Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **431,** 3373–3395 (2013)
Advance Access publication 2013 April 3

doi:10.1093/mnras/stt415

# Simulations of the galaxy population constrained by observations from $z = 3$ to the present day: implications for galactic winds and the fate of their ejecta

Bruno M. B. Henriques,[1]⋆ Simon D. M. White,[1] Peter A. Thomas,[2] Raul E. Angulo,[3] Qi Guo,[4] Gerard Lemson[1] and Volker Springel[5,6]

[1]*Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Str. 1, D-85741 Garching bei München, Germany*
[2]*Astronomy Centre, University of Sussex, Falmer, Brighton BN1 9QH, UK*
[3]*Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, Menlo Park, CA 94025, USA*
[4]*Partner Group of the Max-Planck-Institut für Astrophysik, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China*
[5]*Heidelberger Institut für Theoretische Studien, Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany*
[6]*Zentrum für Astronomie der Universität Heidelberg, ARI, Mönchhofstr. 12-14, D-69120 Heidelberg, Germany*

Accepted 2013 March 5. Received 2013 February 21; in original form 2012 December 10

## ABSTRACT

We apply Monte Carlo Markov Chain (MCMC) methods to large-scale semi-analytic simulations of galaxy formation in a Λ cold dark matter cosmology in order to explore how star formation and feedback are constrained by the observed luminosity and stellar mass functions of galaxies. We build models jointly on the Millennium and Millennium-II dark matter-only simulations, applying fast sampling techniques which allow observed galaxy abundances over the ranges $7 < \log M_\star/\mathrm{M}_\odot < 12$ and $0 \leq z \leq 3$ to be used simultaneously as constraints in the MCMC analysis. When $z = 0$ constraints alone are imposed, we reproduce the results of previous modelling by Guo et al., but no single set of parameter values can reproduce observed galaxy abundances at all redshifts simultaneously, reflecting the fact that low-mass galaxies form too early and thus are overabundant at high redshift in this model. The data require the efficiency with which galactic wind ejecta are reaccreted to vary with redshift and halo mass quite differently than previously assumed. We propose a specific model in which reincorporation time-scales vary inversely with halo mass and are independent of redshift. This produces an evolving galaxy population which fits observed abundances as a function of stellar mass, *B*- and *K*-band luminosity at all redshifts simultaneously. It also produces a significant improvement in two other areas where previous models were deficient. It leads to present day dwarf galaxy populations which are younger, bluer, more strongly star forming and more weakly clustered on small scales than before, although the passive fraction of faint dwarfs remains too high.

**Key words:** methods: analytical – methods: statistical – galaxies: evolution – galaxies: formation – galaxies: high-redshift.

## 1 INTRODUCTION

The semi-analytic modelling technique uses a set of parametrized, physically based equations to describe each of the astrophysical processes that affect galaxy formation and evolution (White 1989; Cole 1991; Lacey & Silk 1991; White & Frenk 1991). The development of the galaxy population as a whole is followed by applying these recipes to a set of merger trees representing the hierarchical assembly of non-linear structure in the dark matter distribution. Such halo merger trees can be constructed efficiently using extensions of the Press & Schechter (1974) formalism (e.g. Cole 1991; Kauffmann, White & Guiderdoni 1993; Cole et al. 1994; Somerville & Primack 1999) but if spatial and kinematic information is needed, for example, for analysis of galaxy clustering, it is more effective to extract them directly from numerical simulations (Roukema et al. 1997; Kauffmann et al. 1999; Hatton et al. 2003; Helly et al. 2003). With increasing numerical resolution it becomes possible to base the trees not on the haloes themselves but rather on the individual gravitationally self-bound subhaloes from which they are built (Springel et al. 2001, 2005; Kang et al. 2005; De Lucia et al. 2006). This increases the fidelity with which the galaxy distribution is

⋆ E-mail: bhenriques@mpa-garching.mpg.de

simulated, but introduces a dependence on numerical resolution which must be tested through careful convergence studies (e.g. Springel et al. 2001; Guo et al. 2011).

The great advantage of such semi-analytic simulations is their decoupling of the computation of dark matter evolution from that of the baryonic components. The physically consistent but relatively simple models adopted for complex and poorly understood baryonic processes allow a wide range of possible descriptions to be explored. The optimal parameter values for each can be determined in an acceptable amount of computer time, allowing the different modelling assumptions to be tested by detailed comparison with the relevant observational data. At present, semi-analytic methods are the only technique able to simulate the evolution of the galaxy population on a scale and with a precision which allows detailed interpretation of modern surveys of the galaxy population. The price, of course, is that they provide at best only very crude information about the internal structure of galaxies or the structure of the intergalactic medium.

The latest version of the Munich semi-analytic model (Guo et al. 2011, 2013, hereafter G11-*WMAP*1, G11-*WMAP*7, respectively) was tested and its parameter values adjusted by comparison with low-redshift galaxy data, namely, stellar mass and optical luminosity functions, gas fractions, colours, gas-phase metallicities, bulge-to-disc ratios, bulge and disc sizes and the black hole–bulge mass relation. Predictions were then tested by looking at satellite abundances, clustering, the evolution of the stellar mass function and the integrated star formation history. This analysis revealed a good match between model and data, particularly for the present-day universe, but some discrepancies remained. Most notable were an excess of lower mass galaxies at high redshift (also visible in the *K*-band luminosity function of Henriques et al. 2012), a substantial population of red dwarfs at low redshift, where very few such galaxies are observed, and excessive small-scale clustering of low-redshift dwarf galaxies. G11-*WMAP*7 showed that only the last of these problems is significantly alleviated if a *Wilkinson Microwave Anisotropy Probe* 7-year (*WMAP*7) cosmology is adopted instead of the *Wilkinson Microwave Anisotropy Probe* 1-year (*WMAP*1) cosmology of the original simulations. Similar problems are found in most recent semi-analytic galaxy formation models (Fontanot et al. 2009; Cirasuolo et al. 2010; Henriques et al. 2011; Bower, Benson & Crain 2012; Somerville et al. 2012; Weinmann et al. 2012), and can thus be viewed as generic.

In this paper we revisit the predictions of the G11-*WMAP*7 semi-analytic model using the Monte Carlo Markov Chain (MCMC) methods introduced by Henriques et al. (2009) and Henriques & Thomas (2010) (see Kampakoglou, Trotta & Silk 2008; Bower et al. 2010; Lu et al. 2011b, 2012 for related methods). We constrain parameters using observational estimates of the stellar mass function and of the *K*- and *B*-band luminosity functions at a variety of redshifts. The MCMC approach offers a systematic and objective means to identify the regions of parameter space consistent with the data and to obtain robust best-fitting estimates of parameter values with associated confidence ranges. This can yield new insight into how the data constrain galaxy formation physics.

We start by investigating the uniqueness of the choices of parameter values of G11-*WMAP*7, finding that they are indeed nearly optimal when representing low-redshift data. We then use the MCMC technique to identify the parameter regions preferred by the observational data at other redshifts, testing whether they overlap with each other and with the region preferred by the low-redshift data. Although almost all the G11-*WMAP*7 parameter values fall in the allowed region at all redshifts, the rate at which gas is reincorporated after ejection in a wind is required to change. Re-accretion must be less efficient than originally assumed at early times, and more efficient at late times. We use the MCMC scheme to identify a simple modification of the treatment of this process which can explain the data simultaneously at all redshifts. We then test this prescription by comparing its predictions with a broad set of observations including stellar metallicities, galaxy colours, star formation rates and ages, clustering and the evolution of the stellar mass–halo mass (SM–HM) relation.

This work uses merger trees from two very large dark matter simulations, which are well adapted to study galaxy formation physics on a variety of scales. The Millennium Simulation (MS; Springel et al. 2005) follows $10^{10}$ particles in a cube of side $500\,h^{-1}$ Mpc, implying a particle mass of $8.6 \times 10^8\,h^{-1}\,\mathrm{M}_\odot$, while the Millennium-II Simulation (MS-II; Boylan-Kolchin et al. 2009) uses the same number of particles in a region a fifth the linear size, resulting in 125 times better mass resolution. Combined, the two simulations produce a galaxy formation model with a dynamic range of five orders of magnitude in stellar mass. The distribution of physical properties converges for galaxies with $10^{9.5} < M_\star < 10^{11.5}\,\mathrm{M}_\odot$. Throughout the text we split our predictions by stellar mass using MS-II-based galaxies for $M_\star < 10^{9.5}\,\mathrm{M}_\odot$ and MS-based galaxies for higher masses.

The simulations can be scaled from the original *WMAP*1 cosmology to the currently preferred *WMAP*7 cosmology using the technique of Angulo & White (2010). Galaxy populations almost identical to those in the original G11-*WMAP*1 model can then be obtained through small adjustments to the semi-analytic modelling parameters (G11-*WMAP*7). This scaling changes the box-size and particle mass of the simulations, as well as shifting their time axis (see G11-*WMAP*7 for details). Below we work in the *WMAP*7 cosmology using the scaled merger trees.

The physical models used here for processes such as cooling, star formation and feedback have been developed gradually within the 'Munich' galaxy formation model over many years (White & Frenk 1991; Kauffmann et al. 1993, 1999; Springel et al. 2001, 2005). The current G11-*WMAP*1/G11-*WMAP*7 version includes a robust treatment of supernova (SN) feedback (following De Lucia, Kauffmann & White 2004 rather than Croton et al. 2006, see also Benson et al. 2003). It follows the growth of black holes through accretion and merging (Kauffmann & Haehnelt 2000) and the quenching of star formation by active galactic nucleus (AGN) feedback (Croton et al. 2006). It treats separately dust extinction from the interstellar medium (ISM) and from young birth clouds (De Lucia & Blaizot 2007) and it can predict luminosities over a wide wavelength range using two different stellar population synthesis models (see Henriques et al. 2011, 2012). G11-*WMAP*1/G11-*WMAP*7 changed several of the baryonic physics recipes in order to improve the treatment of dwarf and satellite galaxies. SN feedback was increased in low-mass galaxies, and environmental effects on satellites were treated more realistically. They also improved the tracking of angular momentum as material moves between the various gaseous and stellar components, allowing a better treatment of the sizes of discs and bulges.

This paper is organized as follows. In Section 2, we describe our implementation of the MCMC method. Section 3 applies this method to the G11-*WMAP*7 model, constraining parameters independently with data at each of a set of redshifts. No parameter set is consistent with the data at all redshifts. Guided by these MCMC results, we formulate in Section 4 a modification of the G11-*WMAP*7 model which can fit all redshifts simultaneously. The predictions of this new model are compared to the constraining observations

in Section 5, while Section 6 tests it against observations of additional properties that were not used as constraints. In Section 7, we summarize our conclusions. Appendix A sets out the equations that define the parameters explored using MCMC sampling, while Appendix B describes how we construct a representative sample of dark matter trees which allows us to obtain results rapidly for any specific semi-analytic model. Appendix C gives details of the observational data we use and of the uncertainties we assume for them, when using MCMC techniques to constrain model parameters.

A recent study by Mutch, Poole & Croton (2013) used MCMC techniques to explore similar issues using the model of Croton et al. (2006) applied to the MS in its original *WMAP*1 cosmology. Because they used stellar mass functions alone, and did not consider epochs earlier than $z = 0.83$, the evolutionary problem addressed here was only weakly present in their observational constraints. As a result they came to somewhat different conclusions than we do in this paper.

Throughout this work we use the Maraston (2005) stellar population synthesis model, but we have checked that for all the properties considered here the Charlot & Bruzual (private communication) code gives very similar results. As noted above, we assume a *WMAP*7 cosmology with specific parameters $h = 0.704$, $\Omega_{\rm m} = 0.272$, $\Omega_\Lambda = 0.728$, $n = 0.961$ and $\sigma_8 = 0.81$, using the MS and MS-II merger trees scaled exactly as in G11-*WMAP*7.

## 2 MONTE CARLO MARKOV CHAINS

Semi-analytic models simulate a large number of complex physical processes whose interplay shapes the distribution of galaxy properties. As a result, it is not straightforward to find the model that best fits a given set of observations, and to understand how unique a specific prescription or parameter set is. This becomes even more problematic when treatments are modified to allow interpretation of new kinds of observation, because such modifications often destroy the match to existing data. In addition, when reasonable agreement appears impossible to achieve, it is hard to know whether this reflects a failure to find the correct parameter set in a high-dimensional space, or overly simplistic modelling of one or more of the processes treated, or the omission of an important process or an issue with the parent dark matter simulation or the structure formation paradigm underlying it.

Many of these problems can be mitigated by combining constraints from multiple observations of a wide range of galaxy properties with proper sampling of the high-dimensional model parameter space. MCMC algorithms can be used to clarify how individual parameters influence specific galaxy properties, to obtain confidence limits for parameters, and to quantify the agreement between model and observation in a statistically robust way. They were first applied to the Munich galaxy formation model by Henriques et al. (2009) and Henriques & Thomas (2010). Such algorithms sample the allowed parameter regions at a rate proportional to the posterior probability of the model conditioned by the observational constraints, making them very efficient in analysing high-dimensional spaces. Section 3 of Henriques et al. (2009) gives a full description of the Metropolis–Hastings algorithm used here (Metropolis et al. 1953; Hastings 1970) which builds chains in which each step to a new set of parameters depends only on the current parameter set, not on the previous history. Appendix B gives some technical details of our implementation. Of particular relevance is our identification of a subset of merger trees which allows rapid convergence to results which are statistically representative of the full cosmological volume. With this approach, we find $\sim$1 per cent of the full data to be sufficient to reproduce the stellar mass function of the combined MS and MS-II to the accuracy we need, speeding up the MCMC calculations by about a factor of 100 and allowing us to constrain the model using the observed abundance of galaxies with stellar masses ranging from $10^7$ to $10^{12}\,{\rm M}_\odot$.

### 2.1 Observational constraints

A crucial aspect of semi-analytic modelling, and of MCMC analysis in particular, is the choice and characterization of the observational data used as constraints. The allowed regions in parameter space and the ability to find a good simultaneous fit to multiple data sets depend sensitively on the error bars assigned to the observations. Purely statistical errors are well defined and accurately stated in many observational studies, but it is much harder to account appropriately for residual systematic uncertainties. Unrecognized systematics can lead to apparent inconsistencies between different determinations of the same population property, jeopardizing a meaningful comparison with theoretical predictions.

Different approaches to this problem have been adopted in earlier work. For example, Bower et al. (2010) tested their model against a single data set to which they assigned a confidence region reflecting their own assessment of its systematic uncertainty. Henriques et al. (2009) and Henriques & Thomas (2010) use multiple 'good' determinations of each observational property, taking the scatter among them (together with the quoted statistical errors) to indicate likely systematic uncertainties. Here we adopt a version of the latter approach, noting that it still involves arbitrary judgements based on incomplete understanding, so that formal levels of agreement or disagreement between theory and observation should be treated with caution. In Section 5, theoretical models are compared with combined data sets where we have used the quoted purely statistical errors and the scatter among different 'good' measurements to make a subjective assessment of the effective uncertainties. In Appendix C we show the individual data sets and provide further details on how they were combined to provide constraints for each property at each redshift.

The allowed region of parameter space for a given model depends not only on the uncertainties assigned to each observation, but also, of course, on the particular observational quantities taken as constraints. This choice should recognize both the discriminative power and the robustness of the data. Use of complementary types of data can increase the aspects of the model which are significantly constrained. In the analysis of this paper we use the stellar mass function, and the *K*- and *B*-band luminosity functions at redshifts 3, 2, 1 and 0. The abundance of galaxies as a function of stellar mass is one of the most fundamental properties of the population, but it has the disadvantage that it is not directly observable, requiring assumptions about stellar populations and reddening. The *i*-band luminosity function would be a useful surrogate, since it is insensitive to dust and is unaffected by emission from stars on the poorly understood thermally pulsing asymptotic giant branch (TP-AGB), but few observational studies are available at redshifts other than zero, making it hard to assess the systematic uncertainties. We therefore use determinations of the stellar mass function itself. When needed, we follow Domínguez Sánchez et al. (2011) in applying a correction of $\Delta M_* = -0.14$ to go from Bruzual & Charlot (2003) to Maraston (2005) stellar populations at $z \geq 1$. As previously mentioned, the combined dynamical range of the MS and MS-II allows us to compare model and data directly, even for the lowest stellar masses observed at $z = 0$.

A critical test of galaxy formation models comes from comparison with the observed, redshift-dependent distributions of colour and star formation rate. We do not include direct observational estimates of these properties in our MCMC sampling, however, since colours are very sensitive to the details of dust modelling and observational estimates of star formation rates have large and redshift-dependent systematic uncertainties which are difficult to characterize due to the small number of observational studies available. Instead, we include colour information indirectly by supplementing our stellar mass function data with estimates of rest-frame *B*- and *K*-band luminosity functions. This provides some leverage on the star formation history, dust content, metallicity and TP-AGB emission of galaxies.

For each property, we combine determinations based on wide and narrow surveys in order to obtain constraints over a large dynamic range. The constraints used for the stellar mass, *K*- and *B*-band luminosity functions at different redshifts are shown in Figs 4, 5 and 6, respectively. Appendix C explains how these were obtained from various observational studies. The likelihood of the model given the adopted observational constraints on each individual property is defined as

$$\mathcal{L} \propto \exp\left(-\chi^2/2\right), \tag{1}$$

where $\chi^2$ is given by

$$\chi^2 = \sum_{i=0}^{N_\phi} \left( \frac{\phi_{\rm model}(i) - \phi_{\rm obs}(i)}{\sigma_{\rm obs}(i)} \right)^2, \tag{2}$$

where $\phi_{\rm model}$ and $\phi_{\rm obs}$ are, respectively, the logarithms of the number densities of galaxies in each bin as predicted by the model and as determined observationally. $\sigma_{\rm obs}(i)$ is our characterization of the observational uncertainty in $\phi_{\rm obs}(i)$ (see Appendix C). The final likelihood we assign to the model is then the product of the likelihoods for each galaxy property, either at a single redshift, or combining all redshifts:

$$\mathcal{L} = \prod_z \left(\mathcal{L}[{\rm SMF}(z)]\mathcal{L}[{\rm LF}_K(z)]\mathcal{L}[{\rm LF}_B(z)]\right) \propto \exp\left(-\chi_{\rm t}^2/2\right), \tag{3}$$

where $\chi_{\rm t}^2$ is the sum of the $\chi^2$ values for all the individual stellar mass and luminosity functions.[1] For a single redshift, this approach is a formal version of that used, for example, by G11-*WMAP*1/G11-*WMAP*7. When all redshifts are combined, the galaxy formation physics are constrained in a statistically robust way which takes full advantage of the fact that the semi-analytic model links galaxy populations self-consistently across cosmic time using a deterministic and physically based treatment of the evolution of individual galaxies.

