\,\mathrm{Gpc}^3$ simulation box. The plot of surface brightness is colour-coded by the luminosity-weighted mean $B - V$ colour of the galaxies. On large scales light clearly follows mass, but non-trivial biases become evident on smaller scales, especially in 'void' regions. The redder colour of galaxies in high-density regions is also very clear.

3.2 Gas infall and cooling

We follow the standard paradigm set out by White & Frenk (1991) as adapted for implementation on high-resolution N-body simulations by Springel et al. (2001a) and De Lucia, Kauffmann & White (2004). This assumes that as each dark matter halo collapses, its own 'fair share' of cosmic baryons collapse with it (but see Section 3.3 below). Thus in our model the mass fraction in baryons associated with every halo is taken to be $f_\mathrm{b} = 17$ per cent, consistent with the first-year *WMAP* result (Spergel et al. 2003). Initially these baryons are in the form of diffuse gas with primordial composition, but later they include gas in several phases as well as stars and heavy elements. The fate of the infalling gas depends on redshift and on the depth of the halo potential (Silk 1977; Rees & Ostriker 1977; Binney 1977; White & Rees 1978). At late times and in massive systems the gas shocks to the virial temperature and is added to a quasi-static hot atmosphere that extends approximately to the virial radius of the dark halo. Gas from the central region of this atmosphere may accrete on to a central object through a cooling flow. At early times and in lower mass systems the infalling gas still shocks to the virial temperature, but its post-shock cooling time is sufficiently short that a quasi-static atmosphere cannot form. Rather the shock occurs at much smaller radius and the shocked gas cools rapidly and settles on to a central object, which we assume to be a cold gas disc. This may in turn be subject to gravitational instability, leading to episodes of star formation.

More specifically, the cooling time of a gas is conventionally taken as the ratio of its specific thermal energy to the cooling rate per unit volume,

$$t_\mathrm{cool} = \frac{3}{2}\frac{\bar{\mu} m_\mathrm{p} k T}{\rho_\mathrm{g}(r)\Lambda(T, Z)}. \tag{3}$$

Here $\bar{\mu} m_\mathrm{p}$ is the mean particle mass, k is the Boltzmann constant, $\rho_\mathrm{g}(r)$ is the hot gas density, and $\Lambda(T, Z)$ is the cooling function. The latter depends both on the metallicity Z and on the temperature of the gas. In our models we assume the post-shock temperature of the infalling gas to be the virial temperature of the halo, $T = 35.9\,(V_\mathrm{vir}/\mathrm{km\,s^{-1}})^2$ K. When needed, we assume that the hot gas within a static atmosphere has a simple 'isothermal' distribution,

$$\rho_\mathrm{g}(r) = \frac{m_\mathrm{hot}}{4\pi R_\mathrm{vir} r^2}, \tag{4}$$

where m_hot is the total hot gas mass associated with the halo and is assumed to extend to its virial radius R_vir.

To estimate an instantaneous cooling rate on to the central object of a halo, given its current hot gas content, we define the cooling radius, r_cool, as the radius at which the local cooling time (assuming the structure of equation 4) is equal to a suitably defined age for the halo. White & Frenk (1991) took this age to be the Hubble time t_H, while Cole et al. (1994, 2000) used the time-scale over which the main progenitor last doubled its mass. Somerville & Primack (1999) argued that the time interval since the last major merger is more appropriate since such mergers redistribute hot gas within the halo. Here we follow Springel et al. (2001a) and De Lucia et al. (2004) and define the cooling radius as the point where the local cooling time is equal to the halo dynamical time, $R_\mathrm{vir}/V_\mathrm{vir} = 0.1H(z)^{-1}$. This is an order of magnitude smaller than t_H and so results in substantially smaller cooling radii and cooling rates (typically by a factor of 3) than the assumption of White & Frenk (1991). Our choice is justified by the tests of Yoshida et al. (2002), who verified explicitly that it results in good object-by-object agreement between the amount of gas predicted to condense in galaxy mass haloes and the amount that actually condensed in their high-resolution smoothed particle hydrodynamics (SPH) simulations of the formation of a galaxy cluster and its environment. These tests assumed primordial abundances in the cooling function. When we implement a chemical enrichment model consistent with the observed element abundances in intracluster gas (see Section 3.9), cooling rates in galaxy-mass haloes are substantially enhanced and we find (as did De Lucia et al. 2004) that a smaller coefficient than used in the original White & Frenk cooling model is required to avoid excessive condensation of gas.

Using the above definition, a cooling rate can now be determined through a simple continuity equation,

$$\dot{m}_\mathrm{cool} = 4\pi\rho_\mathrm{g}(r_\mathrm{cool}) r_\mathrm{cool}^2 \dot{r}_\mathrm{cool}. \tag{5}$$

Despite its simplicity, this equation is a good approximation to the rate at which gas is deposited at the centre in the Bertschinger (1989) similarity solution for a cooling flow. Putting it all together, we take the cooling rate within a halo containing a hot gas atmosphere to be

$$\dot{m}_\mathrm{cool} = 0.5 m_\mathrm{hot} \frac{r_\mathrm{cool} V_\mathrm{vir}}{R_\mathrm{vir}^2}. \tag{6}$$

We assume this equation to be valid when $r_\mathrm{cool} < R_\mathrm{vir}$. This is the criterion that White & Frenk (1991) proposed to separate the *static hot halo regime* from the *rapid cooling regime* (see below).

In low-mass haloes or at high redshifts the formal cooling radius lies outside the virial radius $r_\mathrm{cool} > R_\mathrm{vir}$. The post-shock gas then cools in less than one sound crossing time and cannot maintain the pressure needed to support an accretion shock at large radius. The high-resolution spherical infall simulations of Forcada-Miró & White (1997) show that in this situation the accretion shock moves inwards, the post-shock temperature *increases* and the mass stored in the post-shock hot atmosphere *decreases*, because the post-shock gas rapidly cools on to the central object. In effect, all infalling material is accreted immediately on to the central disc. In this 'rapid cooling regime' we therefore set the cooling rate on to the central object to be equal to the rate at which new diffuse gas is added to the halo.

3.2.1 Rapid cooling or cold accretion?

Although much simplified, the above model was shown by Yoshida et al. (2002) and Helly et al. (2003) to give reasonably accurate, object-by-object predictions for the cooling and accumulation of gas within the galaxies that formed in their N-body+SPH simulations. These neglected star formation and feedback effects in order to test the cooling model alone. They also assumed primordial abundances in the cooling function. In Fig. 2 we show the ratio $r_\mathrm{cool}/R_\mathrm{vir}$ as a

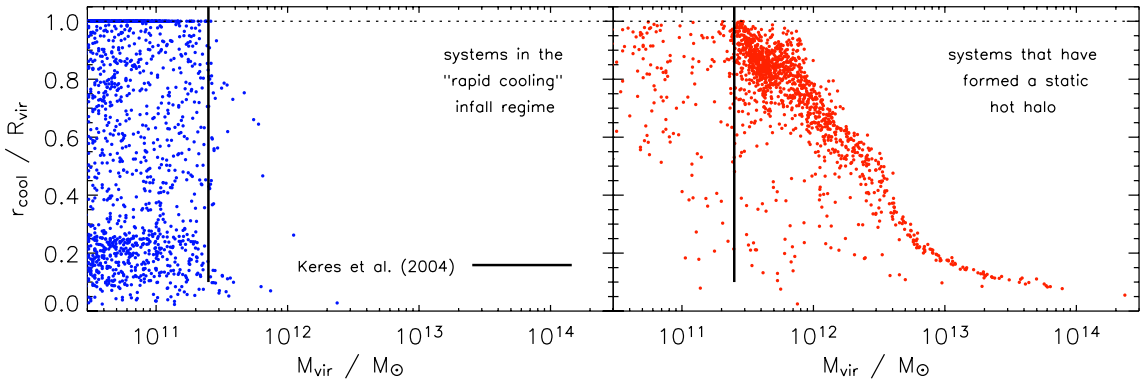

Figure 2. The ratio of the cooling radius to virial radius for a random selection of virialized systems at $z = 0$ plotted against their dark matter virial mass. Systems identified to be in the 'rapid cooling' regime are shown in the left-hand panel, while those that have formed a static hot halo are shown on the right (Section 3.2). A sharp transition between the two regimes is seen close to that found by Keres et al. (2004), marked by the solid vertical line.

function of virial mass for haloes in the 'rapid cooling regime' (left-hand panel) and in the 'static hot halo regime' (right-hand panel) at $z = 0$ for our 'best' galaxy formation model. The two regimes are distinguished by the dominant gas component in each halo: when the mass of hot halo gas exceeds that of cold disc gas, we say that the galaxy has formed a static halo, otherwise the system is taken to be in the rapid cooling phase. Many haloes classified as 'rapidly cooling' by this criterion have $r_{\rm cool} < R_{\rm vir}$, which would apparently indicate a static hot halo. This is misleading, however, as systems where cooling is rapid deposit infalling gas on to the central galactic disc on a short time-scale until they have a low-mass residual halo which satisfies $r_{\rm cool} \sim R_{\rm vir}$. This then persists until the next infall event. Averaging over several cycles of this behaviour, one finds that the bulk of the infalling gas cools rapidly. This is why we choose to classify systems by their dominant gas component. Note also that a massive hot halo forms immediately once cooling becomes inefficient, just as in the 1D infall simulations of Forcada-Miró & White (1997) and Birnboim & Dekel (2003). Our classification is thus quite robust.

The transition between the 'rapid cooling' and 'static halo' regimes is remarkably well defined. At $z = 0$ it occurs at a halo virial mass of 2–3 $\times$ 10^{11} $\mathrm{M}_{\odot}$, and is approximately independent of redshift out to at least $z = 6$. This is close to the transition mass found for the same cosmology by Birnboim & Dekel (2003) using spherically symmetric simulations, and by Keres et al. (2004) using fully 3D simulations. This is, in fact, a coincidence since both sets of authors assume cooling functions with no metals, whereas our 'best' model includes heavy elements which substantially enhance cooling at the temperatures relevant for the transition. Enrichment in the 'rapid cooling' regime arises from the mixing of infalling primordial gas with supernova-enriched gas ejected in a galactic wind (see Section 3.6). If we modify our model to assume a zero-metal cooling function, we find the transition mass to shift downwards by about a factor of 2–3, resulting in a lower cooling rate in comparison with the work of Birnboim & Dekel (2003) and Keres et al. (2004).

The reason for the more efficient (we would argue overly efficient) cooling appears to be different in these two studies. The spherical infall simulations of Forcada-Miró & White (1997) showed good agreement with a transition at the point predicted using the original cooling radius definition of White & Frenk (1991) rather than our revised definition which was checked explicitly against SPH simulations by Yoshida et al. (2002). Spherical models thus predict more efficient cooling than occurs in typical 3D situations. This explains, perhaps, why Birnboim & Dekel (2003) find a higher transition mass than we would predict for zero metallicity. Yoshida et al. (2002) also showed that the density estimation in the implementation of SPH used by Keres et al. (2004) leads to overcooling of gas in galaxy-mass objects as compared with their own entropy- and energy-conserving SPH scheme; the effect is large enough to explain why Keres et al. (2004) find a higher transition mass than we find for their assumed cooling function.

Both Birnboim & Dekel and Keres et al. refer to the 'rapid cooling' regime as 'cold infall'. This is, in fact, a misnomer. In this mode the accretion shock occurs closer to the central object, and so deeper in its potential well than when there is a static hot halo. As a result, the pre-shock velocity of infalling gas is greater, resulting in a *larger* post-shock temperature. The two modes do not differ greatly in the temperature to which infalling gas is shocked, but rather in how long (compared with the system crossing time) the gas spends at the post-shock temperature before its infall energy is lost to radiation. Finally, we note that the existence and importance of these two modes were the major insights of the original work of Silk (1977), Binney (1977) and Rees & Ostriker (1977) and have been built into all modern theories for galaxy formation. A detailed discussion can be found, for example, in White & Frenk (1991).

3.3 Reionization

Accretion and cooling in low-mass haloes is required to be inefficient to explain why dwarf galaxies contain a relatively small fraction of all condensed baryons (White & Rees 1978). This inefficiency may in part result from photoionization heating of the intergalactic medium (IGM) which suppresses the concentration of baryons in shallow potentials (Efstathiou 1992). More recent models identify the possible low-redshift signature of such heating in the faint end of the galaxy luminosity function, most notably in the abundance of the dwarf satellite galaxies in the Local Group (e.g. Tully et al. 2002; Benson et al. 2002).

Gnedin (2000) showed that the effect of photoionization heating on the gas content of a halo of mass $M_{\rm vir}$ can be modelled by defining a characteristic mass scale, the so called *filtering mass*, $M_{\rm F}$, below which the gas fraction $f_{\rm b}$ is reduced relative to the universal value:

$$f_{\rm b}^{\rm halo}(z, M_{\rm vir}) = \frac{f_{\rm b}^{\rm cosmic}}{[1 + 0.26 M_{\rm F}(z)/M_{\rm vir}]^3}. \quad (7)$$

The filtering mass is a function of redshift and changes most significantly around the epoch of reionization. It was estimated by Gnedin using high-resolution softened Lagrangian hydrodynamics (SLH–P^3M) simulations (but see Hoeft et al. 2005). Kravtsov, Gnedin & Klypin (2004) provided an analytic model for these results which distinguishes three 'phases' in the evolution of the IGM: $z > z_0$, the epoch before the first H II regions overlap; $z_0 < z < z_r$, the epoch when multiple H II regions overlap; and $z < z_r$, the epoch when the medium is almost fully reionized. They find that choosing $z_0 = 8$ and $z_r = 7$ provides the best fit to the numerically determined filtering mass. We adopt these parameters and keep them fixed throughout our paper. This choice results in a filtering mass of $4 \times 10^9\,\mathrm{M}_\odot$ at $z = z_r$, and $3 \times 10^{10}\,\mathrm{M}_\odot$ by the present day. See appendix B of Kravtsov et al. (2004) for a full derivation and description of the analytic model.

3.4 Black hole growth, AGN outflows and cooling suppression

There is a growing body of evidence that AGN are a critical piece in the galaxy formation puzzle. Our principal interest in this paper is their role in suppressing cooling flows, thereby modifying the luminosities, colours, stellar masses and clustering of the galaxies that populate the bright end of the galaxy luminosity function. To treat this problem, we first need a physical model for the growth of black holes within our galaxies.

3.4.1 The 'quasar mode'

In our model (which is based closely on that of Kauffmann & Haehnelt 2000), supermassive black holes grow during galaxy mergers both by merging with each other and by accretion of cold disc gas. For simplicity, black hole coalescence is modelled as a direct sum of the progenitor masses and thus ignores gravitational wave losses (including such losses is found to have little effect on the properties of the final galaxy population). We assume that the gas mass accreted during a merger is proportional to the total cold gas mass present, but with an efficiency which is lower for smaller mass systems and for unequal mergers. Specifically,

$$\Delta m_{\mathrm{BH,Q}} = \frac{f'_{\mathrm{BH}} m_{\mathrm{cold}}}{1 + (280\ \mathrm{km\ s^{-1}}/V_{\mathrm{vir}})^2}, \tag{8}$$

where we have changed the original parametrization by taking

$$f'_{\mathrm{BH}} = f_{\mathrm{BH}}(m_{\mathrm{sat}}/m_{\mathrm{central}}). \tag{9}$$

Here $f_{\mathrm{BH}} \approx 0.03$ is a constant and is chosen to reproduce the observed local m_{BH}–m_{bulge} relation (Magorrian et al. 1998; Marconi & Hunt 2003; Häring & Rix 2004). In contrast to Kauffmann & Haehnelt (2000), we allow black hole accretion during both major *and* minor mergers, although the efficiency in the latter is lower because of the $m_{\mathrm{sat}}/m_{\mathrm{central}}$ term. Thus any merger-induced perturbation to the gas disc (which might come from a bar instability or a merger-induced starburst – see Section 3.7) can drive gas on to the central black hole. In this way, minor merger growth of the black hole parallels minor merger growth of the bulge. The fractional contribution of minor mergers to both is typically quite small, so that accretion driven by major mergers is the dominant mode of black hole growth in our model. We refer to this as the 'quasar mode'. [Note that a more schematic treatment of black hole growth would suffice for the purposes of this paper, but in Springel et al. (2005b) and in future work we wish to examine the build-up of the black hole population within galaxies in considerably more detail.]

There is substantial evidence for strong hydrodynamic and radiative feedback from optical/ultraviolet and X-ray AGN (Arav et al. 2001; de Kool et al. 2001; Reeves, O'Brien & Ward 2003; Crenshaw, Kraemer & George 2003). We have not yet explicitly incorporated such feedback in our modelling, and it may well turn out to be important [see, for example, the recent simulations of Di Matteo, Springel & Hernquist (2005), Springel et al. (2005a) and Hopkins et al. (2005)]. We assume 'quasar mode' accretion to be closely associated with starbursts, so this feedback channel may be partially represented in our models by an enhanced effective feedback efficiency associated with star formation and supernovae (see Section 3.6).

3.4.2 The 'radio mode'

In our model, low-energy 'radio' activity is the result of hot gas accretion on to a central supermassive black hole once a static hot halo has formed around the host galaxy of the black hole. We assume this accretion to be continual and quiescent and to be described by a simple phenomenological model:

$$\dot{m}_{\mathrm{BH,R}} = \kappa_{\mathrm{AGN}} \left(\frac{m_{\mathrm{BH}}}{10^8\,\mathrm{M}_\odot}\right) \left(\frac{f_{\mathrm{hot}}}{0.1}\right) \left(\frac{V_{\mathrm{vir}}}{200\ \mathrm{km\ s^{-1}}}\right)^3, \tag{10}$$

where m_{BH} is the black hole mass, f_{hot} is the fraction of the total halo mass in the form of hot gas, $V_{\mathrm{vir}} \propto T_{\mathrm{vir}}^{1/2}$ is the virial velocity of the halo, and κ_{AGN} is a free parameter with units of $\mathrm{M}_\odot\,\mathrm{yr}^{-1}$ with which we control the efficiency of accretion. We find below that $\kappa_{\mathrm{AGN}} = 6 \times 10^{-6}\,\mathrm{M}_\odot\,\mathrm{yr}^{-1}$ accurately reproduces the turnover at the bright end of the galaxy luminosity function. Note that $f_{\mathrm{hot}} V_{\mathrm{vir}}^3 t_{\mathrm{H}}$ is proportional to the total mass of hot gas, so that our formula is simply the product of the hot gas and black hole masses multiplied by a constant efficiency and divided by the Hubble time t_{H}. In fact, we find f_{hot} to be approximately constant for $V_{\mathrm{vir}} \gtrsim 150\,\mathrm{km\,s^{-1}}$, so the dependence of $\dot{m}_{\mathrm{BH,R}}$ on this quantity has little effect. The accretion rate given by equation (10) is typically orders of magnitude below the Eddington limit. In Section 5 we discuss physical accretion models that may reproduce this phenomenology.

We assume that 'radio mode' feedback injects sufficient energy into the surrounding medium to reduce or even stop the cooling flow described in Section 3.2. We take the mechanical heating generated by the black hole accretion of equation (10) to be

$$L_{\mathrm{BH}} = \eta \dot{m}_{\mathrm{BH}} c^2, \tag{11}$$

where $\eta = 0.1$ is the standard efficiency with which mass is assumed to produce energy near the event horizon, and c is the speed of light. This injection of energy compensates in part for the cooling, giving rise to a modified infall rate (equation 6) of

$$\dot{m}'_{\mathrm{cool}} = \dot{m}_{\mathrm{cool}} - \frac{L_{\mathrm{BH}}}{\frac{1}{2} V_{\mathrm{vir}}^2}. \tag{12}$$

For consistency we never allow $\dot{m}'_{\mathrm{cool}}$ to fall below zero. It is worth noting that $\dot{m}_{\mathrm{cool}} \propto f_{\mathrm{hot}}^{3/2} \Lambda(V_{\mathrm{vir}})^{1/2} V_{\mathrm{vir}}^2 t_{\mathrm{H}}^{-1/2}$ (equation 6) and $\dot{m}_{\mathrm{heat}} \equiv 2L_{\mathrm{BH}}/V_{\mathrm{vir}}^2 \propto m_{\mathrm{BH}} f_{\mathrm{hot}} V_{\mathrm{vir}}$ (equation 12), so that

$$\frac{\dot{m}_{\mathrm{heat}}}{\dot{m}_{\mathrm{cool}}} \propto \frac{m_{\mathrm{BH}} t_{\mathrm{H}}^{1/2}}{f_{\mathrm{hot}}^{1/2} \Lambda(V_{\mathrm{vir}})^{1/2} V_{\mathrm{vir}}}. \tag{13}$$

(These scalings are exact for an Einstein–de Sitter universe; we have omitted weak coefficient variations in other cosmologies.) Thus in our model the effectiveness of radio AGN in suppressing cooling flows is greatest at late times and for large values of black hole mass. This turns out to be the qualitative behaviour needed for the suppression of cooling flows to reproduce successfully the luminosities, colours and clustering of low-redshift bright galaxies.

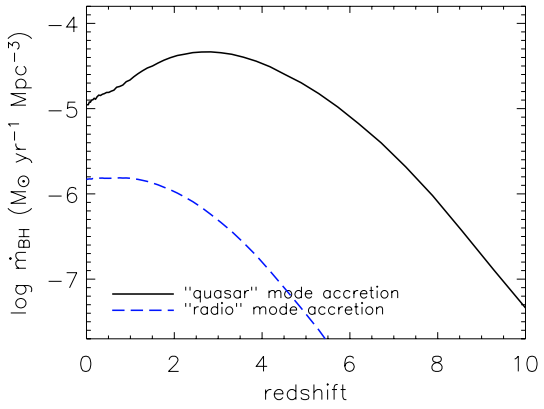

Figure 3. The black hole accretion rate density, $\dot{m}_{\rm BH}$, as a function of redshift for both the 'quasar' and the 'radio' modes discussed in Section 3.4. This figure shows that the growth of black holes is dominated by the 'quasar mode' at high redshift and falls off sharply at $z \lesssim 2$. In contrast, the 'radio mode' becomes important at low redshifts where it suppresses cooling flows, but is not a significant contributor to the overall black hole mass budget.