### 2.2 Parameters

We allow a total of 11 parameters to vary when sampling the parameter space of the G11-*WMAP*7 model. These produce the most significant differences in the distributions of stellar masses and luminosities used as constraints in our study: the star formation efficiency (SFE; $\alpha_{\rm SF}$); the black hole growth efficiency ($f_{\rm BH}$); the AGN radio mode efficiency ($k_{\rm AGN}$); three parameters governing the reheating and injection of cold disc gas into the hot halo phase by SNe, the gas reheating efficiency ($\epsilon$), the reheating cut-off velocity ($V_{\rm reheat}$) and the slope of the reheating dependence on $V_{\rm vir}$ ($\beta_1$); three parameters governing the ejection of hot halo gas to an external reservoir, the gas ejection efficiency ($\eta$), the ejection cut-off velocity ($V_{\rm eject}$) and the slope of the ejection dependence on $V_{\rm vir}$ ($\beta_2$); the efficiency of reincorporating gas from the external reservoir to the hot halo ($\gamma$) or the gas return time in our new reincorporation model ($\gamma'$) and the yield of metals returned to the gas phase by stars ($z_{\rm yield}$). The equations describing these processes and defining the parameters are described concisely in Appendix A. The very fact that such a high-dimensional parameter space can be analysed at all in a reasonable amount of computer time emphasizes a key strength of the MCMC method. Using naive techniques would require computing $\sim 10^{11}$ models to survey the space with just 10 points along each parameter dimension. In contrast, we achieve convergence for chains with as few as 30 000 steps, although we always run at least 100 000 steps and discard the first 10 per cent as a 'burn in' phase.

The semi-analytic model contains additional unknown parameters not included in our MCMC analysis. In general these have a weaker impact on the galaxy properties considered and are left fixed at the G11-*WMAP*7 values. The parameters not sampled are either directly derived from simulations of specific galaxy processes, such as the merger starburst efficiencies (A4), the dynamical friction efficiency during mergers and the redshift of beginning and end of reionization; are derived from stellar evolution theory, such as the total energy released by SN and the amount of stellar material returned to the ISM per star formation episode; or have been adjusted over the years as a result of a comparison to a wider range of observational properties, such as the major merger threshold and the cold-gas-to-dust ratio.

## 3 TESTING THE G11-*WMAP*7 MODEL

We begin by using our MCMC method to explore the regions of parameter space for which the G11-*WMAP*7 galaxy formation model is consistent with the observational constraints at each redshift, *independent* of the constraints at the other redshifts. It is interesting that in each case the high-likelihood region is quite localized in the full 11-dimensional space. The observational data at a single redshift are already sufficient to constrain all the parameters without major degeneracies. Furthermore, at $z = 0$ the high-likelihood region includes the parameter set chosen in the earlier paper, showing that the less formal tuning procedure adopted there did, in fact, result in near-optimal parameter values. Comparing preferred parameter space regions at different redshifts we find very substantial overlap. Only one of the 11 parameters, the efficiency $\gamma$ with which ejected gas from the external reservoir is reincorporated into the hot halo, is required to take substantially different values at different times.

This can be seen clearly in Fig. 1 which shows the posterior likelihood distributions at the four redshifts marginalized into four different two-parameter subspaces: SFE and gas reincorporation efficiency in the upper left-hand panels; black hole growth efficiency and AGN radio mode efficiency in the upper right-hand panels; gas reheating efficiency and gas ejection efficiency in the lower left-hand panels and SFE and metal yield in the lower right-hand panels. In each panel, equal likelihood contours containing 68 and 95 per cent of the marginalized posterior distribution are shown as thick magenta lines, while colours indicate the maximum likelihood over all the MCMC samples in each pixel. In fact, the 95 per cent regions for 10 of the 11 parameter values overlap at least marginally across all four redshifts, whereas the reincorporation efficiency ($\gamma$) is

[1] Note that we neglect correlations in the scatter between different data points and different data sets, even though these undoubtedly exist and may be important.

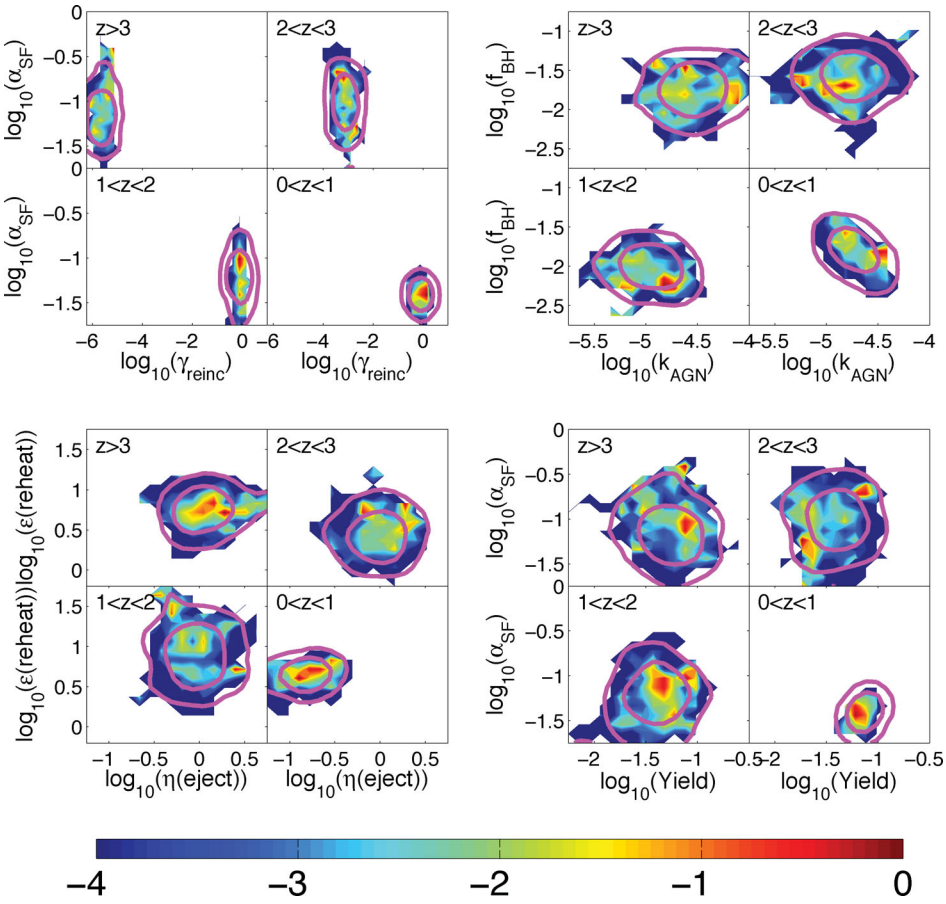



**Figure 1.** Allowed likelihood regions for various semi-analytic parameter pairs at four different redshifts. SFE ($\alpha_{\rm SF}$) and gas reincorporation efficiency ($\gamma$) are shown in the four upper left-hand panels; black hole growth efficiency ($f_{\rm BH}$) and AGN radio mode efficiency ($k_{\rm AGN}$) in the four upper right-hand panels; gas reheating efficiency ($\epsilon$) and gas ejection efficiency ($\eta$) in the four lower left-hand panels; SFE ($\alpha_{\rm SF}$) and metal yield ($z_{\rm yield}$) in the four lower right-hand panels. The G11-*WMAP*7 model is constrained independently at each of the four redshifts using observational estimates of the stellar mass function and the *K*- and *B*-band luminosity functions. Each (logarithmic) parameter space is divided into equal-area pixels and 2D histograms made from our MCMC sampling. Solid magenta lines then show equal likelihood contours containing 68 and 95 per cent of all samples (e.g. of the marginalized 2D posterior distribution). Colours indicate the maximum likelihood value in each bin (the profile distribution) normalized by the global maximum likelihood. The colour scale is labelled by $\log_{10}L/L_{\rm max}$.

required to evolve strongly. The observations indicate that very little ejected gas is reincorporated at high redshift, but that reincorporation is substantially more effective at $z \leq 2$. This appears required within the G11-*WMAP*1/G11-*WMAP*7 representation of galaxy formation physics in order to obtain the observed low abundances of lower mass galaxies at $z \geq 2$ while maintaining the higher abundances observed for similar mass galaxies at lower redshift. There is some indication that the SN ejection efficiency $\eta$ should be lower at $z \sim 0$ than at higher redshift (*x*-axis on the bottom left-hand panels) but the variation required is much smaller than the four orders of magnitude needed for $\gamma$ and a reasonable compromise value can be found on the edge of the 95 per cent regions ($\log \eta \sim -0.3$).

It is worth stressing the discerning power of this procedure. It indicates that most of the physical assumptions made in the G11-*WMAP*1/G11-*WMAP*7 model are consistent with the observed evolution of the galaxy population over $3 > z > 0$. In addition, it clearly indicates which of the assumptions need to be modified and how they should be corrected. The apparent lack of degeneracies also suggests that, for most of the efficiency parameters, evolutionary trends other than those currently assumed are disfavoured by the observational constraints we have adopted. For example, a SFE that increases with decreasing redshift, as proposed by Wang, Weinmann & Neistein (2012) to match the evolution of the number density of dwarfs, is not consistent with the data in the context of *our* model. Indeed, in this model SFE primarily affects the gas-to-star ratios of dwarfs rather than their stellar masses or star formation rates. As in some hydrodynamical simulations (Oppenheimer & Davé 2008; Haas et al. 2012), gas simply accumulates in a dwarf until it is sufficient to fuel star formation and ejection rates which balance the gas supply through infall. Wang et al. (2012), as well as previous work by Neistein & Weinmann (2010), fit the observed evolution of the stellar mass function by implementing cooling rates and star formation rates that scale with time as well as halo mass, although they give no specific physical motivation for it. Changing the scaling of wind properties alone is also unable to produce the required evolution, as recently found also by Bower et al. (2012).

## 4 A NEW MODEL FOR REINCORPORATION OF EJECTED GAS

The G11-*WMAP*1/G11-*WMAP*7 version of the Munich semi-analytic model assumed the gas ejected from haloes to be stored in an external reservoir and to be reincorporated into the system on

a time-scale which depends on halo mass and redshift. The particular model adopted was

$$\dot{M}_{\rm ejec} = -\frac{M_{\rm ejec}}{t_{\rm reinc}}, \tag{4}$$

with

$$t_{\rm reinc} = \frac{1}{\gamma} t_{\rm dyn,h} \left( \frac{V_{\rm vir}}{220\,{\rm km\,s^{-1}}} \right)^{-1}, \tag{5}$$

where $M_{\rm ejec}$ is the current amount of gas in the reservoir, $t_{\rm dyn,\,h} = 0.1H(z)^{-1}$ is the dynamical time of the halo, $V_{\rm vir}$ is its virial velocity and $\gamma$ is a free parameter. The $V_{\rm vir}$ dependence was introduced by G11-*WMAP*1 to model the effect of higher wind velocities (relative to escape velocity) in lower mass systems. The scaling with $t_{\rm dyn,\,h}$ imposes a redshift dependence similar to that of the dark matter growth time of the haloes, which is nearly independent of mass but decreases quite strongly at earlier times (e.g. Guo & White 2008). Our MCMC analysis shows, however, that the data require $\gamma$ to be much smaller (and thus reincorporation to be much less effective) at $z \geq 2$ than at lower redshifts. In order to find a description of the reincorporation process that can reproduce the data at all times, we re-write equation (5) as

$$t_{\rm reinc} = \frac{1}{\gamma}(1+z)^{\delta_1} V_{\rm vir}^{-\delta_2} t_{\rm dyn,h}, \tag{6}$$

and we re-run our MCMC chains constrained simultaneously by the observations at all redshifts and including $\delta_1$ and $\delta_2$ as free parameters.

The results of this procedure are shown in Fig. 2. After marginalizing over other parameters the high likelihood region is quite localized in the $\delta_1$–$\delta_2$ plane. The maximum likelihood values of these parameters and their $\pm 1\,\sigma$ range are $\delta_1 = 2.40^{+0.21}_{-0.33}$ and $\delta_2 = 3.07^{+0.26}_{-0.26}$. All other parameter values remain in their plausible ranges for this maximum likelihood model. If we adopt $\delta_1 = 2.4$ and $\delta_2 = 3$ we find, using $M_{\rm vir} \propto V_{\rm vir}^3/H(z)$ and $t_{\rm dyn,\,h} \propto H(z)^{-1}$, that

$$t_{\rm reinc} \propto \frac{1}{M_{\rm vir}} \frac{(1+z)^{2.4}}{\Omega_{\rm m}(1+z)^3 + (1-\Omega_{\rm m})}, \tag{7}$$

where $\Omega_{\rm m} \sim 0.25$ is the present-day matter density. The redshift-dependent ratio on the right-hand side of this equation varies by less than a factor of 2 for $0 \leq z \leq 6$ so we neglect it and adopt a simpler reincorporation model in which the relevant time-scale is inversely proportional to halo mass and independent of redshift:

$$t_{\rm reinc} = -\gamma' \frac{10^{10}\,{\rm M_\odot}}{M_{\rm vir}}, \tag{8}$$

where the constant $\gamma'$ now has units of time. Using our MCMC chains to identify the best fit of this new model to our constraining data at all four redshifts, we find $\chi^2_{\rm t} = 127$ for 134 degrees of freedom and 11 adjusted parameters. For comparison, applying the same procedure to the G11-*WMAP*7 model yields $\chi^2_{\rm t} = 475$ for the same number of adjusted parameters. Thus our modified assumptions about reaccretion of ejected gas substantially improve the ability of the model to represent the observational data.

The best-fitting value of the reincorporation time-scale, $\gamma' = 1.8 \times 10^{10}$ yr, implies that haloes with $M_{\rm vir} \geq 10^{11.5}\,{\rm M_\odot}$ have very short reincorporation times, hence are able to eject very little gas at any relevant redshift. In contrast, very little of the gas ejected from haloes with $M_{\rm vir} \leq 10^{10}\,{\rm M_\odot}$ ever returns. For a fixed halo mass, ejection is most effective at high redshift because at given halo mass, cooling, star formation and feedback are all typically stronger at earlier times. We note that higher halo masses at higher redshift (see Fig. 10) result in galaxies of given stellar mass having gas return times which are shorter at early times than at $z = 0$. However, the redshift dependence is now much weaker than in G11-*WMAP*1/G11-*WMAP*7, and the reincorporation time-scales in low-mass galaxies are longer, ensuring that gas is predominantly returned at low redshift. The changes are illustrated in the right-hand panel of Fig. 3; at $z = 3$ the reincorporation time is longer in the new model than in G11-*WMAP*7 for all haloes with $V_{\rm vir} < 140\,{\rm km\,s^{-1}}$, whereas at $z = 0$ this is only true for $V_{\rm vir} < 30\,{\rm km\,s^{-1}}$.

Our proposed dependence of the gas return time-scale on halo mass can also be compared with that found in recent direct numerical simulations of galaxy formation. Oppenheimer & Davé (2008) found that the gas ejected in winds is affected not only by gravitational forces, but also by ram pressure against the circumgalactic medium. As a result, material is less likely to escape in denser environments. These hydrodynamical simulations suggest that the time for ejected material to return to a galaxy scales inversely with the total mass of the system rather than with the local dynamical time, a very similar result to that we find here. More specifically, Oppenheimer et al. (2010) found return times to scale with $M_{\rm vir}^{-1.5}$ for constant wind speed and with $M_{\rm vir}^{-0.6}$ for wind speeds proportional to the virial velocity of the system. The latter is more similar to our feedback implementation (see Appendix A for details). Indeed, we obtain a compatible scaling of the return times in our new reincorporation model. Our preferred value for $\gamma'$ ($1.8 \times 10^{10}$ yr) results in a reincorporation time that varies from $1.8 \times 10^{10}$ yr for haloes with $M_{\rm vir} = 10^{10}\,{\rm M_\odot}$ to $1.8 \times 10^{8}$ yr for haloes with $M_{\rm vir} = 10^{12.}\,{\rm M_\odot}$. These correspond to the time-scales for ~10 per cent of the gas to be reincorporated in the simulation of Oppenheimer et al. (2010). As expected, our time-scales are shorter since they correspond to the

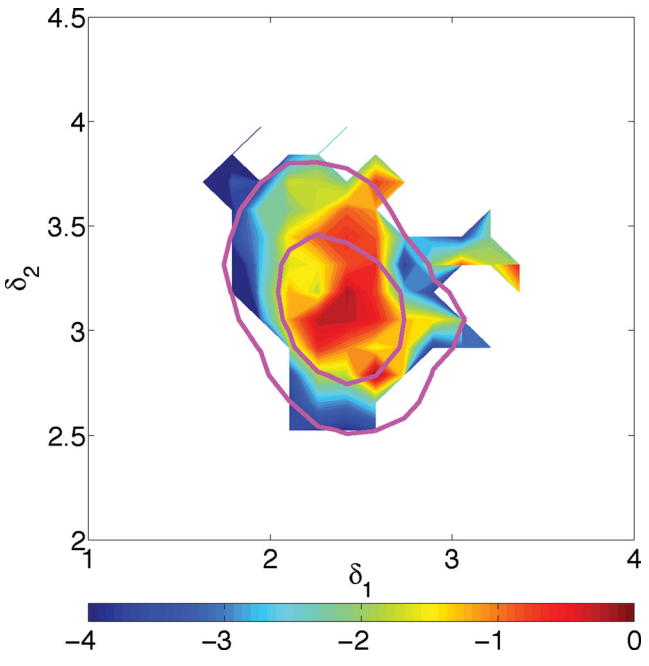



**Figure 2.** High likelihood region for the $\delta_1$ and $\delta_2$ parameter values in the reincorporation model described by equation (6). The model is constrained using the observed stellar mass functions and the *K*- and *B*-band luminosity functions at $z = 0$, 1, 2 and 3 simultaneously. The solid lines are equal likelihood contours containing 68 and 95 per cent of the marginalized 2D posterior distribution, while colours indicate the maximum likelihood value over all MCMC parameter samples projected within each pixel. The logarithmic colour scale is normalized to the global maximum likelihood value.