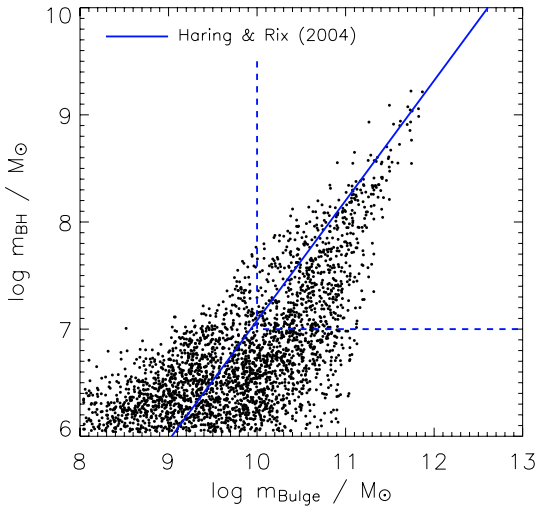

Figure 4. The black hole–bulge mass relation for model galaxies at the present day. The local observational result of Häring & Rix (2004) is given by the solid line, where the dashed box shows the approximate range over which their fit was obtained.

3.4.3 The growth of supermassive black holes

Figure 3 shows the evolution of the mean black hole accretion rate per unit volume averaged over the entire Millennium Simulation box. We separate the accretion into the 'quasar' and 'radio' modes described above (solid and dashed lines respectively). Black hole mass growth in our model is dominated by the merger-driven 'quasar mode', which is most efficient at redshifts of 2–4, dropping by a factor of 5 by redshift zero. This behaviour has similar form to but is weaker than the observed evolution with redshift of the bright quasar population (e.g. Hartwick & Schade 1990). (See also the discussion in Kauffmann & Haehnelt 2000.) In contrast, our 'radio mode' is significant only at late times, as expected from the scaling discussed above, and for the high feedback efficiency assumed in equation (11) it contributes only 5 per cent of the final black hole mass density. We will show, however, that the outflows generated by this accretion can have a major impact on the final galaxy properties. Finally, integrating the accretion rate density over time gives a present-day black hole mass density of $3 \times 10^5\,{\rm M}_\odot\,{\rm Mpc}^{-3}$, consistent with recent observational estimates (Yu & Tremaine 2002; Merloni 2004).

The relationship between black hole mass and bulge mass is plotted in Fig. 4 for the local galaxy population in our 'best' model. In this figure, the solid line shows the best fit to the observations given by Häring & Rix (2004) for a sample of 30 nearby galaxies with well-measured bulge and black hole masses. Their results only probe masses over the range bounded by the dashed lines. Our model galaxies produce a good match to these observations with comparable scatter in the observed range (see their fig. 2).

3.5 Star formation

We use a simple model for star formation similar to those adopted by earlier authors. We assume that all star formation occurs in cold disc gas, either quiescently or in a burst (see Section 3.7). Based on the observational work of Kennicutt (1998), we adopt a threshold surface density for the cold gas below which no stars form, but above which gas starts to collapse and form stars. According to Kauffmann (1996), this critical surface density at a distance R from the galaxy centre may be approximated by

$$\Sigma_{\rm crit}(R) = 120\left(\frac{V_{\rm vir}}{200\ {\rm km\ s^{-1}}}\right)\left(\frac{R}{\rm kpc}\right)^{-1}\ {\rm M_\odot\ pc^{-2}}. \qquad (14)$$

We convert this critical surface density into a critical mass by assuming the cold gas mass to be evenly distributed over the disc. The resulting critical cold gas mass is

$$m_{\rm crit} = 3.8\times10^9\left(\frac{V_{\rm vir}}{200\ {\rm km\ s^{-1}}}\right)\left(\frac{r_{\rm disc}}{10\ {\rm kpc}}\right){\rm M_\odot}, \qquad (15)$$

where we assume the disc scalelength to be $r_{\rm s} = (\lambda/\sqrt{2})R_{\rm vir}$ (Mo, Mao & White 1998), and set the outer disc radius to $r_{\rm disc} = 3r_{\rm s}$, based on the properties of the Milky Way (van den Bergh 2000). Here λ is the spin parameter of the dark halo in which the galaxy resides (Bullock et al. 2001), measured directly from the simulation at each time-step. When the mass of cold gas in a galaxy is greater than this critical value we assume the star formation rate to be

$$\dot{m}_* = \alpha_{\rm SF}(m_{\rm cold} - m_{\rm crit})/t_{\rm dyn,disc}, \qquad (16)$$

where the efficiency $\alpha_{\rm SF}$ is typically set so that 5–15 per cent of the gas is converted into stars in a disc dynamical time $t_{\rm dyn,disc}$, which we define to be $r_{\rm disc}/V_{\rm vir}$. This star formation model produces a global star formation history consistent with the observed star formation density of the universe out to at least $z = 2$, as shown in Fig. 5 (note that this figure also includes star formation through starbursts – see Section 3.7) .

When implemented in our model, equation (16) leads to episodic star formation that self-regulates so as to maintain the critical surface density of equation (14). This naturally reproduces the observed spiral galaxy gas fractions without the need for additional parametrization, as we demonstrate in the top panel of Fig. 6 using model Sb/c galaxies identified as objects with bulge-to-total luminosity $1.5 \leqslant M_{\rm I,bulge} - M_{\rm I,total} \leqslant 2.5$ (bulge formation is described in Section 3.7).

3.6 Supernova feedback

As star formation proceeds, newly formed massive stars rapidly complete their evolution and end their life as supernovae. These

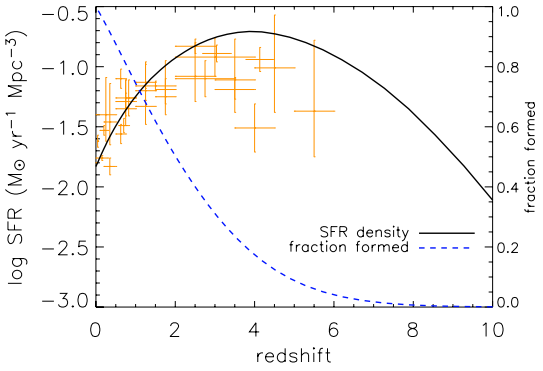

Figure 5. The star formation rate density of the Universe as a function of redshift. The symbols show a compilation of observational results taken from fig. 12 of Springel & Hernquist (2003b). The solid line shows our 'best' model, which predicts that galaxies form much of their mass relatively early. The dashed line (and right axis) indicate the increase in stellar mass with redshift. Approximately 50 per cent of all stars form by $z = 2$.

events inject gas, metals and energy into the surrounding medium, reheating cold disc gas and possibly ejecting gas even from the surrounding halo.

The observations of Martin (1999) suggest modelling the amount of cold gas reheated by supernovae as

$$\Delta m_{\rm reheated} = \epsilon_{\rm disc} \Delta m_*, \tag{17}$$

where Δm_* is the mass of stars formed over some finite time interval and $\epsilon_{\rm disc}$ is a parameter which we fix at $\epsilon_{\rm disc} = 3.5$ based on the observational data. The energy released in this interval can be approximated by

$$\Delta E_{\rm SN} = 0.5 \epsilon_{\rm halo} \Delta m_* V_{\rm SN}^2, \tag{18}$$

where $0.5 V_{\rm SN}^2$ is the mean energy in supernova ejecta per unit mass of stars formed, and $\epsilon_{\rm halo}$ parametrizes the efficiency with which this energy is able to reheat disc gas. Based on a standard initial stellar mass function and standard supernova theory we take $V_{\rm SN} = 630\,{\rm km\,s^{-1}}$. In addition, for our 'best' model we adopt $\epsilon_{\rm halo} = 0.35$. If the reheated gas were added to the hot halo without changing its specific energy, its total thermal energy would change by

$$\Delta E_{\rm hot} = 0.5 \Delta m_{\rm reheated} V_{\rm vir}^2. \tag{19}$$

Thus the *excess* energy in the hot halo after reheating is just $\Delta E_{\rm excess} = \Delta E_{\rm SN} - \Delta E_{\rm hot}$. When $\Delta E_{\rm excess} < 0$ the energy transferred with the reheated gas is insufficient to eject any gas out of the halo and we assume that all hot gas remains associated with the halo. When excess energy is present, i.e. $\Delta E_{\rm excess} > 0$, we assume that some of the hot gas is ejected from the halo into an external 'reservoir'. Specifically, we take

$$\Delta m_{\rm ejected} = \frac{\Delta E_{\rm excess}}{E_{\rm hot}} m_{\rm hot} = \left(\epsilon_{\rm halo} \frac{V_{\rm SN}^2}{V_{\rm vir}^2} - \epsilon_{\rm disc} \right) \Delta m_* \,, \tag{20}$$

where $E_{\rm hot} = 0.5 m_{\rm hot} V_{\rm vir}^2$ is the total thermal energy of the hot gas, and we set $\Delta m_{\rm ejected} = 0$ when this equation gives negative values (implying $\Delta E_{\rm excess} < 0$ as discussed above). This is similar to the traditional semi-analytic feedback recipe, $\Delta m_{\rm ejected} \propto \Delta m_* / V_{\rm vir}^2$, but with a few additions. For small $V_{\rm vir}$ the entire hot halo can be ejected and then $\Delta m_{\rm ejected}$ must saturate at $\Delta m_{\rm reheated}$. Conversely, no hot gas can be ejected from the halo for $V_{\rm vir}^2 > \epsilon_{\rm halo} V_{\rm SN}^2 / \epsilon_{\rm disc}$, i.e. for halo circular velocities exceeding about $200\,{\rm km\,s^{-1}}$ for our favoured parameters.

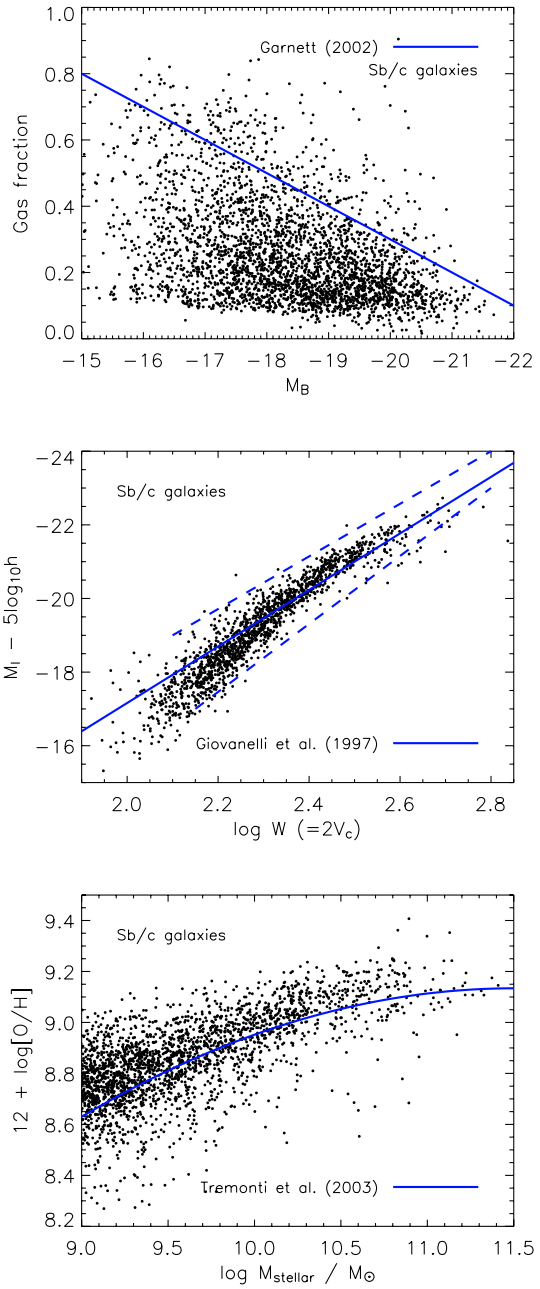

Figure 6. Selected results for Sb/c galaxies (identified by bulge-to-total luminosity, see Section 3.5) for our best model. Top: gas fractions as a function of B-band magnitude. The solid line is a representation of the mean behaviour in the (incomplete) sample of Garnett (2002). Middle: the Tully–Fisher relation, where the disc circular velocity, $V_{\rm c}$, is approximated by $V_{\rm vir}$ for the dark halo. The solid line with surrounding dashed lines represents the mean result and scatter found by Giovanelli et al. (1997). Bottom: cold gas metallicity as a function of total stellar mass. The solid line represents the result of Tremonti et al. (2004).