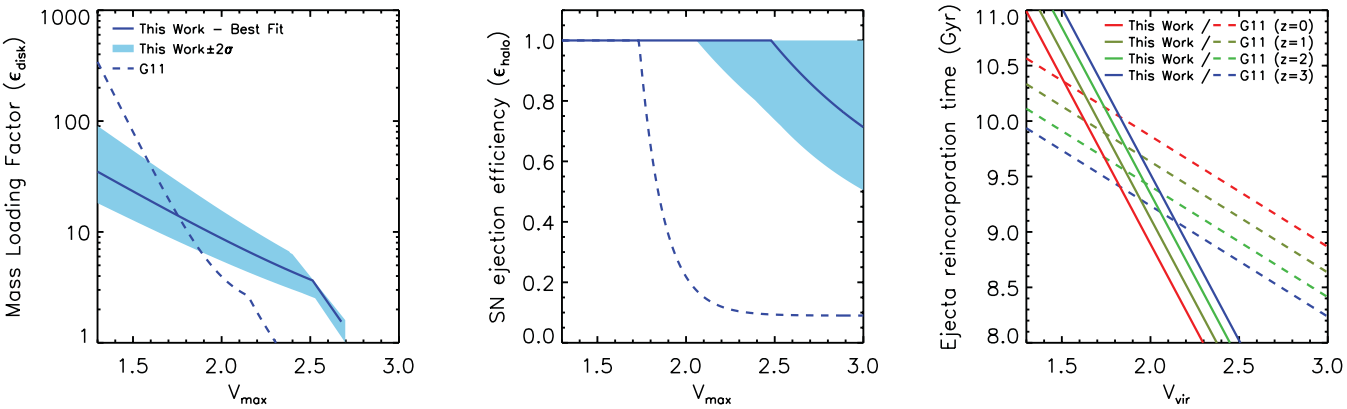



**Figure 3.** Illustration of the major changes in the dependence of feedback on galaxy properties between the G11-*WMAP*7 model and the model of this paper. The left-hand panel shows the disc reheating efficiency $\epsilon_{\rm disc}$ as a function of maximum circular velocity $V_{\rm max}$. Often referred to as the mass-loading factor, this is the ratio of the star formation rate to the rate at which ISM material is heated and injected into the hot halo. The middle panel shows the halo ejection efficiency $\epsilon_{\rm halo}$ as a function of $V_{\rm max}$. This is the fraction of the available SN energy which is used in reheating disc gas and in ejecting hot gas from the halo. The right-hand panel shows the reincorporation time-scale $t_{\rm reinc}$ as a function of halo virial velocity $V_{\rm vir}$ and of redshift. In each panel dashed lines refer to the G11-*WMAP*7 model and solid lines to our new model with its best-fitting parameter values. The blue shaded regions in the left two panels give the $2\sigma$ range allowed by our MCMC sampling. Colours in the right-hand panel indicate redshift as shown by the label.

time for the gas to return from the ejected to the hot phase, whereas in the hydrodynamical simulations they correspond to the time for a full cycle, beginning when the material leaves the star-forming region and ending when it is reincorporated into the ISM of the galaxy. As a caveat, it should be noted that the Oppenheimer & Davé (2008) simulations are not fully hydrodynamical as the hydro forces are turned off for interactions between winds and gas with density greater than 0.01 particles per cm$^3$. This density regime includes a large fraction of the circumgalactic medium, particularly at high redshift, for which the reincorporation times might closer resemble a semi-analytic calculation.

In passing we note that our new version of the semi-analytic code also includes a number of technical changes mostly related to book-keeping. These are designed to assure accurate mass and metal conservation at all times, and improved memory management at run-time. These changes can have an impact on a few individual galaxies but they leave the global properties of the population unchanged. We have also introduced a minor modification to the physics of infall. In previous versions of the model, if the $M_{\rm vir}$ value of a friends-of-friends (FoF) halo drops during its evolution, then hot gas (including metals) was removed so that the total baryon content remains equal to the cosmic baryon fraction times $M_{\rm vir}$.[2] We relax this assumption here and allow each FoF halo to retain all its associated baryons even if this puts its baryon fraction temporarily above the cosmic value. Its baryon content then begins to increase again only after its $M_{\rm vir}$ rises above the previous maximum value. This change simplifies the treatment of chemical enrichment, allowing detailed conservation of the mass of every element as material shifts between the different baryonic components in each FoF halo. It also has very little impact on global galaxy properties.

### 4.1 Best-fitting parameters

As noted above, with the new reincorporation model of equation (8) it is possible to find parameter values that produce acceptable agreement with the observational constraints at all the redshifts we consider. The resulting best-fitting parameters values are listed in Table 1, together with their 95 per cent confidence ranges. They are compared to the parameters values preferred by G11-*WMAP*7 for a fit of the previous model to the $z = 0$ data alone, assuming a *WMAP*7 cosmology. All the original parameter values are outside the $2\sigma$ regions for the new best fit. The most substantial differences occur for the SFE, which is considerably higher in the new model, and for the SN feedback parameters, with high reheating and ejection efficiencies in the new model extending to higher mass systems. These changes combine with the new reincorporation scheme to ensure that the number density of galaxies below $M_\star$ is strongly suppressed at early times, but grows through the return of ejected gas at later times so that the observed abundance of these galaxies at $z = 0$ is still reproduced. This later formation increases the low-redshift star formation rates of low-mass galaxies, making their colours bluer. In Fig. 3 we present a graphical comparison of the changes in the relevant efficiencies between G11-*WMAP*7 and this work.

In our new model the mass loading of the winds which eject gas from galaxy discs into their hot haloes depends less strongly on the rotation velocity of the discs and is larger for all but the smallest discs than in the G11-*WMAP*7 model. (This is $\epsilon_{\rm disc}$ in equation A5; its scaling with $V_{\rm max}$ is illustrated for the old and new models in the left-hand panel of Fig. 3.)

As can be seen from the middle panel of Fig. 3, the values now required for the parameters in equation (A8) imply that all the available energy from SNe is used to reheat and eject gas in typical disc galaxies (i.e. $\epsilon_{\rm halo} = 1$ for $V_{\rm max} < 300\,{\rm km\,s^{-1}}$). This efficiency might be unsatisfactorily high, but we note that within our $2\sigma$ allowed region $\epsilon_{\rm halo}$ can be as low as 60 per cent for galaxies like the Milky Way, and in any case the total energy available to drive winds could exceed that we have assumed by a factor of 2 or more. For example, we have considered only SN feedback, neglecting input due to radiation and winds from massive stars.

Finally, as already mentioned briefly above, the third panel of Fig. 3 shows that the reincorporation times for gas ejected from a halo depend more strongly on halo virial velocity $V_{\rm vir}$ than in G11-*WMAP*1/G11-*WMAP*7, and that at fixed $V_{\rm vir}$ the dependence on redshift is reversed. Thus at $z = 0$ reincorporation is faster in

[2] The $M_{\rm vir}$ values of FoF haloes generally increase with time but can drop temporarily through numerical fluctuations or if they pass close to a larger halo.

**Table 1.** Statistics from the MCMC parameter estimation. The best-fitting values of parameters and their '$2\sigma$' confidence limits are compared with the values published in G11-*WMAP*7 for their *WMAP*7 model constrained by $z = 0$ data alone.

| | G11-*WMAP*7 | New model | $2\sigma$ lower limit | $2\sigma$ upper limit |
|---|---|---|---|---|
| $\alpha_{\rm SF}$ (SF efficiency) | 0.01 | 0.055 | 0.027 | 0.060 |
| $k_{\rm AGN}$ (radio feedback efficiency) | $7.0 \times 10^{-6}$ | $3.2 \times 10^{-5}$ | $2.7 \times 10^{-5}$ | $4.1 \times 10^{-5}$ |
| $f_{\rm BH}$ (BH growth efficiency) | 0.03 | 0.015 | 0.011 | 0.017 |
| $\epsilon$ (mass-loading efficiency) | 4.0 | 2.1 | 1.8 | 2.6 |
| $V_{\rm reheat}$ (mass-loading scale) | 80 | 405 | 315 | 473 |
| $\beta_1$ (mass-loading slope) | 3.2 | 0.92 | 0.82 | 1.14 |
| $\eta$ (SN heating efficiency) | 0.2 | 0.65 | 0.52 | 0.87 |
| $V_{\rm eject}$ (SN heating scale) | 90 | 336 | 263 | 430 |
| $\beta_2$ (SN heating slope) | 3.2 | 0.46 | 0.38 | 0.59 |
| $\gamma'$ (ejecta reincorporation) | Not applicable | $1.8 \times 10^{10}$ yr | $1.8 \times 10^{10}$ yr | $1.9 \times 10^{10}$ yr |
| $Z_{\rm yield}$ (metals yield) | 0.03 | 0.047 | 0.040 | 0.051 |

the new model than in G11-*WMAP*7 for all galaxies with $V_{\rm vir} > 30\,{\rm km\,s^{-1}}$, whereas at $z = 3$ this is only true for $V_{\rm vir} > 150\,{\rm km\,s^{-1}}$ – reincorporation is substantially slower at virial velocities typical of dwarf galaxies.

The increase in SFE relative to the model of G11-*WMAP*1/G11-*WMAP*7 primarily affects intermediate-mass galaxies since objects of lower masses have very short cooling times and star formation rates are driven primarily by the infall of new gas and the reincorporation of ejecta. At high redshift, the increased SFE compensates for the more efficient SN feedback to maintain the abundance of $M_\star$ objects. At lower redshift, the effect is neutralized by the higher AGN radio mode efficiency in the new model, so that the high mass cut-off in the mass function is unchanged. The SFE parameter $\alpha_{\rm SF}$ regulates the conversion of cold gas into stars, so it is important to note that, despite the new efficiency being $\sim$3 times higher than in G11-*WMAP*1 and $\sim$5 times higher than in G11-*WMAP*7, the predicted cold gas fractions of low-redshift spiral and irregular galaxies are quite comparable. The higher conversion rate from gas into stars is balanced by the increased amount of gas available due to the delayed reincorporation of ejected gas. Finally, the new best fit requires a higher value for the metal yield than in G11-*WMAP*1/G11-*WMAP*7. This parameter is primarily constrained by the relative abundances of bright galaxies in the *K* and *B* bands, i.e. by the colours of massive elliptical galaxies.

## 5 CONSISTENCY WITH THE INPUT CONSTRAINTS

In this section we compare the theoretical predictions of our new model with the observational properties used as constraints in our MCMC analysis: the stellar mass function, and the *K*- and *B*-band luminosity functions from $z = 3$ to $z = 0$. All magnitudes are AB, the stellar populations are Maraston (2005) and the *WMAP*7 cosmology is adopted. For all properties we also compare with the predictions of the G11-*WMAP*7 model tuned to the $z = 0$ data alone with re-calculated luminosities for Maraston (2005) stellar populations. All the results presented in this section are recovered using Charlot & Bruzual (private communication) stellar evolution models while Bruzual & Charlot (2003) would result in a lack of bright *K*-band galaxies at $z \geq 2$ (leaving the remaining properties unchanged).

### 5.1 Stellar mass functions

G11-*WMAP*1 found good agreement between the stellar mass function of their model and the $z = 0$ observations that they used to tune its parameters. However, an increasing excess of low-mass objects was found when model predictions were compared to data at higher redshifts. This excess was not a consequence of the high value of $\sigma_8$ adopted for the original MS, since it remained equally strong when the simulations were scaled to a *WMAP*7 cosmology by G11-*WMAP*7.

The observed number densities of massive and infrared-bright galaxies are reproduced equally well in both cosmologies, once allowance is made for the substantial random errors expected in the stellar masses of high-redshift objects (Henriques et al. 2012). Within the G11-*WMAP*1/G11-*WMAP*7 model it appears impossible to retain this agreement, while fixing the abundance evolution at lower mass. Here we test how well our new model for the reincorporation of ejected gas addresses this issue. Note that we are attempting to build an a priori physical simulation which reproduces the observed abundances of galaxies over five orders of magnitude in stellar mass at $z = 0$ and throughout the last five sixths of the age of the Universe over two orders of magnitude in stellar mass.

In Fig. 4 we show the evolution of the stellar mass function from $z = 3$ (lower right-hand panel) to $z = 0$ (top left-hand panel). Results from G11-*WMAP*7 and from our new model with its best-fitting parameter values are shown as dashed and solid red lines, respectively. The dotted red line represents the new model before convolution with a log-normal distribution of width 0.25 dex representing random uncertainties in the observed stellar mass estimates. The shift between dotted and solid red lines thus shows the predicted effect of Eddington bias on the observed distribution. This effect is not included at $z = 0$ since the observational errors should be considerably smaller, particularly those arising from the uncertainties in the stellar emission from young populations, photometry and redshift determinations. The models are compared with the observational constraints adopted for our MCMC analysis, which are derived from a comprehensive set of high-quality data. The individual data sets are shown in Appendix C.

The effect of modifying the treatment of gas reincorporation time is clearly visible in Fig. 4. The delayed return of gas ejected in winds reduces the amount of gas available for star formation at early times. The effect is pronounced in intermediate- and low-mass galaxies ($8 < \log M_\star/{\rm M}_\odot < 11$), leading to a reduction in

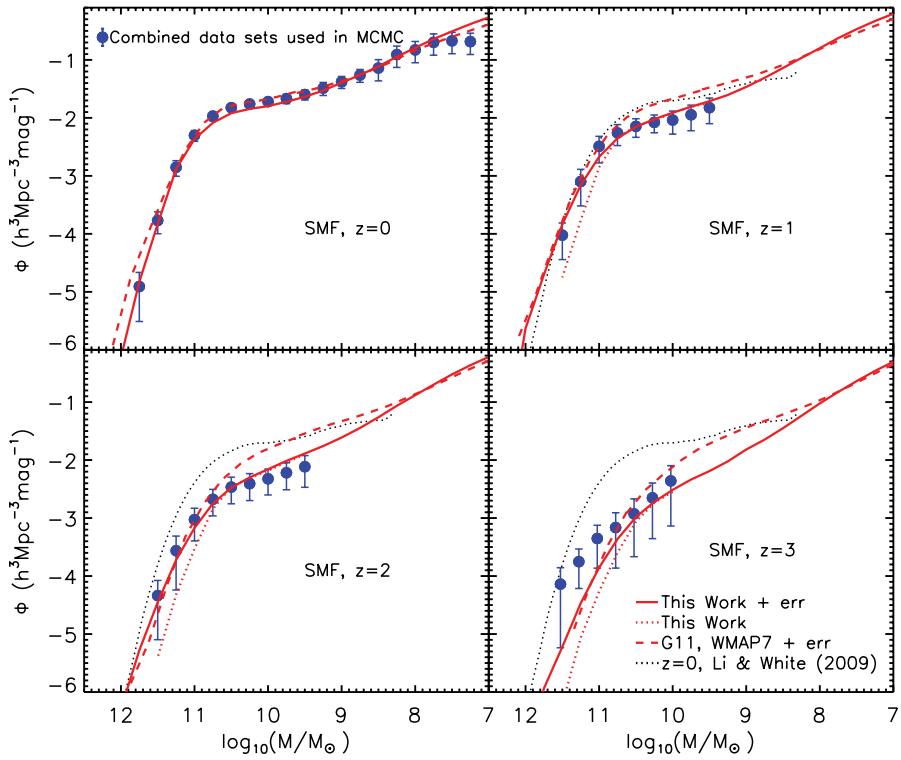



**Figure 4.** Evolution of the stellar mass function from $z = 3$ to $z = 0$. Theoretical predictions from our new model and from G11-*WMAP*7 are shown as solid and dashed red lines, respectively. The dotted red lines in the $z \geq 1$ panels are for the new model before convolution with a log-normal distribution of width 0.25 dex representing random uncertainties in the observed stellar masses. Blue dots with error bars show the observational constraints and their $1\sigma$ uncertainties as adopted for the MCMC analysis. These result from combining a variety of data sets at each redshift. The individual data points are shown in Fig. C1. The $z = 0$ results of Li & White (2009) are repeated at all redshifts as a black dotted line.

the number density at given stellar mass by almost a factor of 2 at $z \geq 2$. The enhanced reincorporation of ejected material at later times provides additional fuel, allowing low-mass galaxies to form enough stars to continue to match the observed abundances at $z = 0$.

The model also matches the moderate but still noticeable evolution of the abundance of massive galaxies. This agreement contrasts with earlier work where the evolution in the observed abundances of high-mass galaxies appeared much weaker than predicted by hierarchical models (e.g. Caputi et al. 2006; Fontana et al. 2006; Marchesini et al. 2009). The discrepancy can be explained by the combination of two effects. For high-redshift galaxies, photometric redshifts and stellar mass estimates obtained by SED fitting have substantial uncertainties which extend the high-mass tail of the distribution of *estimated* stellar masses. In addition, if masses are derived using population synthesis models that omit the near-infrared emission expected from intermediate-age stars, the fluxes of high-redshift galaxies will be interpreted as indicating overly high masses. Once these two effects are taken into account, both in correcting the data and in tuning the model, quite reasonable agreement is obtained.

### 5.2 The rest-frame *K*-band luminosity function

Most previous galaxy formation models have failed to match the observed evolution of the faint end of the *K*-band luminosity function (Kitzbichler & White 2007; Cirasuolo et al. 2010; Henriques et al. 2011, 2012; Somerville et al. 2012). This is a direct consequence of the overly early build up of the dwarf galaxy population seen in the stellar mass function. At the other extreme, model predictions for the evolution in abundance of the brightest *K*-band galaxies have only quite recently been reconciled with observations (Bower, McCarthy & Benson 2008; Fontanot & Monaco 2010; Tonini et al. 2010; Henriques et al. 2011, 2012; Somerville et al. 2012).

In Fig. 5, we compare the predicted evolution of the *K*-band luminosity function out to redshift $z = 3$ with the observational data we used as constraints for our MCMC chains. Results from our new model and from G11-*WMAP*7 are shown as solid and dashed red lines, respectively, in both cases assuming Maraston (2005) stellar populations. The observational constraints are based on combining results from a number of studies, which are shown individually in Appendix C. Once the new reincorporation prescription is included, our model agrees reasonably well with the observational data over the full luminosity and redshift ranges considered. The reduced number density of dwarfs at high redshift caused by the delayed reincorporation of gas has a direct impact on the abundances of faint near-infrared galaxies. This corrects the previous excess of faint objects while maintaining agreement at the bright end.

The most obvious remaining discrepancy is a residual overabundance of faint galaxies at $z = 2$. Within our present model this cannot be corrected without worsening the good fit to the stellar mass and *B*-band luminosity functions of the corresponding $z = 2$ galaxies and their $z = 3$ progenitors. It is plausible that the problem stems from residual systematic errors in one or more of these observational data sets, but it could also reflect a problem in the stellar population modelling of 1 Gyr old populations.

Despite the considerable uncertainties that still remain in theoretical predictions, it is interesting to note that the evolution of the bright tail of the near-infrared luminosity function of galaxies can be matched by a model that simultaneously reproduces the observed evolution of the high-mass tail of the stellar mass function. Until

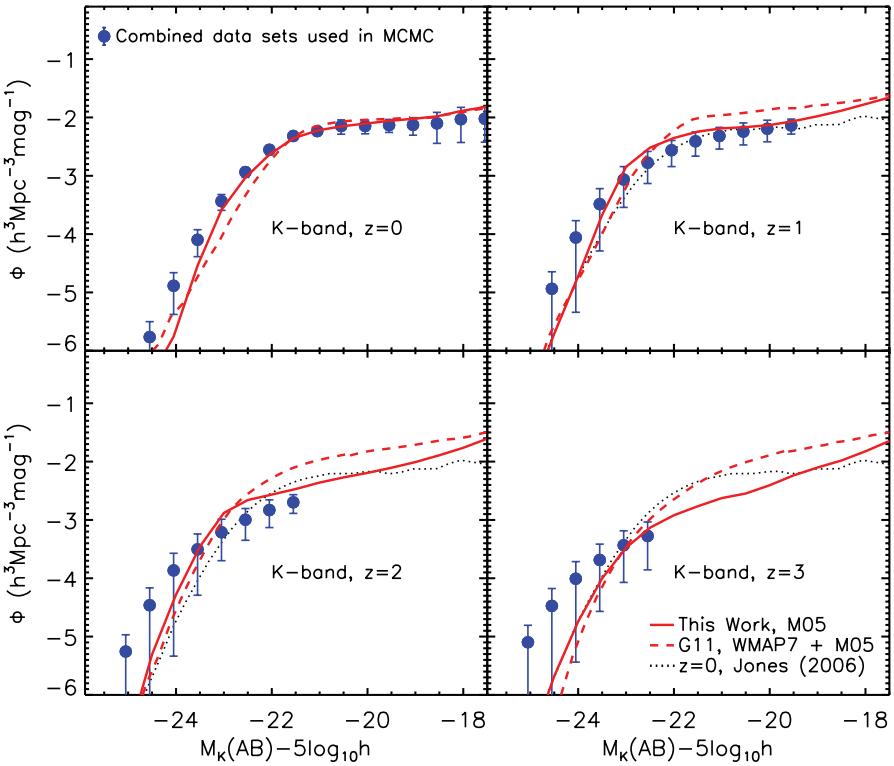



**Figure 5.** Evolution of the rest-frame $K$-band luminosity function from $z = 3$ to $z = 0$. Results for our new model and for G11-*WMAP*7 are shown as solid and dashed red lines, respectively. Blue dots with error bars show the observational constraints and their $1\sigma$ uncertainties as adopted for the MCMC analysis. These result from combining a variety of data sets at each redshift. The individual data sets are shown in Fig. C2. The $z = 0$ results of Jones et al. (2006) are repeated at all redshifts as a black dotted line.

recently, stellar population synthesis models assumed near-infrared emission to come predominantly from old populations. If this were true, the weak evolution in the characteristic luminosity $L_\star$ at rest-frame $K$ from $z = 0$ to $z = 3$, would imply substantially less evolution of the characteristic mass of the stellar mass function over this same redshift range than is expected in hierarchical models of the kind we study here. If, however, intermediate-age populations also emit significantly in the near-infrared (Maraston 2005; Conroy, Gunn & White 2009; Charlot & Bruzual, private communication), then the observed luminosities of the high-redshift galaxies correspond to significantly lower stellar masses, and the observed evolution of the stellar mass and near-infrared luminosity functions can be matched simultaneously. (This is already apparent from the combined results of G11-*WMAP*1 and Henriques et al. 2012.) Acceptable results are obtained for both the Maraston (2005) and Charlot & Bruzual (private communication) stellar population models, but using the older Bruzual & Charlot (2003) models leads to a significant underprediction of $L_\star$ in the $K$-band at higher redshift.

### 5.3 The rest-frame *B*-band luminosity function

Much of the light emitted in the $B$ band is produced by young stars in star-forming regions. In addition, populations become bluer with decreasing metallicity and so emit proportionately more in the $B$ band. Finally, star-forming galaxies also contain dust which absorbs strongly at $B$. Thus, matching observations of the $B$-band luminosity function and its evolution requires detailed modelling of star formation, chemical enrichment and dust production. In our semi-analytic model, star formation happens both quiescently and in merger-induced bursts, and dust is modelled separately for the general ISM and for the birth clouds of young stars. Dust extinction is taken to be proportional to the column density of cold gas in both cases (De Lucia & Blaizot 2007). Metals are returned to the cold gas phase after each star formation episode and later follow the gas as it is transferred between the various baryonic phases (De Lucia et al. 2004). Kitzbichler & White (2007) found that although an earlier version of the Munich model agreed well with the observed $B$-band luminosity function at $z = 0$, it significantly underpredicted $B$-band luminosity functions at high redshift. They corrected this in an ad hoc way by lowering the dust-to-cold gas ratios adopted at early times. The evolution of this ratio with time was slightly modified by Guo & White (2009) and included in the G11-*WMAP*1/G11-*WMAP*7 version of the Munich model, leading to similar results for high-redshift luminosity functions (Henriques et al. 2012), and it is also included here.

In Fig. 6, we compare predictions for $B$-band luminosity functions from $z = 3$ to $z = 0$ with the observational constraints used in our MCMC study. The individual data sets combined to obtain these constraints are shown in Appendix C. Results from our new model with its best-fitting parameter values and from the original G11-*WMAP*7 model are shown as solid and dashed red lines, respectively. At $z \sim 0$, the model is compared to the $b_j$-band luminosity function of the 2dF survey assuming $b_j = B - 0.267(B - V)$ (Norberg et al. 2002), at higher redshift the luminosity functions are for Johnson $B$. The delayed formation of low-mass objects in our new model reduces the amplitude of the high-redshift luminosity functions at fainter magnitudes, but the large observational uncertainties mean that both models are consistent with observation up to $z = 2$. At $z = 3$, both models underpredict the observed abundance at all magnitudes, especially the brightest ones. Errors in the

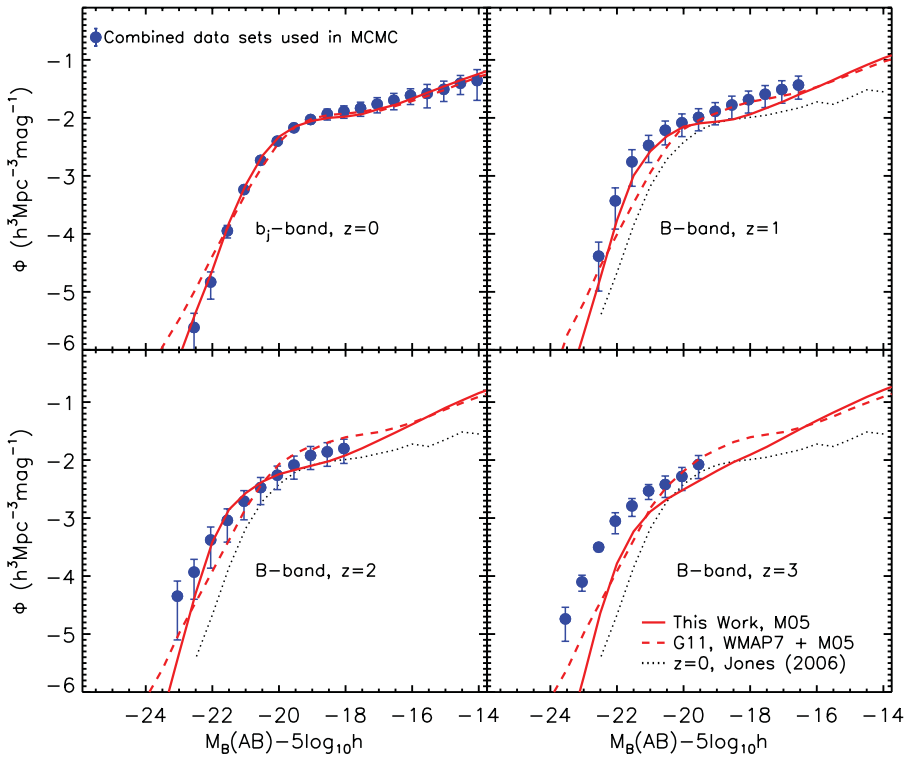



**Figure 6.** Evolution of the rest-frame *B*-band luminosity function from $z = 3$ to $z = 0$. Blue dots with error bars show the observational constraints and their $1\sigma$ uncertainties as adopted for the MCMC analysis. These result from combining a variety of data sets at each redshift. The individual data sets are shown in Fig. C3. The $z = 0$ results of Jones et al. (2006) are repeated at all redshifts as a black dotted line.

treatments of star formation, enrichment and dust obscuration, as well as uncertainties in the stellar population models, may all contribute to this discrepancy.

## 6 MODEL PREDICTIONS

In this section we show model predictions for galaxy properties other than those used as constraints in our MCMC sampling. These include rest-frame colours, star formation rates, ages, stellar metallicities, clustering and the relation between halo mass and stellar mass. We focus on low-redshift data, except that we analyse the evolution of the halo-mass–stellar-mass relation from $z = 3$ to $z = 0$. In most cases, we compare the predictions with data from the Sloan Digital Sky Survey (SDSS) because the very large volume of the survey results in small statistical and cosmic variance uncertainties. Our goal is to investigate whether our improved representation of the evolution of stellar mass and luminosity functions affects any of the other discrepancies between the G11-*WMAP*1/G11-*WMAP*7 model and observed galaxy populations. We find that at $z = 0$ low-mass galaxies in our new model have higher star formation rates, younger stellar populations, are more often blue and are less clustered on small scales than in the G11-*WMAP*1/G11-*WMAP*7 model. In almost all cases this improves the agreement with the observed SDSS galaxy population. Since these properties were not used as constraints when setting parameters, this should be viewed as a significant success for the new model.

### 6.1 Rest-frame colours

Galaxy colours at optical wavelengths are affected by a large number of physical factors, making them difficult to model. In particular, current specific star formation rates, star formation and chemical enrichment histories, the later stages of stellar evolution and the amount and distribution of dust can all have substantial effects on optical colours. Despite this complexity, the observed population of low-redshift galaxies separates fairly cleanly into a 'red sequence' of relatively massive galaxies and a 'blue cloud' of predominantly low-mass dwarfs (Kauffmann et al. 2003; Baldry et al. 2004; Weinmann et al. 2006). In the G11-*WMAP*1/G11-*WMAP*7 model and its predecessors, the majority of dwarf galaxies are much redder than observed, while massive galaxies are not red enough (G11-*WMAP*1). Clearly, star formation is truncated too early in the majority of model dwarfs, an effect which is little influenced by switching from a *WMAP*1 to a *WMAP*7 cosmology (G11-*WMAP*7). The massive galaxies do form most of their stars at high redshift (e.g. De Lucia et al. 2006) so their overly blue colours most likely reflect a metallicity which is too low.

Fig. 7 shows the distributions of $u - i$ colour for eight stellar mass ranges from $10^8$ to $10^{12}\,\mathrm{M}_\odot$ at $z = 0.1$. The new model with its best-fitting parameter values (solid red line) is compared with the G11-*WMAP*7 model (dashed red line). These histograms use data from the MS-II for $10^8 < M_\star < 10^{9.5}\,\mathrm{M}_\odot$ and from the MS for $10^{9.5} < M_\star < 10^{12}\,\mathrm{M}_\odot$. As in G11-*WMAP*1 and G11-*WMAP*7 they are compared with data from SDSS/Data Release 7 (DR7), including $1/V_{\max}$ corrections so that the thin black histograms correspond to volume-limited statistics. All histograms are normalized to have unit integral. There is a clear distinction in the observed distribution between dwarfs, which are almost all blue, and massive galaxies, which are almost all red. Only around the transition at $M_* \sim 10^{10.0}\,\mathrm{M}_\odot$ do both populations appear to exist in comparable numbers. Our new best-fitting model reproduces these trends significantly better than the G11-*WMAP*1/G11-*WMAP*7 model. The

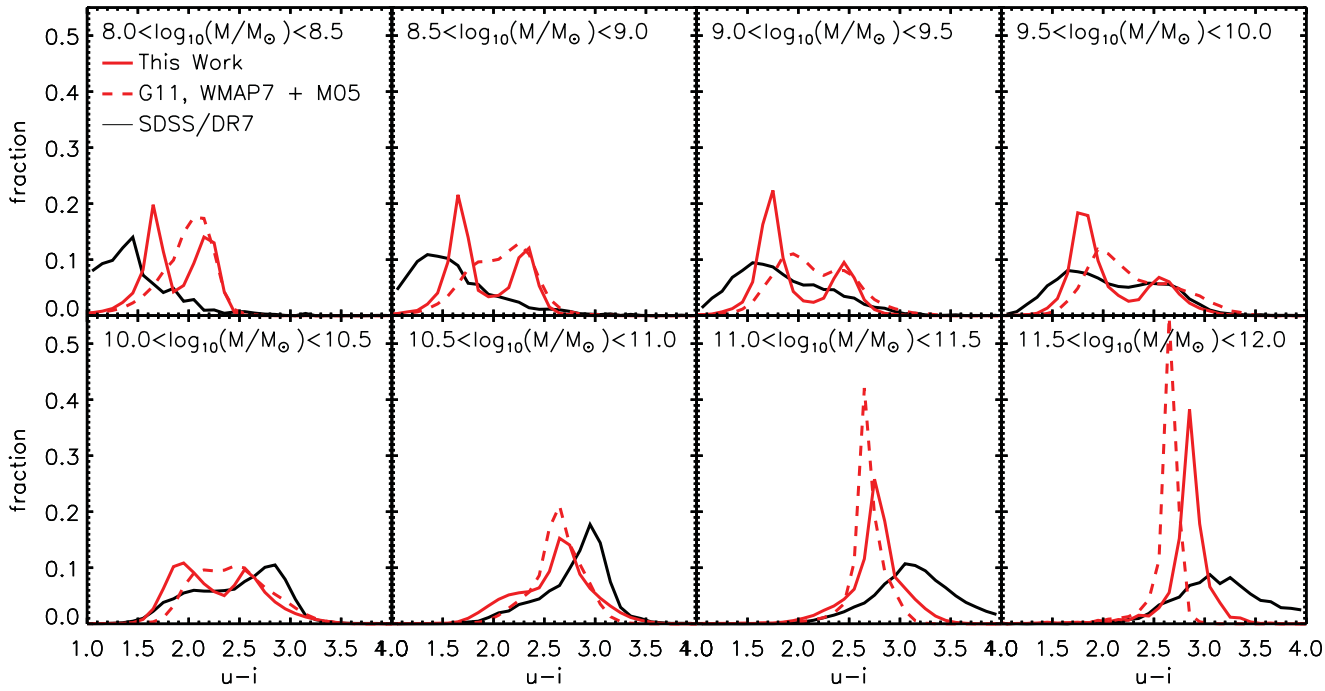



**Figure 7.** Histograms of $u - i$ colour for $z = 0.1$ galaxies in eight disjoint stellar mass ranges. Results for our new best-fitting model (solid red lines) are compared with results from G11-*WMAP*7 (dashed red lines) and with low-redshift data from SDSS/DR7 (solid black lines). The models are based on the MS-II for stellar masses below $10^{9.5}$ $M_{\odot}$ and on the MS for higher masses.

delayed reincorporation of gas and hence the increased fuel supply at late times results in the majority of low-mass galaxies continuing to form stars until $z = 0$. Despite this, the number of passive dwarf galaxies in the model still substantially exceeds that observed.

This excess of passive dwarfs reflects, at least in part, the fact that we impose a threshold for star formation (see equations A1, A2 and A3) since a substantial fraction of the red low-mass galaxies have cold gas masses close to $m_{\rm crit}$. These equations date back to Croton et al. (2006), who based them on results from Kennicutt (1998) and Kauffmann (1996), and they clearly need updating to take account of recent improvements in our observational understanding of the regularities underlying large-scale star formation in galaxies (e.g. Bigiel et al. 2008). First steps in this direction have been made by Fu et al. (2012) who include a prescription for atomic to molecular gas conversion, and assume star formation to be proportional to the molecular content. The observed blue population of low-mass galaxies extends to significantly bluer colours than predicted by the model. This appears to reflect the fact that many of the observed systems have higher specific star formation rates than any star-forming dwarfs in the model (see Fig. 9).

Massive galaxies are still not red enough but the colours we obtain with our new model are closer to those observed than are the colours from G11-*WMAP*1/G11-*WMAP*7. As discussed in the next section, this is a consequence of the higher metal yield in our best-fitting parameter set. This significantly increases the metal abundance of these objects. In general, our theoretical distributions are much narrower than the observed distributions. This may reflect overly narrow metallicity distributions in the model, dust effects or observational errors in the observations.

### 6.2 Stellar metallicities

All chemical elements heavier than lithium are produced and deposited in the gas phase during the later stages of stellar evolution. The amount of metals in future generations of stars is then strongly influenced by mixing processes and by exchanges of material between the cold ISM, the hot gas atmospheres of galaxies, groups and clusters and the circumgalactic medium. It follows that changes in modelling of the physics of gas ejection and reincorporation will be reflected in the stellar mass–metallicity relation. In Fig. 8 we show this relation at $z = 0.1$ for our new model (solid red lines) for the model of G11-*WMAP*7 (dashed red lines) and as estimated from SDSS data by Gallazzi et al. (2005, blue squares and lines). For each sample, the median and the 16 and 84 per cent percentiles are shown as a function of stellar mass. Observations and simulations are normalized by the same solar abundance value of 0.02. Note that Gallazzi et al. (2005) estimate the $1\sigma$ random uncertainties in their SDSS metallicity estimates to range from 0.6 dex at $10^9$ $M_{\odot}$ to 0.2 dex above $10^{11}$ $M_{\odot}$. This can account for much of the difference in scatter between the models and the observations. The optical to near-infrared colours of galaxies are strongly influenced by their metal content, so the constraints we applied on stellar mass and luminosity functions are reflected in the model mass–metallicity relation. As a result, despite the differences in the feedback physics, the predicted distribution shapes are very similar in the two models. Both follow the observed trend approximately, although with a significantly shallower slope.