Ejected gas leaves the galaxy and its current halo in a wind or 'super-wind', but it need not be lost permanently. As the dark halo grows, some of the surrounding ejecta may fall back in and be reincorporated into the cooling cycle. We follow Springel et al. (2001a) and De Lucia et al. (2004) and model this by assuming

$$\dot{m}_{\rm ejected} = -\gamma_{\rm ej} m_{\rm ejected} / t_{\rm dyn}, \tag{21}$$

where $\gamma_{\rm ej}$ controls the amount of reincorporation per dynamical time; typically we take $\gamma_{\rm ej} = 0.3$–1.0. Such values imply that all the ejected gas will return to the hot halo in a few halo dynamical times.

The prescriptions given in this section are simple, as well as physically and energetically plausible, but they have little detailed justification either from observation or from numerical simulation. They allow us to track in a consistent way the exchange of each halo's baryons between our four phases (stars, cold disc gas, hot halo gas, ejecta), but should be regarded as a rough attempt to model the complex astrophysics of feedback which will surely need significant modification as the observational phenomenology of these processes is explored in more depth.

Supernova feedback and star formation act together to regulate the stellar growth of a galaxy. This is especially important for $L < L^*$ galaxies, where feedback can eject most of the baryons from the system, reducing the supply of star-forming material for time periods much longer than the cooling/supernova heating cycle. In the middle panel of Fig. 6 we plot the Tully–Fisher relation for model Sb/c galaxies (see Section 3.5). The Tully–Fisher relation is strongly influenced by the link between star formation and supernova heating. The circular velocity of a galactic disc is (to first order) proportional to the virial velocity of the host dark matter halo and thus to its escape velocity. In our model (and most others) this is closely related to the ability of the galaxy to blow a wind. The luminosity of a galaxy is determined by its ability to turn its associated baryons into stars. The overall efficiency of this process in the face of supernova and AGN feedback sets the amplitude of the Tully–Fisher relation, while the way in which the efficiency varies between systems of different circular velocity has a strong influence on the slope.

To predict a Tully–Fisher relation for our model we need to assign a maximum rotation velocity to each of our galaxy discs. For central galaxies we simply take this velocity to be the $V_{\rm vir}$ of the surrounding halo, while for satellite galaxies we take it to be the $V_{\rm vir}$ of the surrounding halo at the last time the galaxy was a central galaxy. This is a crude approximation, and for realistic halo structures it is likely to be an underestimate both because of the concentration of the dark matter distribution and because of the gravitational effects of the baryons (e.g. Mo et al. 1998). Obtaining a good simultaneous fit to observed luminosity functions and Tully–Fisher relations remains a difficult problem within the ΛCDM paradigm (see for example Cole et al. 2000). Our unrealistic assumption for the disc rotation velocity actually produces quite a good fit to the observational data of Giovanelli et al. (1997), demonstrating that our simple star formation and feedback recipes can adequately represent the growth of stellar mass across a wide range of scales. We find clear deviations from power-law behaviour for $\log W \lesssim 2.3$ (approximately $V_{\rm c} \lesssim 100\,{\rm km\,s^{-1}}$), where the efficiency of removing gas from low-mass systems combines with our threshold for the onset of star formation to reduce the number of stars that can form. The resulting downward bend is qualitatively similar to that pointed out in real data by McGaugh et al. (2000). These authors show that including the gaseous component to construct a 'baryonic' Tully–Fisher relation brings the observed points much closer to a power law, and the same is true in the model that we present here. Limiting star formation in galaxies that inhabit shallow potentials has a strong effect on the faint end of the galaxy luminosity function, as will be seen in Section 4.2.

3.7 Galaxy morphology, merging and starbursts

In the model that we discuss here, the morphology of a galaxy is assumed to depend only on its bulge-to-total luminosity ratio, which in turn is determined by three distinct physical processes: disc growth by accretion, disc buckling to produce bulges, and bulge formation through mergers. We treat disc instabilities using the simple analytic stability criterion of Mo et al. (1998); the stellar disc of a galaxy becomes unstable when the following inequality is met:

$$\frac{V_{\rm c}}{({\rm G}m_{\rm disk}/r_{\rm disk})^{1/2}} \leqslant 1, \tag{22}$$

where $m_{\rm disk}$ and $r_{\rm disk}$ are the mass and radius of the disk respectively (see Section 3.5), and we again approximate the rotation velocity of the disc $V_{\rm c}$ by $V_{\rm vir}$. For each galaxy at each time-step we evaluate the left-hand side of equation (22), and if it is smaller than unity we transfer enough stellar mass from disc to bulge (at fixed $r_{\rm D}$) to restore stability.

Galaxy mergers shape the evolution of galaxies, affecting both their morphology and (through induced starbursts) their star formation history. Mergers can occur in our model between the central galaxy of a dark halo or subhalo and a satellite galaxy which has lost its own dark subhalo. Substructure is followed in the Millennium Run down to a 20-particle limit, which means that the orbit of a satellite galaxy within a larger halo is followed explicitly until its subhalo mass drops below $1.7 \times 10^{10}\,h^{-1}\,{\rm M}_{\odot}$. After this point, the position and velocity of the satellite are represented by those of the most bound particle of the subhalo at the last time it was identified. At the same time, however, we start a merger 'clock' and estimate a merging time for the galaxy using the dynamical friction formula of Binney & Tremaine (1987),

$$t_{\rm friction} = 1.17 \frac{V_{\rm vir} r_{\rm sat}^2}{{\rm G}m_{\rm sat} \ln \Lambda}. \tag{23}$$

This formula is valid for a satellite of mass $m_{\rm sat}$ orbiting in an isothermal potential of circular velocity $V_{\rm vir}$ at radius $r_{\rm sat}$. We take $m_{\rm sat}$ and $r_{\rm sat}$ to be the values measured for the galaxy at the last time its subhalo could be identified. The Coulomb logarithm is approximated by $\ln \Lambda = \ln(1 + M_{\rm vir}/m_{\rm sat})$. The satellite is then merged with the central galaxy a time $t_{\rm friction}$ after its own subhalo was last identified. If the main halo merges with a larger system before this occurs, a new value for $t_{\rm friction}$ is calculated and the merger clock is restarted.

The outcome of the merger will depend on the baryonic mass ratio of the two progenitors . When one dominates the process, i.e. a small satellite merging with a larger central galaxy, the stars of the satellite are added to the bulge of the central galaxy and a minor merger starburst (see below) will result. The cold gas of the satellite is added to the disc of the central galaxy along with any stars that formed during the burst. Such an event is called a *minor merger*.

If, on the other hand, the masses of the progenitors are comparable a *major merger* will result. Under these circumstances the starburst is more significant, with the merger destroying the discs of both galaxies to form a spheroid in which all stars are placed. The dividing line between a major and minor merger is given by the parameter $T_{\rm merger}$: when the mass ratio of the merging progenitors is larger than $T_{\rm merger}$ a major merger results, otherwise the event is a minor merger. Following Springel et al. (2001a) we choose $T_{\rm merger} = 0.3$ and keep this fixed throughout.

Our starburst implementation is based on the 'collisional starburst' model of Somerville et al. (2001). In this model, a fraction $e_{\rm burst}$ of the combined cold gas from the two galaxies is turned into stars as a result of the merger:

$$e_{\rm burst} = \beta_{\rm burst} (m_{\rm sat}/m_{\rm central})^{\alpha_{\rm burst}}, \tag{24}$$

where the two parameters are taken as $\alpha_{\rm burst} = 0.7$ and $\beta_{\rm burst} = 0.56$. This model provides a good fit to the numerical results of Cox et al.

(2004) and also Mihos & Hernquist (1994, 1996) for merger mass ratios ranging from 1:10 to 1:1.

3.8 Spectroscopic evolution and dust

The photometric properties of our galaxies are calculated using stellar population synthesis models from Bruzual & Charlot (2003). Our implementation is fully described in De Lucia et al. (2004) and we refer the reader there (and to references therein) for further details.

To include the effects of dust when calculating galaxy luminosities we apply the simple 'plane-parallel slab' model of Kauffmann et al. (1999). This model is clearly oversimplified, but it permits us to make a reasonable first-order correction for dust extinction in actively star-forming galaxies. For the details of this model we refer the reader to Kauffmann et al. (1999) and to references therein.

3.9 Metal enrichment

Our treatment of metal enrichment is essentially identical to that described in De Lucia et al. (2004). In this model a yield Y of heavy elements is returned for each solar mass of stars formed. These metals are produced primarily in the supernovae which terminate the evolution of short-lived, massive stars. In our model we deposit them directly into the cold gas in the disc of the galaxy. (An alternative would clearly be to add some fraction of the metals directly to the hot halo. Limited experiments suggest that this makes little difference to our main results.) We also assume that a fraction R of the mass of newly formed stars is recycled immediately into the cold gas in the disc, the so-called 'instantaneous recycling approximation' (see Cole et al. 2000). For full details on metal enrichment and exchange processes in our model, see De Lucia et al. (2004). In the bottom panel of Fig. 6 we show the metallicity of cold disc gas for model Sb/c galaxies (selected, as before, by bulge-to-total luminosity, as described in Section 3.5) as a function of total stellar mass. For comparison, we show the result of Tremonti et al. (2004) for mean H II region abundances in SDSS galaxies.

4 RESULTS

In this section we examine how the suppression of cooling flows in massive systems affects galaxy properties. As we will show, the effects are only important for high-mass galaxies. Throughout our analysis we use the galaxy formation model outlined in the previous sections with the parameter choices of Table 1 except where explicitly noted.