A noticeable difference between the two models is the higher overall metallicity in the new best fit. This produces massive galaxies with abundances closer to those observed, but also somewhat overpredicts the metallicities of dwarfs. This is a consequence of the higher yield preferred for the new model by our MCMC chains (0.047 as opposed to 0.03 in G11-*WMAP*1/G11-*WMAP*7). The higher metallicity is responsible for the redder colours of massive objects, as shown in the previous section, for the increase in the number density of bright $K$-band galaxies, and for a reduction in the abundance of bright $B$-band objects at $z = 0$ (see Sections 5.2 and 5.3, respectively).

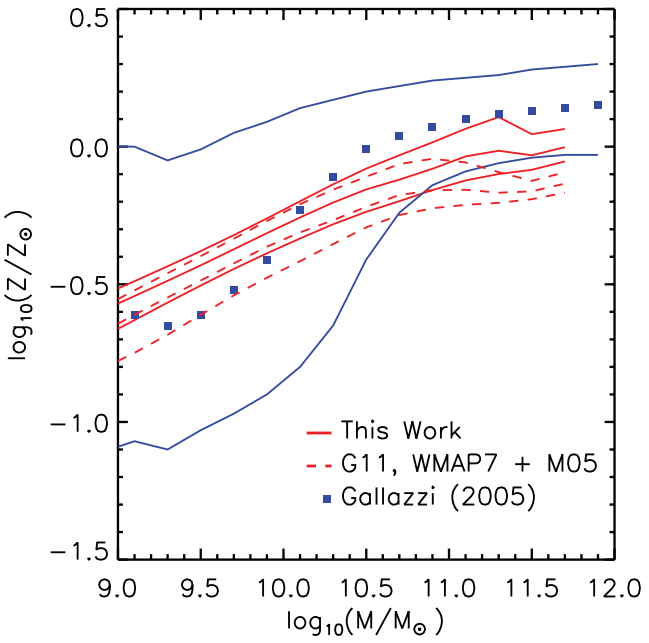



**Figure 8.** The stellar mass–stellar metallicity relation at $z = 0.1$. Our new model with its best-fitting parameter values (solid red lines) is compared with G11-*WMAP*7 (dashed red lines) and with SDSS data from Gallazzi et al. (2005) (blue squares and lines). In all cases, the central lines represent the median metallicity in each mass bin (blue squares for the observations), while the upper and lower lines represent the 16th and 84th percentiles of the distribution.

The steepness of the observed median relation is somewhat puzzling, particularly since the models seem to agree with the observed relation between gas-phase metallicity and stellar mass (see G11-*WMAP*1 – the current model produces similar results). The model slopes in Fig. 8 are quite similar to those found by other published semi-analytic models (e.g. De Lucia & Borgani 2012). At the massive end, a closer fit to observation might be obtained by more effective tidal disruption of satellites (Henriques, Bertone & Thomas 2008; Henriques & Thomas 2010); the high metallicities of stars formed in situ are then less diluted by stars accreted through minor mergers because these are instead assigned to the intracluster light. We note, however, that the G11-*WMAP*1/G11-*WMAP*7 model already includes a substantial intracluster light component in rich clusters. For stellar masses below $\sim 10^{10}\,\mathrm{M}_{\odot}$, strong feedback from SNe is crucial in order to keep the metal abundances low (Bertone, De Lucia & Thomas 2007; Henriques & Thomas 2010; De Lucia & Borgani 2012).

### 6.3 Ages and specific star formation rates in dwarf galaxies

In a recent paper, Weinmann et al. (2012) analysed the overly early build-up of dwarf galaxies in semi-analytic models and in smoothed particle hydrodynamics (SPH) simulations by studying their ages and star formation rates at $z = 0$. As expected, the premature formation of dwarfs leads to older ages and lower star formation rates than observed in the local Universe, with similar discrepancies seen in the G11-*WMAP*1/G11-*WMAP*7 model and in the hydrodynamic simulations of Oppenheimer & Davé (2008). Weinmann et al. (2012) focused on central galaxies, since including satellites would require treatment of additional physical processes and so introduce extra complications into the analysis. Nevertheless, they do note that the disagreement between star formation properties in the models and those observed is considerably worse for satellites than for centrals, as already shown clearly by earlier work (Weinmann et al. 2009; Wang et al. 2012). Using their own semi-analytic model, Bower et al. (2012) found similar trends when looking at the number density of low-mass galaxies at high redshift and their star formation rates at $z = 0$. Our new model correctly predicts the number density of low-mass galaxies as a function of cosmic time, so it is interesting to see if these discrepancies persist. In this section we compare model predictions with data for all types of galaxies, including satellites.

In Fig. 9, we compare the distributions of specific star formation rate (left-hand panel) and *r*-band luminosity-weighted age (right-hand panel) for low-mass galaxies ($10^9 < M_\star < 10^{9.5}\,\mathrm{M}_{\odot}$) with observations at $z = 0.1$. Our new model with its best-fitting parameter values is represented by the solid red lines while results from G11-*WMAP*7 are shown as dashed red lines. SDSS data from the

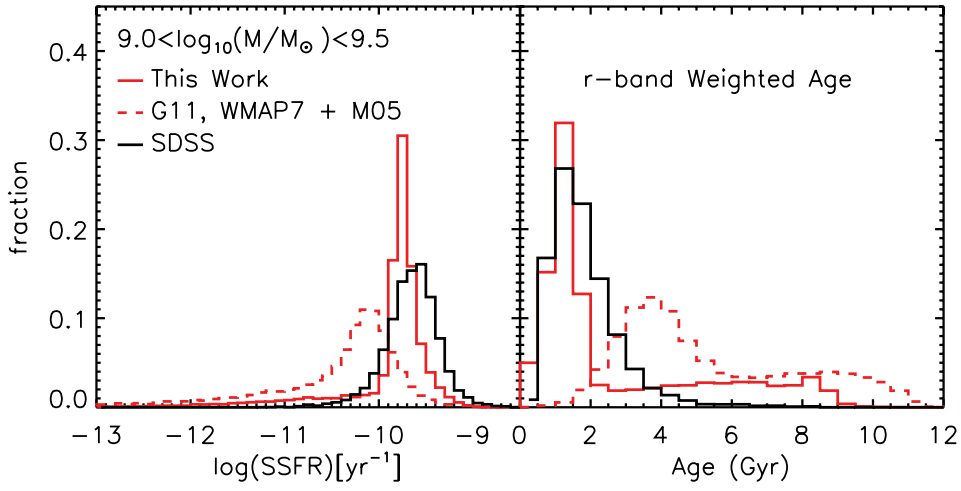



**Figure 9.** The distribution of specific star formation rate (left-hand panel) and *r*-band weighted age (right-hand panel) for galaxies with $10^9 < M_\star < 10^{9.5}\,\mathrm{M}_{\odot}$ at $z = 0.1$. Predictions for our new model with its best-fitting parameter values are shown as solid red lines, while results from G11-*WMAP*7 are shown as dashed red lines (both are based on the MS-II). The models are compared with SDSS data; specific star formation rate estimates come from Salim et al. (2007) and stellar ages from Gallazzi et al. (2005). Both are shown as solid black lines.

MPA-JHU catalogue[3] are plotted as solid black lines after applying a $1/V_{max}$ correction to convert them to volume-limited statistics. The star formation rate data are taken from SDSS-DR7 (Salim et al. 2007) while the luminosity-weighted ages are from SDSS-DR4 (Gallazzi et al. 2005). Stellar mass estimates following Kauffmann et al. (2003) are used in both cases. The longer reincorporation time-scales for ejected gas in our new model delay the growth of low-mass galaxies and shift star formation activity to significantly lower redshifts. As can be seen in the left-hand panel of Fig. 9 this increases specific star formation rates at $z = 0.1$ by almost a factor of 3 and brings them into relatively good agreement with those observed.

This later star formation results in substantially younger luminosity-weighted ages. In the right-hand panel of Fig. 9, the mode of the distribution for low-mass galaxies drops by more than a factor of 2 to ∼1 Gyr. This is in considerably better agreement with SDSS data, despite the long tail to 'old' ages in the model histograms which is almost absent in the observations. This corresponds to the red (primarily) satellite population seen in the relevant panel of Fig. 7. We note that the age discrepancy for the blue population between the G11-*WMAP*1/G11-*WMAP*7 model and the SDSS data appears larger here than in Weinmann et al. (2012) because that paper compared *V*-band weighted ages from the models with *r*-band weighted ages from SDSS.

### 6.4 Stellar-mass–halo-mass relation

The amount of cold gas, and hence of fuel for star formation, in a galaxy depends directly on the mass of its halo. At early times the ratio of baryons to dark matter is almost uniform on the scales which build galaxies, and although later evolution modifies this proportionality considerably, distributing the baryons differently between the stars and the various gas components, it does not erase the strong trend for more massive haloes to have more massive central galaxies. For this reason, a monotonic relation between the stellar mass of a central galaxy and the mass of its halo is expected to be a good first approximation to the outcome of the galaxy formation process. Since the halo of a satellite galaxy can be strongly truncated by tidal effects without a correspondingly large reduction in the mass of its central galaxy, the stellar mass of a satellite is expected to be more tightly related to the maximum past mass of its (sub)halo than to its current subhalo mass (see e.g. Reddick et al. 2012).

Under the assumption of a monotonic relation between central galaxy stellar mass and maximum past halo mass, this SM–HM relation can be obtained explicitly by comparing the observed abundance of galaxies as a function of stellar mass with the simulated abundance of subhaloes as a function of past maximum mass (see Frenk et al. 1988; Vale & Ostriker 2004; Conroy, Wechsler & Kravtsov 2006; Behroozi, Conroy & Wechsler 2010; Guo et al. 2010; Moster et al. 2010; Reddick et al. 2012, for various forms of this abundance matching argument). This assumes, of course, that the real Universe conforms to the Λ cold dark matter (ΛCDM) model, and, in addition, that all real galaxies correspond to a subhalo which survives until the relevant redshift in the dark matter simulation being used. This latter assumption is seriously violated in the galaxy formation models of G11-*WMAP*1/G11-*WMAP*7 and this paper (see e.g. fig. 14 in G11-*WMAP*1) although the violation does not dramatically affect the SM–HM relation.

[3] http://www.mpa-garching.mpg.de/SDSS/

Although the (sub)halo abundance matching method appears very successful at producing mock catalogues which match observation, it has serious drawbacks. In the first place, since the observed abundance of galaxies is used as *input* to the scheme, it cannot be used to constrain the physics of galaxy formation. This has to be done indirectly using the derived SM–HM relation, the clustering of galaxies and their evolution. Secondly, the stellar masses of galaxies undoubtedly do depend on aspects of their assembly history other than the maximum past mass of their haloes, for example, whether they are currently central or satellite galaxies, whether they recently experienced a major merger and so on. These additional dependences may bias mock catalogues constructed by abundance matching, and thus the cosmological parameter values inferred by comparing their large-scale structure to that observed. Such biases can only be identified through models which explicitly follow galaxy formation. Here we compare the SM–HM relation predicted by our new model to that inferred by Moster, Naab & White (2013) who applied an abundance matching technique self-consistently across multiple redshifts using the MS and MS-II. Note that they corrected for 'missing' subhaloes using the G11-*WMAP*1 model, rather than matching to the subhaloes detected in the simulations alone.

Fig. 10 compares results from Moster et al. (2013) for redshifts 0, 1, 2 and 3 (solid lines) to the SM–HM relation in our new model (filled circles give the median central stellar mass at each halo mass, while 'error bars' indicate the $\pm 1\sigma$ scatter). Different colours encode redshift as indicated by the label. Model predictions are based on the MS for $M_{vir} > 10^{11}$ M$_\odot$ and on the MS-II for smaller masses. For comparison, we show the relation for the G11-*WMAP*7 model at $z = 3$ as a dotted blue line (at $z = 0$ the G11-*WMAP*1/G11-*WMAP*7

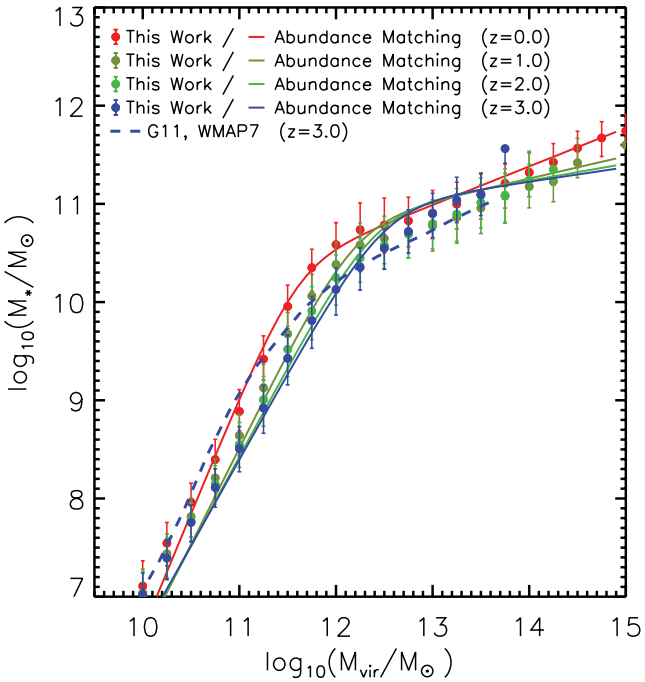



**Figure 10.** Stellar mass as a function of maximum past halo mass for the galaxy populations present at redshifts 3, 2, 1 and 0. Filled circles show median values at each halo mass for our new model with its best-fitting parameter values, with error bars indicating the $\pm 1\sigma$ scatter. For comparison, solid lines show the relations obtained at these same redshifts (coded by colour) using abundance matching techniques by Moster, Naab & White (2013). The dashed blue line is the median relation at $z = 3$ for the G11-*WMAP*7 model.

prediction is very close to that of our new model). For the new model, the agreement with the SM–HM relation inferred directly by Moster et al. (2013) is very good. This is, of course, expected since the new model is a good fit to the high-redshift stellar mass functions we used as constraints, and these are very similar to the observational results used as input by Moster et al. (2013). Notice that the characteristic mass at which the SM–HM relation changes slope now shifts to higher mass with increasing redshift, whereas it remained approximately constant in the G11-*WMAP*1/G11-*WMAP*7 model. As a result, stellar masses are substantially lower in the new model at $z = 3$ for haloes of mass $M_{\rm vir} < 10^{12}\,{\rm M}_\odot$ than in the G11-*WMAP*1/G11-*WMAP*7 model, as required to reduce the abundance of intermediate- and low-mass galaxies to the observed level. It is interesting that the $z = 3$ relation lies below the $z = 0$ relation almost everywhere, showing that galaxy formation was actually less efficient at given halo mass at $z = 3$ than today, despite the 64 times higher density within haloes at that time.

### 6.5 Galaxy clustering

The spatial distribution of galaxies provides many important tests for galaxy formation models. Clustering measurements hold information about the cosmological parameters, the statistics of the initial conditions, the nature of dark matter, the way in which galaxies grow within dark matter haloes, the way in which they are modified by environmental influences and, on small scales, the rates of galaxy collisions and mergers. The latter are of particular relevance for semi-analytic models, since the limited resolution and the absence of baryons in dark matter simulations cause subhaloes to be destroyed by tides well before their central galaxies should themselves be disrupted or merge into the central galaxy of their parent halo.

The G11-*WMAP*1/G11-*WMAP*7 model ties the position of each such orphan galaxy to that of the dark matter particle which was most bound within its subhalo at the last time this could be identified. At later times, the orphan's position is a weighted average of those of the particle and of the central galaxy of the main halo, coinciding with the particle at subhalo disruption and with the central galaxy a dynamical friction time later, at which point orphan and central galaxy merge. This scheme produces better agreement with observed small-scale galaxy clustering than previous cruder models. G11-*WMAP*1/G11-*WMAP*7 found it to reproduce the autocorrelations of galaxies with $\log M_\star/{\rm M}_\odot > 10.8$ on all scales larger than 30 kpc. However for smaller stellar masses they found excess clustering on scales below 1 Mpc. This discrepancy was significantly smaller for a *WMAP*7 cosmology than for the original *WMAP*1 cosmology, but remained noticeable.