4.1 The suppression of cooling flows

We begin with Fig. 7, which shows how our 'radio mode' heating model modifies gas condensation. We compare mean condensation rates with and without the central AGN heating source as a function of halo virial velocity (solid and dashed lines respectively). Recall that virial velocity provides a measure of the equilibrium temperature of the system through $T_{\rm vir} \propto V_{\rm vir}^2$, as indicated by the scale on the top axis. The four panels show the behaviour at four redshifts between 6 and the present day. The vertical dotted line in each panel marks haloes for which $r_{\rm cool} = R_{\rm vir}$, the transition between systems that form static hot haloes and those where infalling gas cools rapidly on to the central galaxy disc (see Section 3.2 and Fig. 2). This transition moves to haloes of lower temperature with time, suggesting a 'down-sizing' of the characteristic mass of actively star-forming galaxies. At lower $V_{\rm vir}$ gas continues to cool rapidly, while at higher

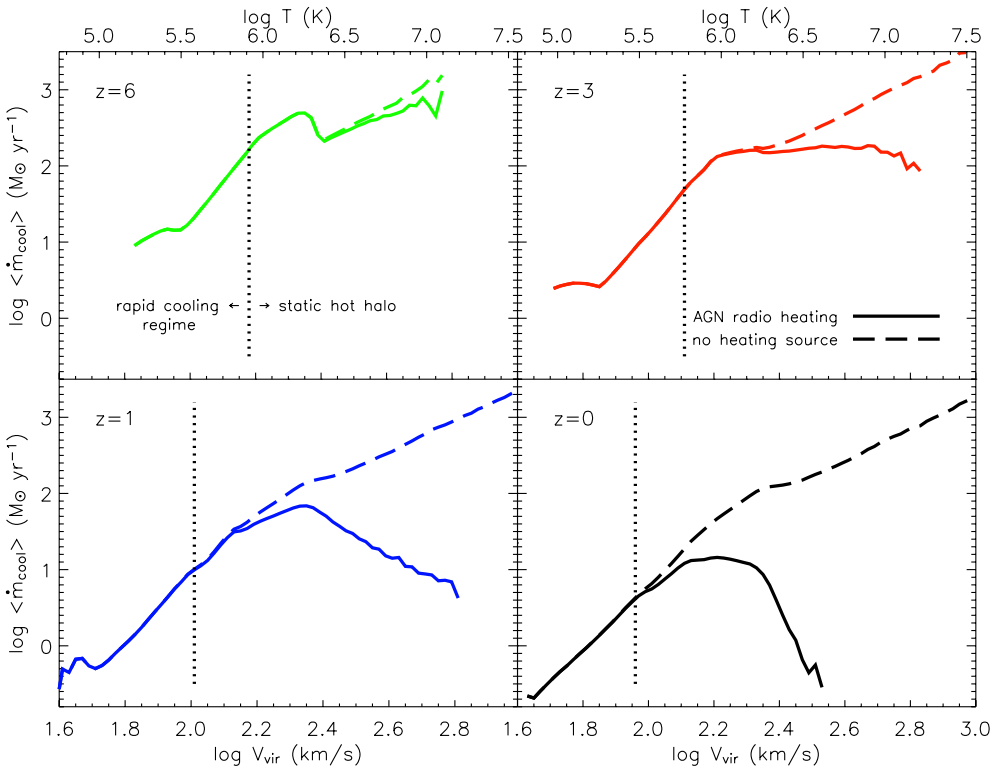

Figure 7. The mean condensation rate, $\langle \dot{m}_{\rm cool} \rangle$ as a function of halo virial velocity $V_{\rm vir}$ at redshifts of 6, 3, 1 and 0. Solid and dashed lines in each panel represent the condensation rate with and without 'radio mode' feedback respectively, while the vertical dotted lines show the transition between the rapid cooling and static hot halo regimes, as discussed in Section 3.2. This figure demonstrates that cooling flow suppression is most efficient in our model for haloes with $V_{\rm vir} > 150\,{\rm km\,s^{-1}}$ and at $z \leqslant 3$.

$V_{\rm vir}$ new fuel for star formation must come from cooling flows which are affected by 'radio mode' heating.

The effect of 'radio mode' feedback is clearly substantial. Suppression of condensation becomes increasingly effective with increasing virial temperature and decreasing redshift. The effects are large for haloes with $V_{\rm vir} \gtrsim 150\,{\rm km\,s^{-1}}$ ($T_{\rm vir} \gtrsim 10^6$ K) at $z \lesssim 3$. Condensation stops almost completely between $z = 1$ and the present in haloes with $V_{\rm vir} > 300\,{\rm km\,s^{-1}}$ ($T_{\rm vir} > 3 \times 10^6$ K). Such systems correspond to the haloes of groups and clusters which are typically observed to host massive elliptical or cD galaxies at their centres. Our scheme thus produces results which are qualitatively similar to the ad hoc suppression of cooling flows assumed in previous models of galaxy formation. For example, Kauffmann et al. (1999) switched off gas condensation in all haloes with $V_{\rm vir} > 350\,{\rm km\,s^{-1}}$, while Hatton et al. (2003) stopped condensation when the bulge mass exceeded a critical threshold.

4.2 Galaxy properties with and without AGN heating

The suppression of cooling flows in our model has a dramatic effect on the bright end of the galaxy luminosity function. In Fig. 8 we present K- and $b_{\rm J}$-band luminosity functions (left- and right-hand panels respectively) with and without 'radio mode' feedback (solid and dashed lines respectively). The luminosities of bright galaxies are reduced by up to two magnitudes when the feedback is switched on, and this induces a relatively sharp break in the luminosity function which matches the observations well. We demonstrate this by overplotting K-band data from Cole et al. (2001) and Huang et al. (2003) in the left-hand panel, and $b_{\rm J}$-band data from Norberg et al. (2002) in the right-hand panel. In both bandpasses the model is quite close to the data over the full observed range. We comment on some of the remaining discrepancies below.

Our feedback model also has a significant effect on bright galaxy colours, as we show in Fig. 9. Here we plot the $B - V$ colour distribution as a function of stellar mass, with and without the central heating source (top and bottom panels respectively). In both panels we have colour-coded the galaxy population by morphology as estimated from bulge-to-total luminosity ratio (split at $L_{\rm bulge}/L_{\rm total} = 0.4$). Our morphological resolution limit is marked by the dashed line at a stellar mass of $\sim 4 \times 10^9\,{\rm M}_\odot$; this corresponds approximately to a halo of 100 particles in the Millennium Run. Recall that the morphology of a galaxy depends both on its past merging history and on the stability of its stellar disc in our model. Both mergers and disc instabilities contribute stars to the spheroid, as described in Section 3.7. The build-up of haloes containing fewer than 100 particles is not followed in enough detail to model these processes robustly.

A number of important features can be seen in Fig. 9. Of note is the bimodal distribution in galaxy colours, with a well-defined red sequence of appropriate slope separated cleanly from a broader 'blue cloud'. It is significant that the red sequence is composed predominantly of early-type galaxies, while the blue cloud is composed mostly of disc-dominated systems. This aspect of our model suggests that that the physical processes that determine morphology (i.e. merging, disc instability) are closely related to those that control star formation history (i.e. gas supply) and thus determine galaxy colour. The red and blue sequences both display a strong metallicity gradient from low to high mass (c.f. Fig. 6), and it is this which induces a 'slope' in the colour–magnitude relations which agrees well with observation (e.g. Baldry et al. 2004).

By comparing the upper and lower panels in Fig. 9 we can see how 'radio mode' feedback modifies the luminosities, colours and morphologies of high-mass galaxies. Not surprisingly, the brightest and most massive galaxies are also the reddest and are ellipticals when cooling flows are suppressed, whereas they are brighter, more massive, much bluer and typically have discs if cooling flows continue to supply new material for star formation. AGN heating cuts off the gas supply to the disc from the surrounding hot halo, truncating star formation and allowing the existing stellar population to redden. However, these massive red galaxies do continue to grow through merging. This mechanism allows the dominant cluster galaxies to gain a factor of 2 or 3 in mass without significant star formation, in apparent agreement with observation (Aragon-Salamanca, Baugh & Kauffmann 1998). This late-stage (i.e. $z \lesssim 1$) hierarchical growth moves objects to higher mass without changing their colours.

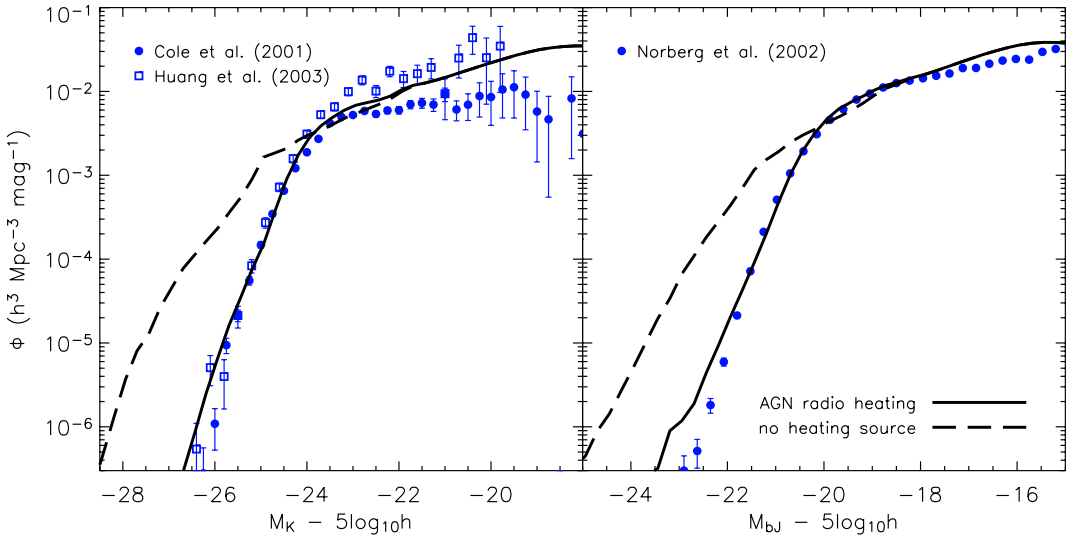

Figure 8. Galaxy luminosity functions in the K (left) and $b_{\rm J}$ (right) photometric bands, plotted with and without 'radio mode' feedback (solid and long-dashed lines respectively – see Section 3.4). Symbols indicate observational results as listed in each panel. As can be seen, the inclusion of AGN heating produces a good fit to the data in both colours. Without this heating source our model overpredicts the luminosities of massive galaxies by about two magnitudes and fails to reproduce the sharp bright-end cut-offs in the observed luminosity functions.

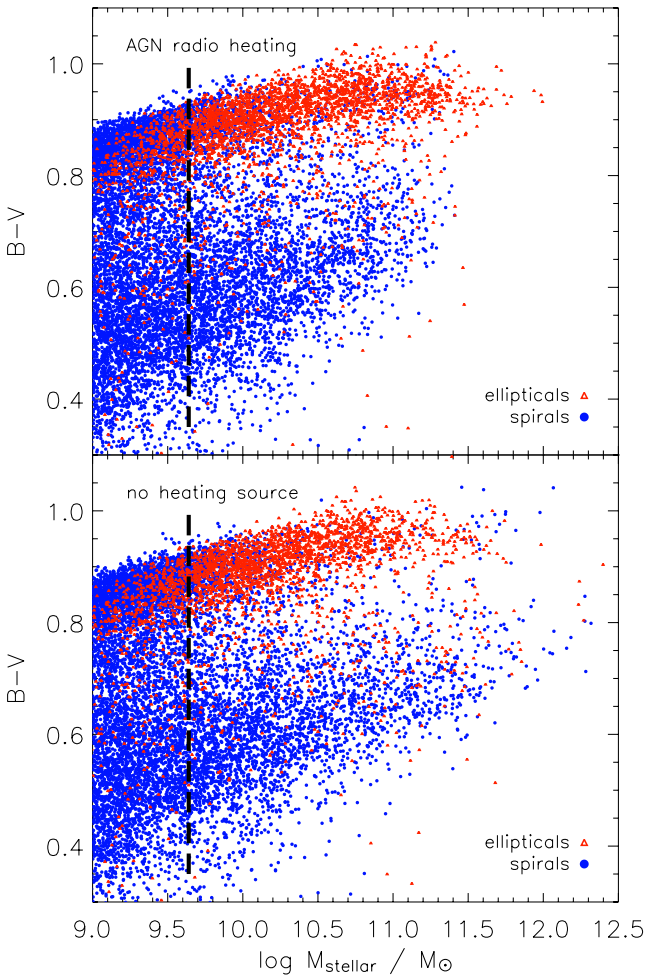

Figure 9. The $B - V$ colours of model galaxies plotted as a function of their stellar mass with (top) and without (bottom) 'radio mode' feedback (see Section 3.4). A clear bimodality in colour is seen in both panels, but without a heating source the most massive galaxies are blue rather than red. Only when heating is included are massive galaxies as red as observed. Triangles (red in the online version of the article) and circles (blue online) correspond to early and late morphological types respectively, as determined by bulge-to-total luminosity ratio (see Section 4.2). The thick dashed lines mark the resolution limit to which morphology can be reliably determined in the Millennium Run.

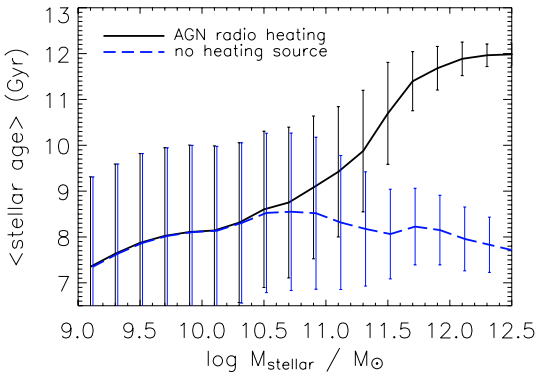

Figure 10. Mean stellar ages of galaxies as a function of stellar mass for models with and without 'radio mode' feedback (solid and dashed lines respectively). Error bars show the rms scatter around the mean for each mass bin. The suppression of cooling flows raises the mean age of high-mass galaxies to large values, corresponding to high formation redshifts.

It is also interesting to examine the effect of AGN heating on the stellar ages of galaxies. In Fig. 10 the solid and dashed lines show mean stellar age as a function of stellar mass for models with and without 'radio mode' feedback, while error bars indicate the rms scatter around the mean. Substantial differences are seen for galaxies with $M_{\rm stellar} \gtrsim 10^{11}\,{\rm M}_{\odot}$: the mean age of the most massive galaxies approaching 12 Gyr when cooling flows are suppressed but remaining around 8 Gyr when feedback is switched off. Such young ages are clearly inconsistent with the old stellar populations observed in the majority of massive cluster ellipticals.

The colour bimodality in Fig. 9 is so pronounced that it is natural to divide our model galaxies into red and blue populations and to study their properties separately. We do this by splitting at $B - V = 0.8$, an arbitrary but natural choice. Fig. 11 shows separate $b_{\rm J}$ luminosity functions for the resulting populations. For comparison we overplot observational results from Madgwick et al. (2002) for 2dFGRS galaxies split by spectral type. Their luminosity functions are essentially identical to those of Cole et al. (2005), who split the 2dFGRS catalogue by $b_{\rm J} - {\rm r_F}$ colour. It thus can serve to indicate the observational expectations for populations of different colour. The broad behaviour of the red and blue populations is similar in the model and in the 2dFGRS. The faint end of the luminosity function is dominated by late-types, whereas the bright end has an excess of early-types. The two populations have equal abundance about 0.5–1 mag brighter than $M^*_{b_{\rm J}}$ (Norberg et al. 2002).

Fig. 11 also shows some substantial differences between model and observations. The red and blue populations differ more in the real data than they do in the model. There is a tail of very bright blue galaxies in the model, which turn out to be objects undergoing strong, merger-induced starbursts. These correspond in abundance, star formation rate and evolutionary state to the observed population of ultraluminous infrared galaxies (ULIRGs), with the important difference that almost all the luminosity from young stars in the real systems is absorbed by dust and re-emitted in the mid- to far-infrared (Sanders & Mirabel 1996). Clearly we need better dust modelling than our simple 'slab' model (Section 3.8) in order to reproduce the properties of such systems adequately. If we suppress starbursts in bright galaxy mergers we find that the blue tail disappears and the observed behaviour is recovered. A second and substantial discrepancy is the apparent overproduction of faint red galaxies in our model, as compared with the 2dFGRS measurements [however, see Popesso et al. (2005) and Gonzalez et al. (2005)]. Further work is clearly needed to understand the extent and significance of this difference.

5 PHYSICAL MODELS OF AGN FEEDBACK

Our phenomenological model for 'radio mode' feedback (Section 3.4.2) is not grounded in any specific model for hot gas accretion on to a black hole or for the subsequent generation and thermalization of energy through radio outflows. Rather it is based on the observed properties of cooling flows and their central radio sources, and on the need for a source of feedback that can suppress gas condensation on to massive galaxies without requiring the formation of new stars. We have so far focused on the effects of such feedback without discussing how it might be realized. In this section we present two physical models which suggest how accretion on to the central black hole may lead to activity in a way that could justify the parameter scalings we have adopted.

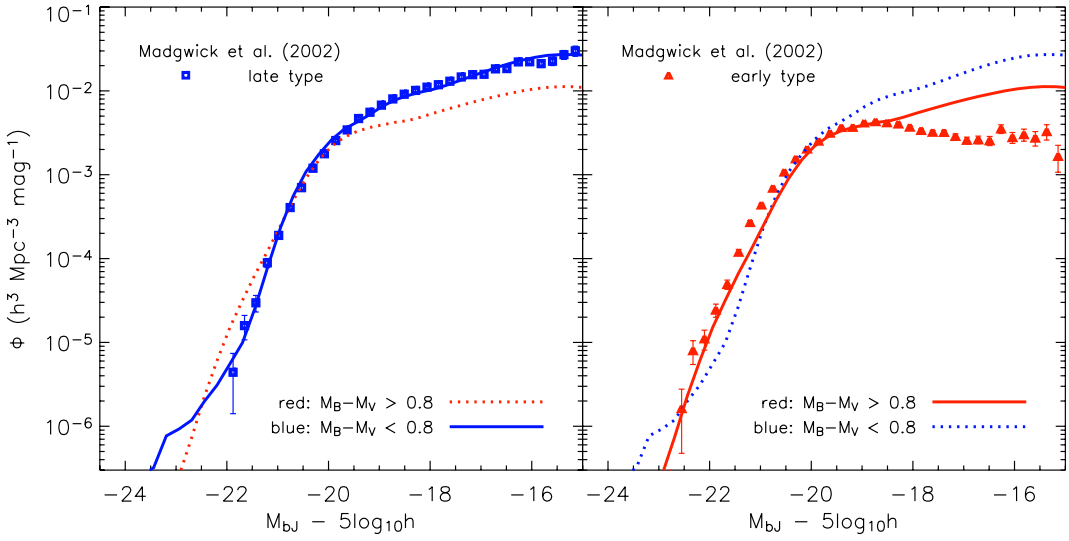

Figure 11. The b_J-band galaxy luminosity function split by colour at $B-V=0.8$ (Fig. 9) into blue (left-hand panel) and red (right-hand panel) sub-populations (solid lines). The dotted lines in each panel repeat the opposite colour luminosity function for reference. Symbols indicate the observational results of Madgwick et al. (2002) for early- and late-type 2dFGRS galaxies, split according to spectral type. Although our model split by colour captures the broad behaviour of the observed type-dependent luminosity functions, there are important differences which we discuss in Section 4.2.

5.1 Cold cloud accretion

A simple picture for cooling flow evolution, based on the similarity solution of Bertschinger (1989) for an unperturbed halo in isolation, can be summarized as follows. Cooling flows develop in any halo where the cooling time of the bulk of the hot gas is longer than the age of the system so that a static hot halo can form. Such haloes usually have a strong central concentration and we approximate their structure by a singular isothermal sphere. The inner regions then have a local cooling time shorter than the age of the system, and the gas that they contain radiates its binding energy and flows inwards. The flow region is bounded by the cooling radius $r_{\rm cool}$ where the local cooling time is equal to the age of the system (see Section 3.2). This radius increases with time as $t^{1/2}$. As Bertschinger showed, the temperature of the gas *increases* by about 20 per cent as it starts to flow inwards, and its density profile flattens to $\rho_{\rm g} \propto r^{-3/2}$. Initially, the flow is subsonic and each gas shell sinks stably and isothermally in approximate hydrostatic equilibrium. As it sinks, however, its inward motion accelerates because its cooling time shrinks more rapidly than the sound traveltime across it, and at the sonic radius, $r_{\rm sonic}$, the two become equal. At this point the shell goes into free fall, its temperature decreases rapidly and it may fragment as a result of thermal instability (Cowie, Fabian & Nulsen 1980; Nulsen 1986; Balbus & Soker 1989). The dominant component of the infalling gas is then in the form of cold clouds and is no longer self-coupled by hydrodynamic forces. Different clouds pursue independent orbits, some with pericentres perhaps orders of magnitude smaller than $r_{\rm sonic}$. If these lie within the zone of influence of the black hole, $r_{\rm BH} = {\rm G}\, m_{\rm BH}/V_{\rm vir}^2$, we assume that some of the cold gas becomes available for fuelling the radio source; otherwise we assume it to be added to the cold gas disc.

The parameter scalings implied by this picture can be estimated as follows. The sound traveltime across a shell at the cooling radius is shorter than the cooling time by a factor $\sim r_{\rm cool}/R_{\rm vir}$. At smaller radii the ratio of cooling time to sound traveltime decreases as $r^{1/2}$ so that $r_{\rm sonic}/r_{\rm cool} \sim (r_{\rm cool}/R_{\rm vir})^2$, implying $r_{\rm sonic} \sim r_{\rm cool}^3/R_{\rm vir}^2$. If we adopt $r_{\rm BH} > 10^{-4}\, r_{\rm sonic}$ as the condition for effective fuelling of the radio source, we obtain

$$m_{\rm BH} > 10^{-4} M_{\rm vir} (r_{\rm cool}/R_{\rm vir})^3 \qquad (25)$$

as the corresponding minimum black hole mass for fragmented clouds to be captured. Under such conditions, only a small fraction ($\sim$0.01 per cent) of the cooling flow mass need be accreted to halt the flow. The ratio in parentheses on the right-hand side of this equation scales approximately as $r_{\rm cool}/R_{\rm vir} \propto (m_{\rm hot}/M_{\rm vir})^{1/2}\, t_{\rm H}^{-1/2}\, V_{\rm vir}^{-1}$, so the minimum black hole mass scales approximately as $(m_{\rm hot}/M_{\rm vir})^{3/2}\, t_{\rm H}^{-1/2}$ and is almost independent of $V_{\rm vir}$. In our model, the growth of black holes through mergers and 'quasar mode' accretion produces a population where mass increases with time and with host halo mass. As a result, effective fuelling takes place primarily in the more massive haloes and at late times for this 'cold cloud' prescription.