In Fig. 11 we show projected autocorrelation functions for galaxies in several disjoint stellar mass ranges at $z = 0.1$. Results for our new model with its best-fitting parameter values are shown as solid red lines while results from G11-*WMAP*7 are shown as dashed red lines. For both models, results from the MS-II are plotted for $\log M_\star/{\rm M}_\odot < 9.77$ and from the MS for higher masses. Model results are compared with the same SDSS/DR7 data used in G11-*WMAP*1 and G11-*WMAP*7. The new model now accurately reproduces the observations for masses above $10^{10.27}\,{\rm M}_\odot$ and for

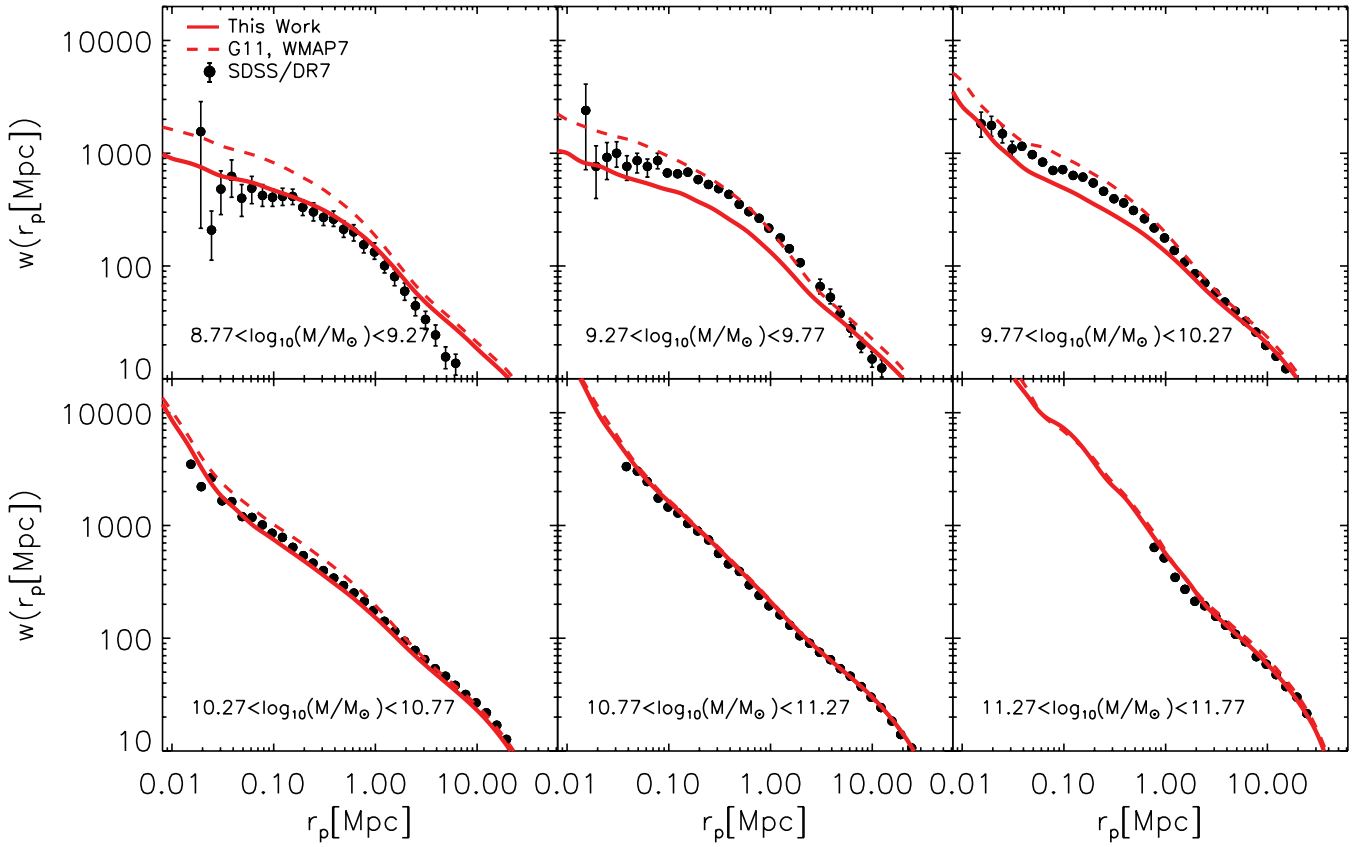



**Figure 11.** Projected two-point autocorrelation functions of galaxies in six disjoint stellar mass ranges at $z = 0.1$. Results from our new model with its best-fitting parameter values (solid red lines) are compared with results from G11-*WMAP*7 (dashed red lines) and with observational data from SDSS/DR7 (solid black lines). The MS-II is used for stellar masses below $10^{9.77}\,{\rm M}_\odot$ and the MS for higher masses.

projected separations larger than 20 kpc. The later assembly of stellar mass caused by our modified reincorporation model reduces the masses of satellite galaxies relative to those of centrals, because the satellites lose their ejecta reservoirs before the gas can be reaccreted. This weakens the 'one-halo' contributions to the autocorrelation functions and eliminates the small-scale clustering excess for $8.77 < \log M_\star/\mathrm{M}_\odot < 9.27$ and $10.27 < \log M_\star/\mathrm{M}_\odot < 10.77$. In the two intermediate-mass bins, the reduction is actually too strong and the new model falls significantly below the observations. Both models significantly overpredict the clustering at $\sim$10 Mpc scales for the lowest mass bin. This can simply result from systematic uncertainties in the data since low-mass galaxies are only observed over a limited volume. These results show that small-scale clustering is quite sensitive to galaxy formation physics, and as a result it is desirable to include clustering data among the constraints when using MCMC sampling to explore galaxy formation issues.

## 7 CONCLUSIONS

Semi-analytic galaxy formation models are designed to explore how the many astrophysical processes which shape the formation and evolution of galaxies are reflected in the systematic properties of the observed galaxy population. In this paper we have extended the MCMC methodology introduced in this context by Henriques et al. (2009) and Henriques & Thomas (2010) to allow population properties at a wide range of redshifts to be used simultaneously as constraints. This exploits one of the strong points of the semi-analytic programme, the fact that its predictions at all redshifts are, by construction, consistent both with the specific physical assumptions of the model and with the growth of structure in a ΛCDM universe.

As observational input, we decided to use stellar mass functions and rest-frame *B*- and *K*-band luminosity functions at $z = 0$, 1, 2 and 3. We combined recent high-quality observational determinations to obtain a uniform set of constraints with plausible and consistent estimates of their uncertainties. We adopted the modelling framework of G11-*WMAP*1/G11-*WMAP*7 and used our optimized MCMC scheme to explore allowed values for 11 of its parameters, those governing the physics of star formation, black hole growth, feedback-driven gas flows and chemical enrichment.

When constrained using data at $z = 0$ alone, the MCMC modelling gave best-fitting parameter values close to those preferred by G11-*WMAP*1/G11-*WMAP*7. Similarly good fits were obtained when the data at any other single redshift were used as constraints. However, no single set of parameter values could provide a good fit at all four redshifts. The allowed regions at different epochs overlap for most parameters, but the time-scale for reincorporation of ejected gas is required to vary differently with redshift than in the G11-*WMAP*1/G11-*WMAP*7 model. This explains why this model, like most other physically based galaxy population models (see e.g. Fontanot et al. 2009; Cirasuolo et al. 2010; Henriques et al. 2011; Bower et al. 2012; Somerville et al. 2012; Weinmann et al. 2012), overpredicts the abundance of lower mass galaxies at high redshift once parameter values are set to fit the present-day galaxy population.

In order to identify the modifications required to match galaxy properties at all epochs, we introduced arbitrary power-law scalings of ejecta return time with redshift and with halo virial velocity, and we reran our MCMC sampling constrained by the abundance data at all four redshifts together. This demonstrated that the extended model could indeed fit the observed stellar mass and luminosity functions at all four redshifts simultaneously. The ejecta return time was required to scale inversely with halo virial mass and to be independent of redshift. This simple result is quite similar to the scaling which Oppenheimer & Davé (2008) and Oppenheimer et al. (2010) found in their cosmological numerical simulations, where the chosen feedback recipes gave reasonable fits to the observed properties of circumgalactic gas. These similarities should nonetheless be taken with caution as the hydroforces between winds and gas are turned off in this simulation for a large fraction of the circumgalactic medium. The gas reincorporation time-scales might therefore closer resemble a semi-analytic calculation.

A comparison of our new best-fitting model with the preferred model of G11-*WMAP*1/G11-*WMAP*7 showed quite substantial changes to the efficiency and scaling of a number of the star formation and feedback processes, despite the fact that the two models produce very similar galaxy populations at $z = 0$. In the new model, the efficiency with which the gas disc makes stars has increased substantially (by more than a factor of 5), SN energy is used at maximum efficiency to eject gas in all but the most massive galaxies, AGN feedback is more efficient (by more than a factor of 3), the mass-loading factor for disc winds depends much more moderately on galaxy rotation velocity and ejecta return times scale more moderately with halo virial velocity $V_{\rm vir}$ but oppositely with redshift (at fixed $V_{\rm vir}$) than in the G11-*WMAP*1/G11-*WMAP*7 model. Combined the modifications ensure a suppression of the formation of low-mass galaxies at early times and a more efficient build-up at $z < 2$, while maintaining the previous evolution of the number density of the most massive objects. An increased yield leads to better reproduction of the metallicities of massive galaxies.

The fact that models with such substantial differences produce similar present-day populations is a dramatic illustration of the interconnectedness of the many processes influencing galaxy formation. Nevertheless, our MCMC analysis demonstrates that there are no significant degeneracies when each model is fit to the observational data it was designed to interpret. All the parameters are required for an adequate fit. The physical interest of the fits lies in the specific scalings and efficiencies implied, in particular, in checking whether the efficiencies lie within their plausible ranges. In this context it may be worrying that our new model requires SN energy to be used so efficiently to eject gas from galaxy discs.

In our new model ejected gas returns to lower mass haloes much less rapidly at high redshift than in the G11-*WMAP*1/G11-*WMAP*7 model. As a result, low-mass galaxies grow more slowly, and their abundance as a function of stellar mass and luminosity is a better fit to the observations. At $z < 1$, however, ejecta return to all but the smallest haloes more rapidly than in the earlier model, leading to stronger late-time star formation and similar $z = 0$ stellar masses. By construction, these changes eliminate one of the main problems which G11-*WMAP*7 identified in their model: its overly early production of lower mass galaxies. It is interesting that the excessive red colours and clustering of low-mass galaxies are also reduced, even though the relevant observations were not used as constraints in our MCMC modelling. As a result, in the new model:

(i) the observed galaxy abundances are correctly represented over fie order of magnitude in mass ($7 < \log M_\star/\mathrm{M}_\odot < 12$) at $z = 0$ and over two order of magnitude in mass throughout the last five sixths of the age of the Universe;

(ii) the fraction of dwarf galaxies ($\log M_\star/\mathrm{M}_\odot < 9.5$) which are blue, young and actively star forming at $z = 0$ is substantially higher than in the model of G11-*WMAP*7;

(iii) the clustering of dwarf galaxies is reduced as result of the increased central to satellite ratio at fixed stellar mass, and is no longer systematically above that observed.

Nevertheless, at the lowest masses, star-forming galaxies are still older and the passive fraction is still significantly larger than observed. This problem may be related to the star formation threshold adopted in the model (see Appendix A1), since many of the dwarfs, while containing considerable gas, are sitting very close to the threshold.

The increased heavy element yield in the new model produces higher metallicities and redder colours for massive galaxies, improving the fit to observations. On the other hand, the slope of the stellar mass–metallicity relation remains shallower than observed, so the stellar metallicities of low-mass galaxies are now overpredicted. Henriques & Thomas (2010) noted that the slope of this relation can be increased by allowing more efficient tidal disruption of satellites. The stellar populations of massive galaxies are then less diluted by accretion of low-metallicity material from merging satellites. It remains to be seen whether this can improve the chemical properties of the model without destroying its ability to fit the luminosities of massive galaxies which currently grow substantially more through merging than through star formation (see Guo & White 2008). A more complete treatment of chemical evolution, allowing detailed comparison with the observational information now available on stellar and gas-phase abundances and abundance ratios, will undoubtedly be a fruitful extension of the kind of galaxy population modelling presented in this paper.

## ACKNOWLEDGEMENTS

The computations used in this paper were performed on the Cosmology Machine supercomputer at the Institute for Computational Cosmology, which is part of the DiRAC Facility jointly funded by STFC, the Large Facilities Capital Fund of BIS and Durham University. The work of BMBH, SDMW, REA and GL was supported by Advanced Grant 246797 'GALFORMOD' from the European Research Council. PAT was supported by the Science and Technology Facilities Council [grant number ST/I000976/1]. QG acknowledges support from the National basic research program of China (973 program under grant No. 2009CB24901), the Young Researcher Grant of National Astronomical Observatories, CAS, the NSFC grants program (No. 11143005), as well as the Partner Group program of the Max Planck Society. VS acknowledges support by the Deutsche Forschungsgemeinschaft through Transregio 33, 'The Dark Universe'. The authors thank Michele Cirasuolo, Danilo Marchesini and Rob Yates for providing their observational data and the referee Joop Schaye for insightful suggestions that helped improving the quality of the paper.

## REFERENCES

Angulo R. E., White S. D. M., 2010, MNRAS, 405, 143
Baldry I. K., Glazebrook K., Brinkmann J., Ivezić Ž., Lupton R. H., Nichol R. C., Szalay A. S., 2004, ApJ, 600, 681
Baldry I. K., Glazebrook K., Driver S. P., 2008, MNRAS, 388, 945
Behroozi P. S., Conroy C., Wechsler R. H., 2010, ApJ, 717, 379
Bell E. F., McIntosh D. H., Katz N., Weinberg M. D., 2003, ApJS, 149, 289
Benson A. J., Bower R., 2011, MNRAS, 410, 2653
Benson A. J., Bower R. G., Frenk C. S., Lacey C. G., Baugh C. M., Cole S., 2003, ApJ, 599, 38
Bertone S., De Lucia G., Thomas P. A., 2007, MNRAS, 379, 1143
Bigiel F., Leroy A., Walter F., Brinks E., de Blok W. J. G., Madore B., Thornley M. D., 2008, AJ, 136, 2846
Bower R. G., McCarthy I. G., Benson A. J., 2008, MNRAS, 390, 1399
Bower R. G., Vernon I., Goldstein M., Benson A. J., Lacey C. G., Baugh C. M., Cole S., Frenk C. S., 2010, MNRAS, 407, 2017
Bower R. G., Benson A. J., Crain R. A., 2012, MNRAS, 422, 2816
Boylan-Kolchin M., Springel V., White S. D. M., Jenkins A., Lemson G., 2009, MNRAS, 398, 1150
Bruzual G., Charlot S., 2003, MNRAS, 344, 1000
Caputi K. I., McLure R. J., Dunlop J. S., Cirasuolo M., Schael A. M., 2006, MNRAS, 366, 609
Chabrier G., 2003, PASP, 115, 763
Cirasuolo M., McLure R. J., Dunlop J. S., Almaini O., Foucaud S., Simpson C., 2010, MNRAS, 401, 1166
Cole S., 1991, ApJ, 367, 45
Cole S., Aragon-Salamanca A., Frenk C. S., Navarro J. F., Zepf S. E., 1994, MNRAS, 271, 781
Cole S. et al., 2001, MNRAS, 326, 255
Conroy C., Wechsler R. H., Kravtsov A. V., 2006, ApJ, 647, 201
Conroy C., Gunn J. E., White M., 2009, ApJ, 699, 486
Croton D. J. et al., 2006, MNRAS, 365, 11
De Lucia G., Blaizot J., 2007, MNRAS, 375, 2
De Lucia G., Borgani S., 2012, MNRAS, 426, L61
De Lucia G., Kauffmann G., White S. D. M., 2004, MNRAS, 349, 1101
De Lucia G., Springel V., White S. D. M., Croton D., Kauffmann G., 2006, MNRAS, 366, 499
Domínguez Sánchez H. et al., 2011, MNRAS, 417, 900
Drory N., Bender R., Feulner G., Hopp U., Maraston C., Snigula J., Hill G. J., 2003, ApJ, 595, 698
Fontana A. et al., 2006, A&A, 459, 745
Fontanot F., Monaco P., 2010, MNRAS, 405, 705
Fontanot F., Somerville R. S., Silva L., Monaco P., Skibba R., 2009, MNRAS, 392, 553
Frenk C. S., White S. D. M., Davis M., Efstathiou G., 1988, ApJ, 327, 507
Fu J., Kauffmann G., Li C., Guo Q., 2012, MNRAS, 424, 2701
Gallazzi A., Charlot S., Brinchmann J., White S. D. M., Tremonti C. A., 2005, MNRAS, 362, 41
Giallongo E., Salimbeni S., Menci N., Zamorani G., Fontana A., Dickinson M., Cristiani S., Pozzetti L., 2005, ApJ, 622, 116
Guo Q., White S. D. M., 2008, MNRAS, 384, 2
Guo Q., White S. D. M., 2009, MNRAS, 396, 39
Guo Q., White S., Li C., Boylan-Kolchin M., 2010, MNRAS, 404, 1111
Guo Q. et al., 2011, MNRAS, 413, 101 (G11-*WMAP*1)
Guo Q., White S., Angulo R. E., Henriques B., Lemson G., Boylan-Kolchin M., Thomas P., Short C., 2013, MNRAS, 428, 1351 (G11-*WMAP*7)
Haas M. R., Schaye J., Booth C. M., Dalla Vecchia C., Springel V., Theuns T., Wiersma R. P. C., 2012, preprint (arXiv:1211.1021)
Hastings W. K., 1970, Biometrika, 57, 97
Hatton S., Devriendt J. E. G., Ninin S., Bouchet F. R., Guiderdoni B., Vibert D., 2003, MNRAS, 343, 75
Helly J. C., Cole S., Frenk C. S., Baugh C. M., Benson A., Lacey C., 2003, MNRAS, 338, 903
Henriques B. M. B., Thomas P. A., 2010, MNRAS, 403, 768
Henriques B. M., Bertone S., Thomas P. A., 2008, MNRAS, 383, 1649
Henriques B. M. B., Thomas P. A., Oliver S., Roseboom I., 2009, MNRAS, 396, 535
Henriques B., Maraston C., Monaco P., Fontanot F., Menci N., De Lucia G., Tonini C., 2011, MNRAS, 415, 3571
Henriques B. M. B., White S. D. M., Lemson G., Thomas P. A., Guo Q., Marleau G.-D., Overzier R. A., 2012, MNRAS, 421, 2904
Ilbert O. et al., 2005, A&A, 439, 863
Ilbert O. et al., 2010, ApJ, 709, 644
Jones D. H., Peterson B. A., Colless M., Saunders W., 2006, MNRAS, 369, 25
Kampakoglou M., Trotta R., Silk J., 2008, MNRAS, 384, 1414
Kang X., Jing Y. P., Mo H. J., Börner G., 2005, ApJ, 631, 21
Kauffmann G., 1996, MNRAS, 281, 487
Kauffmann G., Haehnelt M., 2000, MNRAS, 311, 576