To test this particular model we switch off our standard phenomenological treatment of 'radio mode' feedback (Section 3.4.2), assuming instead that feedback occurs only when equation (25) is satisfied and that in this case it is sufficient to prevent further condensation of gas from the cooling flow. All other elements of our galaxy and black hole formation model are unchanged. The resulting cooling flow suppression is similar to that seen in Fig. 7, and all results presented in Sections 3 and 4 are recovered. An illustration of this is given by Fig. 12, where we compare the K-band luminosity function from this particular model (the dashed line) to the observational data (cf. also Fig. 8). The model works so well, of course, because the numerical coefficient in equation (25) is uncertain and we have taken advantage of this to choose a value that puts the break in the luminosity function at the observed position. This adjustment plays the role of the efficiency parameter $\kappa_{\rm AGN}$ in our standard analysis (see equation 10).

5.2 Bondi–Hoyle accretion

Our second physical model differs from the first in assuming that accretion is not from the dominant, cold cloud component which forms within the sonic radius, but rather from a subdominant hot

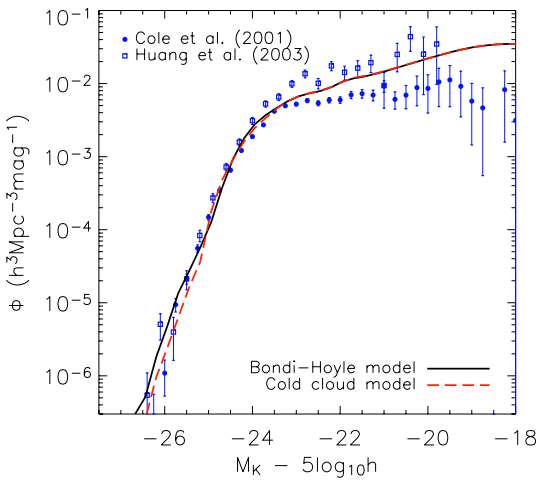

Figure 12. The observed K-band galaxy luminosity function is compared with the results from models using the two physical prescriptions for 'radio mode' accretion discussed in Section 5: the Bondi–Hoyle accretion model (solid line) and the cold cloud accretion model (dashed line). Symbols indicate observational data from Cole et al. (2001) and Huang et al. (2003). Both models can produce a luminosity function that matches observation well.

component which fills the space between these clouds. The clouds themselves are assumed to be lost to the star-forming disc. The density profile of the residual hot component was estimated by Nulsen & Fabian (2000) from the condition that the cooling time of each radial shell should remain equal to the sound traveltime across it as it flows inwards. This requires the density of the hot component to vary as $1/r$ within $r_{\rm sonic}$, and thermal instabilities must continually convert material into condensed clouds in order to maintain this structure as the hot gas flows in.

The rate at which hot gas is accreted on to the black hole can then be estimated from the Bondi–Hoyle formula (Bondi 1952; Edgar 2004):

$$\dot{m}_{\rm Bondi} = 2.5\pi {\rm G}^2 \frac{m_{\rm BH}^2 \rho_0}{c_{\rm s}^3}. \quad (26)$$

Here ρ_0 is the (assumed uniform) density of hot gas around the black hole, and in all that follows we approximate the sound speed, $c_{\rm s}$, by the virial velocity of the halo, $V_{\rm vir}$. Of course, the density distribution of gas surrounding the black hole is *not* uniform so the question immediately arises as to what density we should choose. We follow a suggestion of Churazov et al. (2005) and use the value predicted by the 'maximum cooling flow' model of Nulsen & Fabian (2000) at the Bondi radius $r_{\rm Bondi} \equiv 2GM_{\rm BH}/c_{\rm s}^2 = 2r_{\rm BH}$, the conventional boundary of the sphere of influence of the black hole. We therefore equate the sound traveltime across a shell at this radius to the local cooling time there:

$$\frac{2r_{\rm Bondi}}{c_{\rm s}} \approx \frac{4{\rm G}m_{\rm BH}}{V_{\rm vir}^3} = \frac{3}{2}\frac{\bar{\mu} m_{\rm p} kT}{\rho_{\rm g}(r_{\rm Bondi})\Lambda(T,Z)}. \quad (27)$$

Solving for the density gives

$$\rho_0 = \rho_{\rm g}(r_{\rm Bondi}) = \frac{3\mu m_{\rm p}}{8{\rm G}}\frac{kT}{\Lambda}\frac{V_{\rm vir}^3}{m_{\rm BH}}. \quad (28)$$

Combining equations (28) and (26) provides us with the desired estimate for the hot gas accretion rate on to the black hole:

$$\dot{m}_{\rm Bondi} \approx {\rm G}\mu m_{\rm p}\frac{kT}{\Lambda} m_{\rm BH}. \quad (29)$$

Notice that this rate depends only on the black hole mass and on the virial temperature of the halo. It is independent both of time and of $m_{\rm hot}/M_{\rm vir}$, the hot gas fraction of the halo. It is valid as long as $r_{\rm Bondi} < r_{\rm sonic}$, which is always the case in our models.

To investigate the effects of this model we replace the phenomenological 'radio mode' accretion rate of equation (10) with that given by equation (29). Since the latter has no adjustable efficiency, we use the energy generation parameter η of equation (11) to control the effectiveness of cooling flow suppression. [This was not necessary before since η always appeared in the product $\eta\kappa_{\rm AGN}$, where $\kappa_{\rm AGN}$ is the efficiency parameter of equation (10).] With this change of cooling flow accretion model and taking $\eta = 0.03$, we are able to recover the results of Sections 3 and 4 without changing any other aspects of our galaxy and black hole formation model. The final galaxy population is, in fact, almost identical to that presented in previous sections. This is not surprising, perhaps, since equation (29) has very similar scaling properties to equation (10). In Fig. 12 we illustrate the success of the model by overplotting its prediction for the K-band luminosity function (the solid line) on the observational data and on the prediction of the cold cloud accretion model of the last subsection. The two models agree very closely both with each other and with our standard phenomenological model (see Fig. 8).

6 CONCLUSIONS

AGN feedback is an important but relatively little-explored element in the co-evolution of galaxies and the supermassive black holes at their centres. In this paper we set up machinery to study this co-evolution in unprecedented detail using the very large Millennium Run, a high-resolution simulation of the growth of structure in a representative region of the concordance ΛCDM cosmology. Most of our modelling follows earlier work, but in an important extension we introduce a 'radio' feedback mode, based on simple physical models and on the observed phenomenology of radio sources in cooling flows. This mode suppresses gas condensation at the centres of massive haloes without requiring the formation of new stars. Our modelling produces large catalogues of galaxies and supermassive black holes which can be used to address a very wide range of issues concerning the evolution and clustering of galaxies and AGN. Some clustering results were already presented in Springel et al. (2005b). In the present paper, however, we limit ourselves to presenting the model in some detail and to investigating the quite dramatic effects which 'radio mode' feedback can have on the luminosities and colours of massive galaxies. Our main results can be summarized as follows.

(i) We study the amount of gas supplied to galaxies in each of the two gas infall modes discussed by White & Frenk (1991): the 'static hot halo' mode where post-shock cooling is slow and a quasi-static hot atmosphere forms behind the accretion shock; and the 'rapid cooling' mode where the accretion shock is radiative and no such atmosphere is present. We distinguish these two modes using the criterion of White & Frenk (1991) as modified by Springel et al. (2001a) and tested explicitly using SPH simulations by Yoshida et al. (2002). Our results show a sharp transition between the two regimes at a halo mass of 2–3 $\times$ $10^{11}\,{\rm M}_\odot$. This division depends on the chemical enrichment prescription adopted and moves from higher to lower $V_{\rm vir}$ with time (corresponding to approximately constant $M_{\rm vir}$), suggesting a 'down-sizing' of star formation activity as the bulk of the gas accreted by the haloes of larger systems is no longer available in the interstellar medium of the central galaxy.

(ii) We have built a detailed model for cooling, star formation, supernova feedback, galaxy mergers and metal enrichment based on the earlier models of Kauffmann et al. (1999), Springel et al. (2001a) and De Lucia et al. (2004). Applied to the Millennium Run, this model reproduces many of the observed properties of the local galaxy population: the Tully–Fisher, cold gas fraction/stellar mass and cold gas metallicity/stellar mass relations for Sb/c spirals (Fig. 6); the field galaxy luminosity functions (Figs 8 and 11); the colour–magnitude distribution of galaxies (Fig. 9); and the increase in mean stellar age with galaxy mass (Fig. 10). In addition the model produces a global star formation history in reasonable agreement with observation (Fig. 5). We also show in Springel et al. (2005b) that the $z = 0$ clustering properties of this population are in good agreement with observations.

(iii) Our black hole implementation extends the previous work of Kauffmann & Haehnelt (2000) by assuming three modes of AGN growth: merger-driven accretion of cold disc gas in a 'quasar mode'; merging between black holes; and 'radio mode' accretion when a massive black hole finds itself at the centre of a static hot gas halo. The 'quasar mode' is the dominant source for new black hole mass and is most active between redshifts of 4 and 2. The 'radio mode' grows in overall importance until $z = 0$ and is responsible for the feedback which shuts off the gas supply in cooling flows. This model reproduces the black hole mass/bulge mass relation observed in local galaxies (Fig. 4). The global history of accretion in the 'quasar mode' is qualitatively consistent with the evolution inferred from the optical AGN population (Fig. 3).

(iv) Although the overall accretion rate is low, 'radio mode' outflows can efficiently suppress condensation in massive systems (Fig. 7). As noted by many authors who have studied the problem in more detail than we do, this provides an energetically feasible solution to the long-standing cooling flow 'problem'. Our analysis shows that the resulting suppression of gas condensation and star formation can produce luminosity functions with very similar bright-end cut-offs to those observed (Fig. 8), as well as colour–magnitude distributions in which the most massive galaxies are red, old and elliptical, rather than blue, young and disc-dominated (Figs 9 and 10).

(v) The $B - V$ colour distribution of galaxies is bimodal at all galaxy masses. Galaxies with early-type bulge-to-disc ratios are confined to the red sequence, as are the most massive galaxies, and the most massive galaxies are almost all bulge-dominated, as observed in the real Universe (Fig. 9). This bimodality provides a natural division of model galaxies into red and blue subpopulations. The colour-dependent luminosity functions are qualitatively similar to those found for early- and late-type galaxies in the 2dFGRS (Fig. 11), although there are significant discrepancies. After exhausting their cold gas, massive central galaxies grow on the red sequence through 'burstless' merging, gaining a factor of 2 or 3 in mass without significant star formation (Aragon-Salamanca et al. 1998). Such hierarchical growth does not change the colour of a galaxy significantly, moving it brightward almost parallel to the colour–magnitude relation.

(vi) We present two physical models for black hole accretion from cooling flow atmospheres. We suppose that this accretion is responsible for powering the radio outflows seen at the centre of almost all real cooling flows. The models differ in their assumptions about how gas accretes from the inner regions of the cooling flow, where it is thermally unstable and dynamically collapsing. One assumes accretion of cold gas clouds if these come within the sphere of influence of the black hole, while the other assumes Bondi-like accretion from the residual diffuse hot gas component. Each of the two models can produce $z = 0$ galaxy populations similar both to that of our simple phenomenological model for 'radio mode' feedback and to the observed population (see Fig. 12). Our main results are thus not sensitive to the details of the assumed accretion models.

The presence of heating from a central AGN has long been suspected as the explanation for the apparent lack of gas condensation in cluster cooling flows. We have shown that including a simple treatment of this process in galaxy formation models not only 'solves' the cooling flow problem, but also dramatically affects the properties of massive galaxies, inducing a cut-off similar to that observed at the bright end of the galaxy luminosity function, and bringing colours, morphologies and stellar ages into much better agreement with observation than is the case for models without such feedback. We will extend the work presented here in a companion paper, where we investigate the growth of supermassive black holes and the related AGN activity as a function of host galaxy properties out to high redshift. The catalogues of galaxies and supermassive black holes produced by our modelling machinery are also being used for a very wide range of projects related to understanding formation, evolution and clustering processes, as well as for interpreting observational samples.

ACKNOWLEDGMENTS

DJC acknowledges the financial support of the International Max Planck Research School in Astrophysics PhD Fellowship. GDL thanks the Alexander von Humboldt Foundation, the Federal Ministry of Education and Research, and the Programme for Investment in the Future (ZIP) of the German Government for financial support. Many thanks to Andrea Merloni and Jiasheng Huang. Special thanks to Eugene Churazov for numerous valuable discussions and suggestions. We also thank the anonymous referee for a number of valuable suggestions which helped to improve the quality of this paper. The Millennium Run simulation was carried out by the Virgo Supercomputing Consortium at the Computing Centre of the Max Planck Society in Garching. Catalogues of galaxies from the modelling reported in this paper are available at http://www.mpa-garching.mpg.de/galform/agnpaper.

REFERENCES

Aragon-Salamanca A., Baugh C. M., Kauffmann G., 1998, MNRAS, 297, 427
Arav N. et al., 2001, ApJ, 561, 118
Bagla J. S., 2002, JA&A, 23, 185
Balbus S. A., Soker N., 1989, ApJ, 341, 611
Baldry I. K., Glazebrook K., Brinkmann J., Ivezić Ž., Lupton R. H., Nichol R. C., Szalay A. S., 2004, ApJ, 600, 681
Barnes J., Hut P., 1986, Nat, 324, 446
Begelman M., de Kool M., Sikora M., 1991, ApJ, 382, 416
Bender R., Saglia R. P., 1999, in Merritt D., Sellwood J. A., Valluri M., eds, ASP Conf. Ser. Vol. 182, Galaxy Dynamics – A Rutgers Symposium. Astron. Soc. Pac., San Francisco, p. 113
Benson A. J., Frenk C. S., Lacey C. G., Baugh C. M., Cole S., 2002, MNRAS, 333, 177
Benson A. J., Bower R. G., Frenk C. S., Lacey C. G., Baugh C. M., Cole S., 2003, ApJ, 599, 38
Bertschinger E., 1989, ApJ, 340, 666
Binney J., 1977, ApJ, 215, 483
Binney J., Tabor G., 1995, MNRAS, 276, 663
Binney J., Tremaine S., 1987, Galactic Dynamics. Princeton Univ. Press, Princeton, NJ, p. 747
Birnboim Y., Dekel A., 2003, MNRAS, 345, 349

Bode P., Ostriker J. P., Xu G., 2000, ApJS, 128, 561
Bondi H., 1952, MNRAS, 112, 195
Brüggen M., Kaiser C. R., 2002, Nat, 418, 301
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., 2001, ApJ, 555, 240
Burns J. O., Gregory S. A., Holman G. D., 1981, ApJ, 250, 450
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., Aragon-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
Cole S. et al., 2001, MNRAS, 326, 255
Cole S. et al., 2005, MNRAS, 362, 505
Colless M. et al., 2001, MNRAS, 328, 1039
Cowie L. L., Binney J., 1977, ApJ, 215, 723
Cowie L. L., Fabian A. C., Nulsen P. E. J., 1980, MNRAS, 191, 399
Cowie L. L., Songaila A., Hu E. M., Cohen J. G., 1996, AJ, 112, 839
Cox T. J., Primack J., Jonsson P., Somerville R. S., 2004, ApJ, 607, L87
Crenshaw D. M., Kraemer S. B., George I. M., 2003, ARA&A, 41, 117
Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, ApJ, 292, 371
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
Di Matteo T., Springel V., Hernquist L., 2005, Nat, 433, 604
Edgar R., 2004, New Astron. Rev., 48, 843
Efstathiou G., 1992, MNRAS, 256, 43P
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
Fan X. et al., 2001, AJ, 122, 2833
Forcada-Miró M. I., White S. D. M., 1997, astro-ph/9712204
Garnett D. R., 2002, ApJ, 581, 1019
Ghigna S., Moore B., Governato F., Lake G., Quinn T., Stadel J., 1998, MNRAS, 300, 146
Giovanelli R., Haynes M. P., da Costa L. N., Freudling W., Salzer J. J., Wegner G., 1997, ApJ, 477, L1
Gnedin N. Y., 2000, ApJ, 542, 535
Gonzalez R. E., Lares M., Lambas D. G., Valotto C., 2005, A&A submitted (astro-ph/0507144)
Häring N., Rix H., 2004, ApJ, 604, L89
Hartwick F. D. A., Schade D., 1990, ARA&A, 28, 437
Hatton S., Devriendt J. E. G., Ninin S., Bouchet F. R., Guiderdoni B., Vibert D., 2003, MNRAS, 343, 75
Heckman T. M., Armus L., Miley G. K., 1990, ApJS, 74, 833
Helly J. C., Cole S., Frenk C. S., Baugh C. M., Benson A., Lacey C., Pearce F. R., 2003, MNRAS, 338, 913
Hockney R. W., Eastwood J. W., 1981, Computer Simulation Using Particles. McGraw-Hill, New York
Hoeft M., Yepes G., Gottlober S., Springel V., 2005, MNRAS submitted (astro-ph/0501304)
Hopkins P. F. et al., 2005, ApJ submitted (astro-ph/0506398)
Huang J.-S., Glazebrook K., Cowie L. L., Tinney C., 2003, ApJ, 584, 203
Kaiser C. R., Binney J., 2003, MNRAS, 338, 837
Kauffmann G., 1996, MNRAS, 281, 475
Kauffmann G., Haehnelt M., 2000, MNRAS, 311, 576
Kauffmann G., Colberg J. M., Diaferio A., White S. D. M., 1999, MNRAS, 303, 188
Kennicutt R. C., 1998, ApJ, 498, 541
Keres D., Katz N., Weinberg D. H., Dave R., 2004, MNRAS, 363, 2
Kravtsov A. V., Gnedin O. Y., Klypin A. A., 2004, ApJ, 609, 482
Lacey C., Cole S., 1993, MNRAS, 262, 627
McGaugh S. S., Schombert J. M., Bothun G. D., de Blok W. J. G., 2000, ApJ, 533, L99
McNamara B. R. et al., 2000, ApJ, 534, L135
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
Madau P., Ferguson H. C., Dickinson M. E., Giavalisco M., Steidel C. C., Fruchter A., 1996, MNRAS, 283, 1388
Madgwick D. S. et al., 2002, MNRAS, 333, 133
Magorrian J. et al., 1998, AJ, 115, 2285
Mandelbaum R., McDonald P., Seljak U., Cen R., 2003, MNRAS, 344, 776
Marconi A., Hunt L. K., 2003, ApJ, 589, L21
Martin C. L., 1999, ApJ, 513, 156
Merloni A., 2004, MNRAS, 353, 1035
Mihos J. C., Hernquist L., 1994, ApJ, 425, L13
Mihos J. C., Hernquist L., 1996, ApJ, 464, 641
Mo H. J., Mao S., White S. D. M., 1998, MNRAS, 295, 319
Norberg P. et al., 2002, MNRAS, 336, 907
Nulsen P. E. J., 1986, MNRAS, 221, 377
Nulsen P. E. J., Fabian A. C., 2000, MNRAS, 311, 346
Olive K. A., Steigman G., Walker T. P., 2000, Phys. Rep., 333, 389
Omma H., Binney J., Bryan G., Slyz A., 2004, MNRAS, 348, 1105
Peacock J. A. et al., 2001, Nat, 410, 169
Percival W. J. et al., 2002, MNRAS, 337, 1068
Perlmutter S. et al., 1999, ApJ, 517, 565
Peterson J. R. et al., 2001, A&A, 365, L104
Popesso P., Biviano A., Böhringer H., Romaniello M., 2005, A&A, in press (astro-ph/0506201)
Rees M. J., Ostriker J. P., 1977, MNRAS, 179, 541
Reeves J. N., O'Brien P. T., Ward M. J., 2003, ApJ, 593, L65
Riess A. G. et al., 1998, AJ, 116, 1009
Ruszkowski M., Begelman M. C., 2002, ApJ, 581, 223
Sanders D. B., Mirabel I. F., 1996, ARA&A, 34, 749
Seljak U., Zaldarriaga M., 1996, ApJ, 469, 437
Seljak U. et al., 2005, Phys. Rev. D, 71, 103515
Shaver P. A., Wall J. V., Kellermann K. I., Jackson C. A., Hawkins M. R. S., 1996, Nat, 384, 439
Silk J., 1977, ApJ, 211, 638
Somerville R. S., Primack J. R., 1999, MNRAS, 310, 1087
Somerville R. S., Primack J. R., Faber S. M., 2001, MNRAS, 320, 504
Spergel D. N. et al., 2003, ApJS, 148, 175
Springel V., 2005, MNRAS submitted (astro-ph/0505010)
Springel V., Hernquist L., 2003a, MNRAS, 339, 289
Springel V., Hernquist L., 2003b, MNRAS, 339, 312
Springel V., White S. D. M., Tormen G., Kauffmann G., 2001a, MNRAS, 328, 726
Springel V., Yoshida N., White S. D. M., 2001b, New Astron., 6, 79
Springel V., Di Matteo T., Hernquist L., 2005a, MNRAS, 361, 776
Springel V. et al., 2005b, Nat, 435, 629
Tabor G., Binney J., 1993, MNRAS, 263, 323
Tamura T. et al., 2001, A&A, 365, L87
Tegmark M. et al., 2004, Phys. Rev. D, 69, 103501
Thoul A. A., Weinberg D. H., 1995, ApJ, 442, 480
Tremonti C. A. et al., 2004, ApJ, 613, 898
Tully R. B., Somerville R. S., Trentham N., Verheijen M. A. W., 2002, ApJ, 569, 573
van den Bergh S., 2000, The galaxies of the Local Group. Cambridge Univ. Press, Cambridge
Van Waerbeke L., Mellier Y., Pelló R., Pen U.-L., McCracken H. J., Jain B., 2002, A&A, 393, 369
White S. D. M., 1996, in Schaefer R., Silk J., Spiro M., Zinn-Justin J., eds, Cosmology and Large-Scale Structure. Elsevier, Dordrecht (astro-ph/9410043)
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
Xu G., 1995, ApJS, 98, 355
Yoshida N., Stoehr F., Springel V., White S. D. M., 2002, MNRAS, 335, 762
Yu Q., Tremaine S., 2002, MNRAS, 335, 965

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