Kauffmann G., White S. D. M., Guiderdoni B., 1993, MNRAS, 264, 201
Kauffmann G., Colberg J., Diaferio A., White S., 1999, MNRAS, 303, 188
Kauffmann G. et al., 2003, MNRAS, 341, 33
Kennicutt R. C., Jr, 1998, ApJ, 498, 541
Kitzbichler M. G., White S. D. M., 2007, MNRAS, 376, 2
Lacey C., Silk J., 1991, ApJ, 381, 14
Li C., White S. D. M., 2009, MNRAS, 398, 2177
Lu Y., Kereš D., Katz N., Mo H. J., Fardal M., Weinberg M. D., 2011a, MNRAS, 416, 660
Lu Y., Mo H. J., Weinberg M. D., Katz N., 2011b, MNRAS, 416, 1949
Lu Y., Mo H. J., Katz N., Weinberg M. D., 2012, MNRAS, 421, 1779
Maraston C., 2005, MNRAS, 362, 799
Marchesini D. et al., 2007, ApJ, 656, 42
Marchesini D., van Dokkum P. G., Förster Schreiber N. M., Franx M., Labbé I., Wuyts S., 2009, ApJ, 701, 1765
Marchesini D. et al., 2010, ApJ, 725, 1277
Metropolis N., Rosenbluth A., Rosenbluth M., Teller A., Teller E., 1953, J. Chem. Phys., 21, 1087
Moster B. P., Somerville R. S., Maulbetsch C., van den Bosch F. C., Macciò A. V., Naab T., Oser L., 2010, ApJ, 710, 903
Moster B. P., Naab T., White S. D. M., 2013, MNRAS, 428, 3121
Mutch S. J., Poole G. B., Croton D. J., 2013, MNRAS, 428, 2001
Neistein E., Weinmann S. M., 2010, MNRAS, 405, 2717
Norberg P. et al., 2002, MNRAS, 336, 907
Oppenheimer B. D., Davé R., 2008, MNRAS, 387, 577
Oppenheimer B. D., Davé R., Kereš D., Fardal M., Katz N., Kollmeier J. A., Weinberg D. H., 2010, MNRAS, 406, 2325
Pérez-González P. G. et al., 2008, ApJ, 675, 234
Poli F. et al., 2003, ApJ, 593, L1
Pozzetti L. et al., 2003, A&A, 402, 837
Pozzetti L. et al., 2010, A&A, 523, A13
Press W. H., Schechter P., 1974, ApJ, 187, 425
Reddick R. M., Wechsler R. H., Tinker J. L., Behroozi P. S., 2012, preprint (arXiv:1207.2160)
Roukema B. F., Quinn P. J., Peterson B. A., Rocca-Volmerange B., 1997, MNRAS, 292, 835
Salim S. et al., 2007, ApJS, 173, 267
Salimbeni S. et al., 2008, A&AS, 477, 763
Saracco P. et al., 2006, MNRAS, 367, 349
Somerville R. S., Primack J. R., 1999, MNRAS, 310, 1087
Somerville R. S., Primack J. R., Faber S. M., 2001, MNRAS, 320, 504
Somerville R. S., Gilmore R. C., Primack J. R., Domínguez A., 2012, MNRAS, 423, 1992
Springel V., White S. D. M., Tormen G., Kauffmann G., 2001, MNRAS, 328, 726
Springel V. et al., 2005, Nat, 435, 629
Tonini C., Maraston C., Thomas D., Devriendt J., Silk J., 2010, MNRAS, 403, 1749
Vale A., Ostriker J. P., 2004, MNRAS, 353, 189
Wang L., Weinmann S. M., Neistein E., 2012, MNRAS, 421, 3450
Weinmann S. M., van den Bosch F. C., Yang X., Mo H. J., 2006, MNRAS, 366, 2
Weinmann S. M., Kauffmann G., van den Bosch F. C., Pasquali A., McIntosh D. H., Mo H., Yang X., Guo Y., 2009, MNRAS, 394, 1213
Weinmann S. M., Pasquali A., Oppenheimer B. D., Finlator K., Mendel J. T., Crain R. A., Macciò A. V., 2012, MNRAS, 426, 2797
White S. D. M., 1989, in Frenk C. S., Ellis R. S., Shanks T., Heavens A. R., Peacock J. A., eds, NATO ASIC Proc. 264, The Epoch of Galaxy Formation. Kluwer, Dordrecht, p. 15
White S. D. M., Frenk C. S., 1991, ApJ, 379, 52
Willmer C. N. A. et al., 2006, ApJ, 647, 853
Zucca E. et al., 2009, A&A, 508, 1217

## APPENDIX A: PHYSICAL PRESCRIPTIONS

Our semi-analytic model is built on top of high-resolution *N*-body simulations that provide a detailed model for the evolution of the dark matter distribution. At each time, a FoF algorithm is used to define a set of haloes, each of which is then further divided into a set of subhaloes, disjoint, self-bound particle clumps with a single gravitational potential minimum where a galaxy is assumed to form. Each subhalo at time $i$ is linked to a unique descendant at time $i + 1$ by following its constituent particles. These links produce merger trees which specify the full assembly history of every $z = 0$ subhalo, and are the data structures on which the semi-analytic model operates to follow the formation and evolution of the galaxy that each subhalo contains. As subhaloes form and grow through accretion, new dark matter is assumed to bring the global mean baryon fraction with it. These baryons are transferred among the various baryonic components associated with the subhalo (hot gas atmosphere, cold disc gas, ejected gas, disc and bulge stars) by the semi-analytic model. In very low-mass haloes, gas infall is suppressed by UV heating once the Universe is ionized. At higher mass, accretion of newly acquired gas occurs in two modes: at high redshift and for subhaloes with mass below $\sim 10^{11.5}\,\mathrm{M}_{\odot}$, cooling is rapid and infalling gas accretes directly on to the central galaxy; at later times and in more massive haloes, the gas shocks to form a quasi-static atmosphere from which it gradually accretes on to the galaxy through a cooling flow. Although this spherical model does not allow filamentary inflow to coexist with a hot atmosphere, it produces galaxy accretion rates as a function of halo mass and redshift which seem to agree reasonably with those derived from the detailed gas dynamical simulations of Benson & Bower (2011) (but see also Lu et al. 2011a). In practice, the situation is considerably complicated by the strong galactic winds required in all realistic models.

Star formation occurs in cold disc gas, either quiescently or in merger-induced bursts. A fraction of the stars formed (those with masses above $8\,\mathrm{M}_{\odot}$) is assumed to explode as Type II SNe. This mass is instantaneously returned to the cold phase and the energy released is used to reheat gas from the cold to the hot phase. Left-over energy is used to eject gas from the hot phase into an external reservoir. This ejected gas can later be reincorporated in the hot phase, and it is this aspect of the G11-*WMAP*1/G11-*WMAP*7 model which we modify in this paper, as described in Section 4. The evolution of massive stars also produces new metals that are deposited in the cold phase. Each newly formed galaxy is assigned a central seed black hole that then grows through accretion of cold gas during mergers and through quiescent accretion of hot gas. When the black hole reaches a sufficiently large mass, this quiescent accretion is assumed to power a radio source which offsets cooling in the hot gas, thereby quenching star formation.

In the following subsections, we present the equations describing those aspects of the semi-analytic modelling for which the parameter values have been allowed to vary in our MCMC analysis. These parameters are $\alpha_{\rm SF}$, $k_{\rm AGN}$, $f_{\rm BH}$, $\epsilon$, $V_{\rm reheat}$, $\beta_1$, $\eta$, $V_{\rm eject}$, $\beta_2$, $\gamma'$ and $Z_{\rm yield}$. For a more complete description of the model we refer the reader to G11-*WMAP*1.

### A1 Star formation

Over most of the lifetime of a galaxy, star formation happens in a quiescent mode through the fragmentation of cold gas that reaches a sufficiently high density. The star formation rate in this mode is modelled as

$$\dot{m}_{\star} = \alpha_{\rm SF} \frac{(m_{\rm cold} - m_{\rm crit})}{t_{\rm dyn,disc}}, \qquad \text{(A1)}$$

where $m_{\rm cold}$ is the mass of cold gas and $t_{\rm dyn,\,disc} = r_{\rm disc}/V_{\rm max}$ is the dynamical time of the disc. Note that the mass locked up in stars over a time interval d$t$ is $(1-R)\dot{m}_\star\,{\rm d}t$, the rest being instantaneously returned to the cold disc gas. ($R = 0.43$ here denotes the recycled fraction as determined from the Chabrier 2003 initial mass function.) The threshold gas mass above which star formation is assumed to occur, $m_{\rm crit}$, is derived from a threshold surface density at each radius as given by Kauffmann (1996):

$$\Sigma_{\rm crit}(R) = 120\left(\frac{V_{\rm vir}}{200\,{\rm km\,s^{-1}}}\right)\left(\frac{R}{\rm kpc}\right)^{-1}\,{\rm M_\odot\,pc^{-2}}, \tag{A2}$$

leading to the estimate

$$m_{\rm crit} = 3.8\times10^9\left(\frac{V_{\rm vir}}{200\,{\rm km\,s^{-1}}}\right)\left(\frac{r_{\rm disc}}{10\,{\rm kpc}}\right)\,{\rm M_\odot}, \tag{A3}$$

where $r_{\rm disc}$ is obtained from the specific angular momentum of the cold gas disc which in turn is calculated by summing the contributions to the angular momentum vector from all infalling gas (G11-*WMAP*1).

During mergers it is assumed that the cold gas is disturbed significantly, allowing a large fraction to reach the densities necessary for fragmentation. This phenomenon is modelled (following Somerville, Primack & Faber 2001) as a higher efficiency for star formation during merger events, with the total amount of stars formed given by

$$m_\star^{\rm burst} = 0.56\left(\frac{m_{\rm sat}}{m_{\rm central}}\right)^{0.7} m_{\rm gas}. \tag{A4}$$

The burst parameters are kept fixed in the MCMC sampling; of the parameters in this section, only $\alpha_{\rm SF}$ is allowed to vary.

### A2 Supernova feedback

Massive stars release large amounts of energy into the surrounding medium both through their stellar winds and when they explode as SNe. This energy can reheat cold disc gas and inject it into the hot atmosphere, and it may also eject hot gas from the virialized regions of the galaxy/subhalo system. The ejecta may then fall back into the system at some later time. This feedback-induced cycle is treated by the semi-analytic model as a three stage process involving gas reheating, gas ejection and the reincorporation of ejected gas.

The mass of gas reheated by star formation is assumed to be directly proportional to the amount of stars formed:

$$\Delta m_{\rm reheated} = \epsilon_{\rm disc}\,\Delta m_\star, \tag{A5}$$

where

$$\epsilon_{\rm disc} = \epsilon\left[0.5 + \left(\frac{V_{\rm max}}{V_{\rm reheat}}\right)^{-\beta_1}\right]. \tag{A6}$$

Any SN energy left over after reheating the disc gas is used to eject material from the hot atmosphere into an external reservoir. The total energy input from SN is assumed to be

$$\Delta E_{\rm SN} = \epsilon_{\rm halo}\,\frac{1}{2}\Delta m_\star V_{\rm SN}^2, \tag{A7}$$

where $\frac{1}{2}V_{\rm SN}^2$ is the mean energy in SN ejecta per unit mass of stars formed, $V_{\rm SN}^2 = 630\,{\rm km\,s^{-1}}$ and

$$\epsilon_{\rm halo} = \eta\left[0.5 + \left(\frac{V_{\rm max}}{V_{\rm eject}}\right)^{-\beta_2}\right]. \tag{A8}$$

The ejected mass is then derived from the residual energy after reheating:

$$\frac{1}{2}\Delta M_{\rm eject}V_{\rm vir}^2 = \Delta E_{\rm SN} - \frac{1}{2}\Delta M_{\rm reheat}V_{\rm vir}^2. \tag{A9}$$

As described in G11-*WMAP*1, if this equation gives $\delta M_{\rm ejec} < 0$, the reheated mass is assumed to saturate at $\delta M_{\rm reheat} = \delta E_{\rm SN}/(\frac{1}{2}V_{\rm vir}^2)$. Also, $\epsilon_{\rm halo}$ is forced never to exceed unity. Both conditions ensure that the total amount of energy used in feedback is limited to that available from SNe. In practice this limit should not be reached as it would imply no radiative losses and no work done by the outflows. Material ejected from the hot atmosphere can be reincorporated at later times according to the models described in Section 4.

### A3 AGN feedback

Three processes are assumed to control the growth of massive central black holes and their effects on their environment. Cold gas accretion during mergers fuels 'quasar mode' growth which is the principal route by which the black hole population gains mass. Black holes are also assumed to accrete quiescently from hot gas atmospheres in a 'radio mode' which does not add significantly to their mass but pumps energy into the hot gas, thus counteracting its cooling and quenching the supply of new cold gas for star formation. Finally central black holes are assumed to merge when their parent galaxies merge.

The amount of gas accreted in quasar mode is taken to be proportional to the mass ratio between the two merging galaxies and to their total amount of cold gas:

$$\Delta m_{\rm BH,Q} = \frac{f_{\rm BH}(m_{\rm sat}/m_{\rm central})\,m_{\rm cold}}{1+(280\,{\rm km\,s^{-1}}/V_{\rm vir})^2}. \tag{A10}$$

The rate of radio mode accretion from the hot gas is taken to be

$$\dot{m}_{\rm BH,R} = k_{\rm AGN}\left(\frac{m_{\rm BH}}{10^8\,{\rm M_\odot}}\right)\left(\frac{f_{\rm hot}}{0.1}\right)\left(\frac{V_{\rm vir}}{200\,{\rm km\,s^{-1}}}\right)^3, \tag{A11}$$

where $m_{\rm BH}$ is the black hole mass and $f_{\rm hot}$ is the mass fraction of hot gas in the halo. This accretion is assumed to pump energy back into the hot gas through mechanical heating at a substantial fraction of the Eddington rate

$$L_{\rm BH} = \zeta\dot{m}_{\rm BH,R}c^2, \tag{A12}$$

where $\dot{m}_{\rm BH,R}$ is given by equation (A11), $c$ is the speed of light and $\zeta$ is set to 0.1 (which in combination with $k_{\rm AGN}$ determines the heating rate). This energy is used to offset the cooling of gas from the hot atmosphere on to the cold disc, resulting in a modified cooling rate,

$$\dot{m}'_{\rm cool} = \dot{m}_{\rm cool} - \frac{L_{\rm BH}}{\frac{1}{2}V_{\rm vir}^2}, \tag{A13}$$

which is required to be non-negative.

### A4 Metal enrichment

The metals produced by the stars are assumed to follow the same processing cycle as the gas. For every solar mass of stars formed, a mass $Z_{\rm yield}\,{\rm M_\odot}$ of heavy elements is returned instantaneously to the cold phase, where $Z_{\rm yield}$ is known as the yield and is the only chemical enrichment parameter which we allow to vary in our MCMC analysis. These heavy elements then follow the gas flow. They can be locked in new long-lived stars, or they can be reheated into the hot phase, ejected into the external reservoir, reincorporated

in the hot phase and returned to the cold phase by cooling. Each of the three gas phases is assumed to be chemically homogeneous at each time, i.e. mixing of heavy elements within each phase is assumed to be instantaneous.

## APPENDIX B: MONTE CARLO MARKOV CHAIN ANALYSIS

As discussed in Henriques et al. (2009), implementing the MCMC approach for cosmological galaxy formation modelling on simulations like the MS and MS-II raises significant challenges with respect to computational power. In a brute force approach, the semi-analytic model would need to follow the formation and evolution of up to 20 million galaxies throughout the simulation volume for every parameter set proposed at a MCMC step. Population distributions of interest could then be compared with the observational constraints to compute a likelihood. The size of the calculation and the required number of steps make this a very costly proposition. Henriques et al. (2009) and Henriques & Thomas (2010) therefore restricted themselves to a small but approximately representative subvolume, just 1/512 of the full MS. While reasonably efficient, their approach did not optimize the fidelity of the representation of the full model for given computational resources. Here, we adopt a scheme which defines a representative set of subhalo merger trees such that abundance statistics such as those used as constraints in this paper can be modelled with controlled and approximately uniform precision and at minimum computational cost across the full range of stellar masses (or luminosities) considered. It is also important to ensure that we have a properly representative model at all the redshifts considered. Our scheme allows merger trees from the MS and MS-II to be used simultaneously, significantly increasing the dynamical range in galaxy mass/luminosity which can be used to constrain the model.

Consider the problem of estimating the galaxy luminosity function $\Phi(L)$ of a specific model, which we take to be the function obtained when the model is applied to the full MS and MS-II simulations. We want to find a small subset of the halo merger trees which reproduces $\Phi(L)$ to some desired accuracy over some luminosity range of interest. Let us assume that our simulated volume contains $N$ haloes, partitioned into a total of $I$ mass bins labelled by $i$. Then

$$N = \sum_{i=1}^{I} N_i, \tag{B1}$$

where $N_i$ is the number of haloes in the $i$th mass bin. Further, assume that we wish to estimate a luminosity function defined on a set of $J$ luminosity bins labelled by the index $j$. We can write the luminosity function of the simulated volume as a whole as

$$\Phi_j = \sum_{i=1}^{I} N_i \phi_{ij}, \tag{B2}$$

where $\phi_{ij}$ is the average number of galaxies in luminosity bin $j$ for a randomly selected simulation halo in mass bin $i$. We seek a set of numbers $n_i$ such that choosing $n_i$ haloes at random among the $N_i$ in the $i$th bin is sufficient to estimate the luminosity to function to some pre-assigned accuracy, for example, $\pm 5$ per cent.

For a specific subsample of $n_i$ haloes in each bin $i$, a straightforward estimate of $\Phi_j$ is

$$\tilde{\Phi}_j = \sum_{i=1}^{I} \frac{N_i}{n_i} \sum_{k=1}^{n_i} s_{ijk}, \tag{B3}$$

where $s_{ijk}$ is the number of galaxies in the $j$th luminosity bin for the $k$th of the $n_i$ haloes in mass bin $i$. For different $k$ the $s_{ijk}$ are independent, identically distributed random variables with non-negative integer values and mean $\phi_{ij}$. We will assume that they are Poisson distributed, but we note that this, in fact, overestimates their variance because each halo contains one and only one central galaxy and such central galaxies dominate the bright end of the luminosity function (see e.g. fig. 24 of G11-*WMAP*1). With this Poisson assumption we have

$$\langle \tilde{\Phi}_j \rangle = \Phi_j \tag{B4}$$

and

$$\mathrm{var}\left[\tilde{\Phi}_j\right] = \sum_{i=1}^{I} \frac{N_i^2}{n_i^2} \phi_{ij}. \tag{B5}$$

If we wish the rms uncertainty of our luminosity function estimate to be a fraction $F$ of its true value in each of the luminosity bins, we then require

$$\sum_{i=1}^{I} \frac{N_i^2}{n_i^2} \phi_{ij} = F^2 \Phi_j^2; \quad j = 1, J. \tag{B6}$$

For a specific galaxy formation model implemented on the MS, the G11-*WMAP*7 model, for example, we know the $N_i$, the $\Phi_j$ and the $\phi_{ij}$, so equations (B6) are a set of linear constraint equations on $F^2 n_i$. In the specific case $I = J$, the matrix $\phi_{ij}$ is square and can be inverted to get an explicit solution for the $n_i$,

$$\frac{1}{n_i^2} = \frac{F^2}{N_i^2} \sum_{j=1}^{J} \Phi_j^2 \phi_{ji}^{-1}; \quad i = 1, I. \tag{B7}$$

If $I > J$ and if for every $i$ at least one of the $\phi_{ij}$ is non-zero, the solution space of equations (B6) is $(I - J)$-dimensional and one can normally pick a set of $n_i$ values satisfying the requirements $0 \leq n_i \leq N_i$ and minimizing some computational cost function, for example, the total number of trees chosen, $\sum_i n_i$, or the total number of galaxies modelled, $\sum_{ij} n_i \phi_{ij}$. For our analysis we choose $F = 0.05$ to ensure that the uncertainty of the luminosity and stellar mass functions estimated for our models are smaller than the observational uncertainties on the data we compare them with. In total, less than 1 per cent of all MS trees are needed to obtain estimates consistent with the full set to this level of accuracy. In practice, we use MS-II trees to calculate abundances for galaxies with $\log M_\star/\mathrm{M}_\odot < 9.5$ and MS trees to calculate abundances for more massive galaxies.

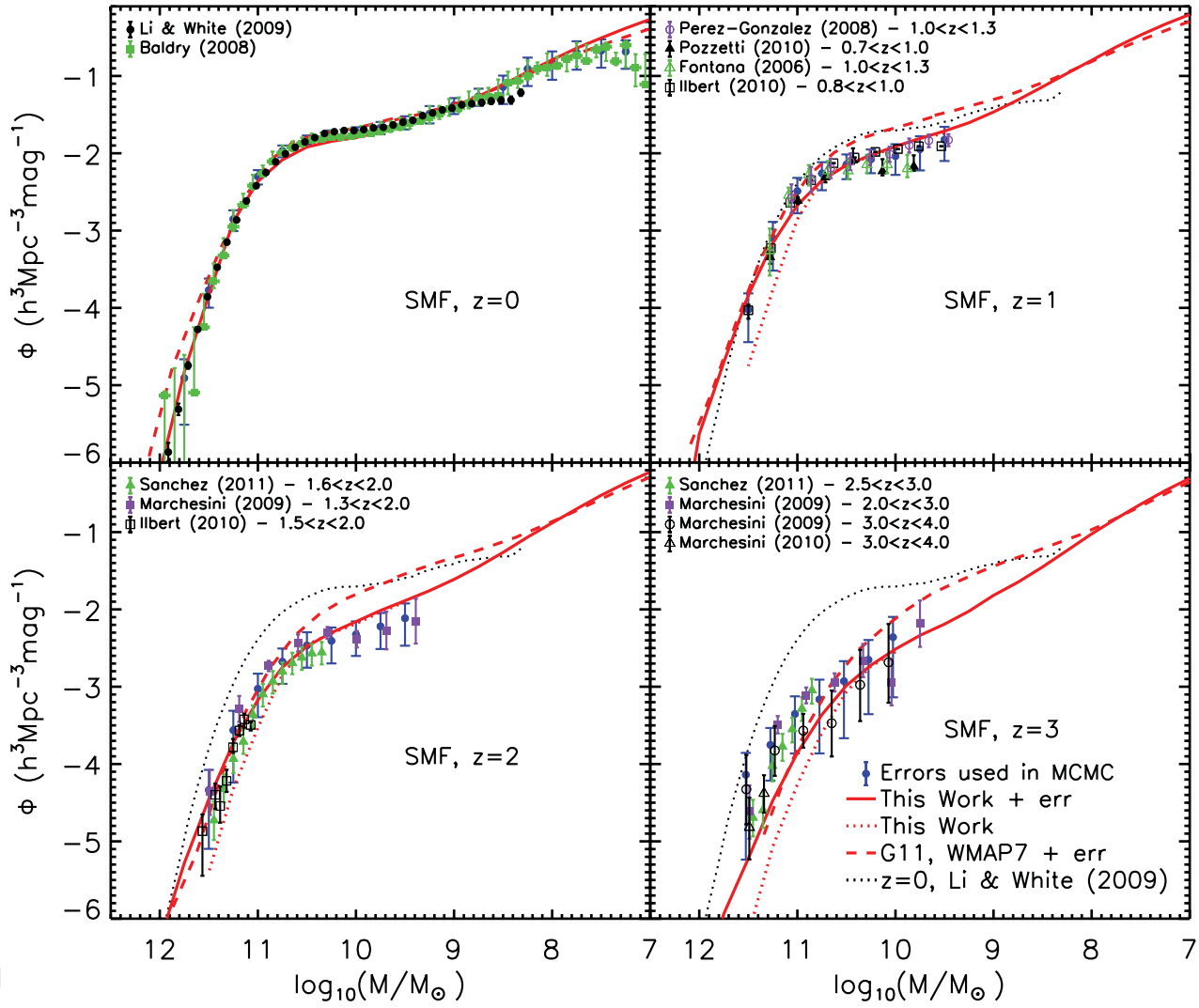
A representative and sufficiently large sample of haloes is needed at all redshifts at which the model will be compared with observation. Starting at $z = 0$, we select the 'optimum' number $n_i$ of FoF haloes at random in each mass bin. When weighted by $N_i/n_i$, these provide a representative sample of the full cosmic halo population. Hence, when combined with the same weights, the galaxies they contain at $z = 0$ reproduce the present-day luminosity and stellar mass functions (equation B3) and the galaxies contained by their progenitors at $z = 1$ (say) reproduce these same functions at that time. It is unclear, however, whether there will be enough progenitor haloes in each mass bin to satisfy our S/N requirements at $z = 1$. We check explicitly whether this is the case. If not, we select sufficient additional $z = 1$ haloes to restore the required precision. Their formation trees are then followed back together with those of the original haloes to $z = 2$ and the process is repeated. Note that the representative halo set and the associated trees and weights are determined once using a fiducial model. Galaxy formation is then followed sequentially from high to low redshift in this predefined

set of trees for every parameter set in our MCMC analysis. We have verified explicitly that these procedures produce luminosity and stellar mass function estimates at all redshifts which reproduce the functions for the full simulations at least as accurately as inferred from the simplified analysis above. This is true for the region in parameter space explored by the MCMC and for the best-fitting model in particular.

## APPENDIX C: OBSERVATIONAL ERRORS

In this appendix we discuss in more detail our method for combining individual data sets to construct observational constraints with realistic error bars for each of the population properties used in the MCMC sampling presented in Section 5. This is a crucial aspect of the MCMC analysis as the allowed regions in parameter space and the goodness of a fit to multiple properties depend strongly on the observational uncertainties. While errors based on counting statistics and cosmic variance are given in most observational studies it is much harder to get a realistic assessment of systematic uncertainties. These can show up as apparent inconsistencies between different determinations of the same population property, and failing to account for them can jeopardize a meaningful comparison with theoretical predictions.

Here we follow Henriques et al. (2009) and Henriques & Thomas (2010), using multiple determinations of each observational property and taking the scatter among them to indicate likely systematic uncertainties. For each observational property, at each redshift, we re-bin the individual estimates into a fixed set of broad bins. In each bin we then discount any determination with statistical uncertainties larger than 50 per cent and take a straight average of the rest as the constraint and their maximum and minimum values as its $\pm 1\sigma$ uncertainty. By not weighting the averages we attempt to account for the fact that systematic errors can affect large and small surveys in similar ways. This can be seen, for example, in the work




[b!]

**Figure C1.** Evolution of the stellar mass function from $z = 3$ to $z = 0$ as in Fig. 4, except with the data points for the individual underlying surveys also shown. These surveys are Baldry, Glazebrook & Driver (2008) and Li & White (2009) at $z = 0$, and the Great Observatories Origins Deep Survey (GOODS)-MUSIC (Fontana et al. 2006), *Spitzer* (Pérez-González et al. 2008), Marchesini et al. (2009), *Spitzer*-Cosmic Evolution Survey (COSMOS; Ilbert et al. 2010), NOAO Extremely Wide-Field Infrared Imager (NEWFIRM; Marchesini et al. 2010), zCOSMOS (Pozzetti et al. 2010) and COSMOS (Domínguez Sánchez et al. 2011) at higher redshift. All mass estimates at $z > 0$, except Domínguez Sánchez et al. (2011) have been shifted by $-0.14$ dex to convert from Bruzual & Charlot (2003) to Maraston (2005) stellar populations (Domínguez Sánchez et al. 2011). The $z = 0$ results of Li & White (2009) are repeated at all redshifts as a black dotted line.

of Marchesini et al. (2009, 2010) for the stellar mass functions and Cirasuolo et al. (2010) for the *K*-band luminosity function. Counting errors are swamped in their overall error budget by combined uncertainties from SED fitting assumptions, photometric redshift errors, photometric errors, extrapolations from the observed photometry and cosmic variance.

Finally, we harmonize the sizes of error bars to avoid dramatic changes between bins, for example, where there is a change in the number of surveys included. Typically, we re-size the error bar to the value of the next consecutive bin towards the knee of the function. More surveys are normally available towards the knee and their counting errors are usually smaller there, making systematics estimates more robust in this region. Clearly our procedure is an art rather than a quantitative science, but we believe it gives realistic estimates of the overall level of uncertainty given the information available.

Our adopted constraints are shown together with the individual data sets on which they are based in Fig. C1 for the stellar mass functions, in Fig. C2 for the *K*-band luminosity functions and in Fig. C3 for the *B*-band luminosity functions. The constraints are shown as blue dots with error bars while other data points represent observational estimates from the individual surveys. Theoretical predictions are shown as red lines. For the stellar mass function, the most massive bin at $z = 0$ and the most massive and least massive bins at $z = 1$ have been re-sized according to the method described in the last paragraph. The same was done for the faintest bin in the $z = 3$ *B*-band luminosity function.

We emphasize that despite our attempt to estimate uncertainties from the data, our method still involves arbitrary judgements about the quality of data and the way in which different surveys should be combined. As a result, formal levels of agreement between theory and observation should be treated with caution.

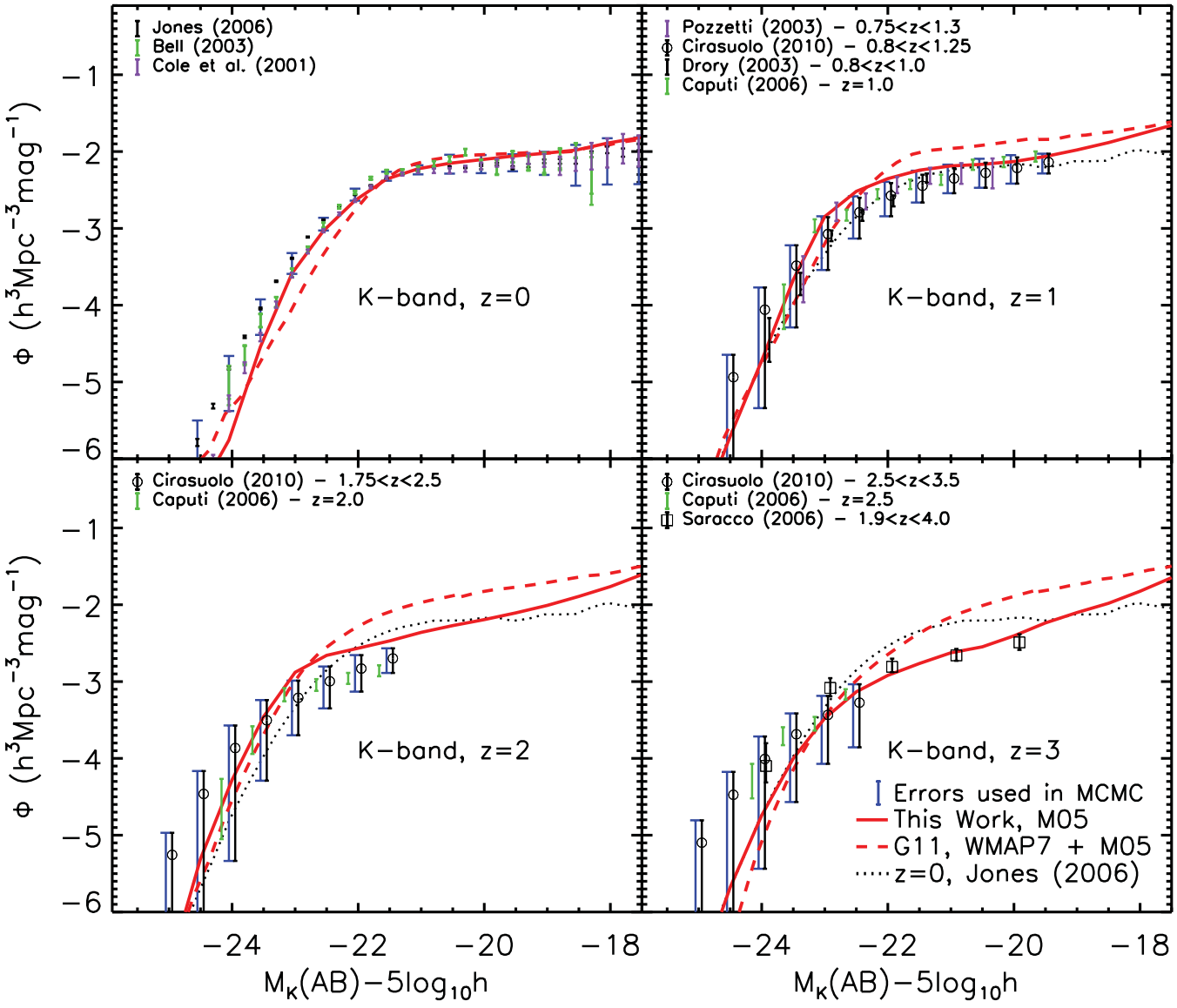



**Figure C2.** Evolution of the rest-frame *K*-band luminosity function from $z = 3$ to $z = 0$ as in Fig. 5, except that data points for the individual surveys are shown. These are three different 2MASS-based determinations at $z = 0$ (Cole et al. 2001; Bell et al. 2003; Jones et al. 2006), and determinations at higher redshift from Munich Near-IR Cluster Survey (MUNICS; Drory et al. 2003), the K20 survey (Pozzetti et al. 2003), the *Hubble Deep Field*-South (HDF-S; Saracco et al. 2006), GOODS-*Chandra Deep Field*-South (CDF-S; Caputi et al. 2006) and the UKIRT Infrared Deep Sky Survey-Ultra Deep Field (UKIDSS-UDF; Cirasuolo et al. 2010). The $z = 0$ results of Jones et al. (2006) are repeated at all redshifts as a black dotted line.

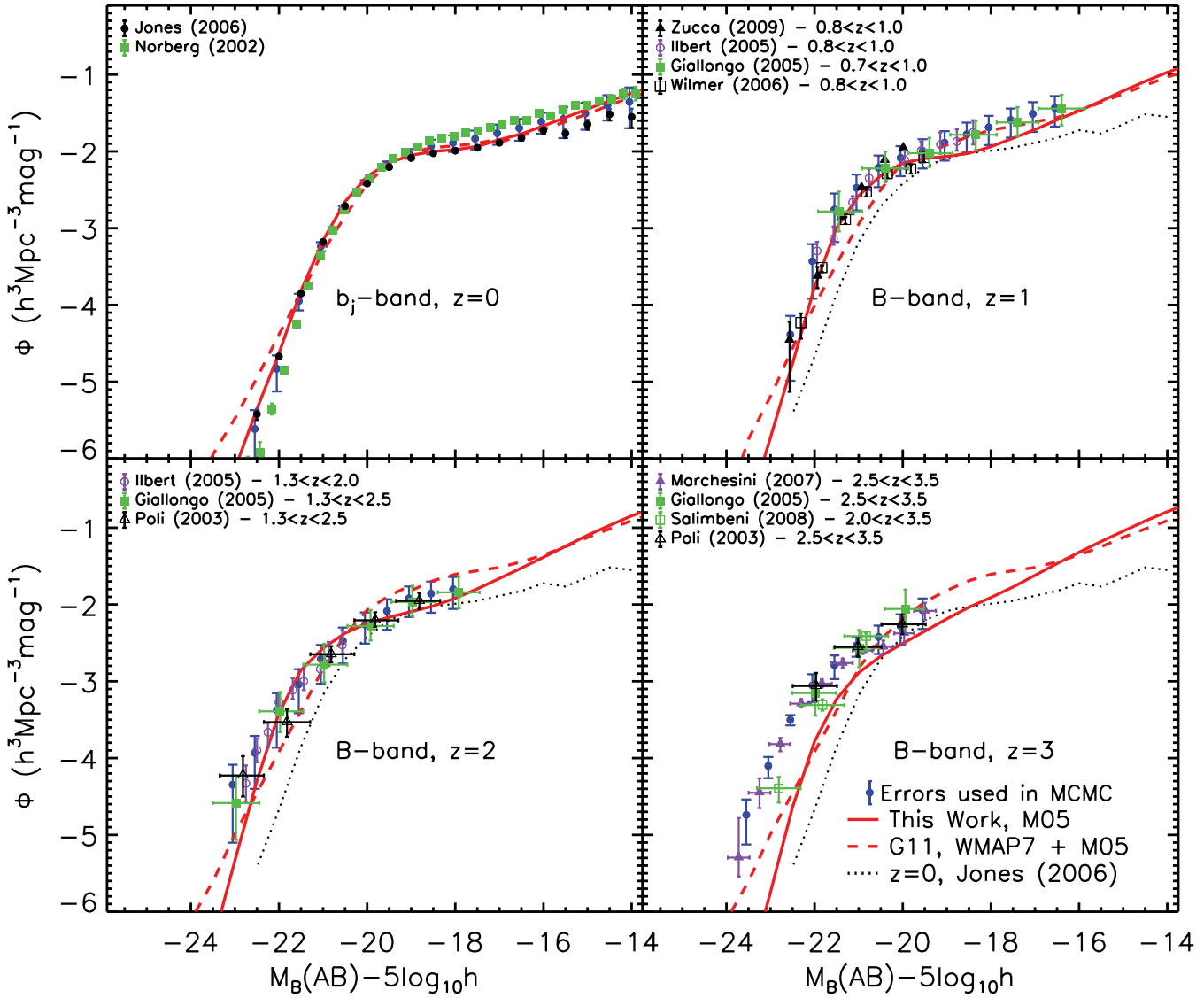



**Figure C3.** Evolution of the rest-frame *B*-band luminosity function from $z = 3$ to $z = 0$ as in Fig. 6, except that data points for the individual surveys are shown. At $z = 0$, these are the 2dF Galaxy Redshift Survey (2dFGRS; Norberg et al. 2002) and the 6dF Galaxy Redshift Survey (6dFGRS; Jones et al. 2006). At higher redshift, they are for HDF-S (Poli et al. 2003), HDF-N (Giallongo et al. 2005), VIMOS-VLT Deep Survey (VVDS; Ilbert et al. 2005), Deep Extragalactic Evolutionary Probe 2 (DEEP2; Willmer et al. 2006), GOODS-MUSYC plus FIRES (Marchesini et al. 2007), GOODS-MUSYC (Salimbeni et al. 2008) and zCOSMOS (Zucca et al. 2009). The $z = 0$ results of Jones et al. (2006) are repeated at all redshifts as a black dotted line.

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