Theoretical Modelling of Gas Cooling and Feedback in Galaxy Formation

HOU, JUN

Theoretical Modelling of Gas Cooling and Feedback in Galaxy Formation

Jun Hou

A Thesis presented for the degree of
Doctor of Philosophy

Institute for Computational Cosmology
Department of Physics
Durham University
United Kingdom

May 2017

Theoretical Modelling of Gas Cooling and Feedback in Galaxy Formation

Jun Hou

Abstract

Semi-analytical (SA) galaxy formation models have wide applications, and they are complementary to hydrodynamical simulations, which are more physically detailed but also much more computationally expensive. It is important to make semi-analytical models as physical as possible for the robustness of their applications. In this work we try to improve the modelling of two important processes, supernova (SN) feedback and gas cooling, in the SA model GALFORM.

We first improve the SN feedback recipe in a phenomenological way, using the constraints from four observations, including the Milky Way (MW) satellite galaxy luminosity function, the faint end of the field galaxy luminosity function, the redshift at which the universe was half reionized and the stellar metallicity of the MW satellites. We find that these observations favour a SN feedback model in which the feedback strength evolves with redshift. We further apply this improved model to investigate some details of reionization.

We then develop a new, more physical model for gas cooling in halos in semi-analytical models. We compare this new cooling model with a cosmological hydrodynamical simulation with stripped-down galaxy formation physics running with the grid-based moving mesh code AREPO, along with two previous models (GFC1 and GFC2) in GALFORM and the models in L-GALAXIES and MORGANA. We find that generally all SA models predict cumulative cool masses close to the simulation, but the mass cooling rates in low redshift massive halos are overestimated. These SA models overpredict the specific angular momenta of the cool gas for low mass halos, while for low redshift massive halos, the predictions from the new cooling model generally agree better with the simulation than the earlier SA cooling models. We also use the simulation to investigate gas cooling in individual halos in more detail.

Declaration

The work in this thesis is based on research carried out at the Institute for Computational Cosmology, the Department of Physics Durham University, between September 2013 and April 2017. Under the supervision of Prof. Cedric Lacey and Prof. Carlos S. Frenk. Chapter 3 has been published in the form of refereed paper,

- *Constraining SN feedback: a tug of war between reionization and the Milky Way satellites* Hou, Jun; Frenk, Carlos. S.; Lacey, Cedric G.; Bose, Sownak, 2016, MNRAS, 463, 1224

No part of this thesis has been submitted elsewhere for any other degree or qualification. It is all my own work unless referenced to the contrary in the text.

Acknowledgements

I would like to thank my supervisors Prof. Cedric Lacey and Prof. Carlos S. Frenk for their supervision. I would also like to thank Dr. John Helly, Dr. Lydia Heck and all other members of the COSMA support team for their help about my various computational issues. Finally I would like to thank Dr. Matthieu Schaller, Dr. Peter Mitchell, Dr. Violeta Gonzalez-Perez and Mr. Sownak Bose for their helpful discussions.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Galaxy formation is an important topic in modern astrophysics. This is firstly because that galaxies are important and interesting astronomical objects, and it is the task of astrophysics to understand their features and evolutions. Secondly, in observations galaxies serve as tracers of the underlying large-scale matter distribution, which is important for constraining cosmology; a better understanding of galaxy formation would help to get more accurate knowledge of cosmology. Thirdly, many small scale processes, such as star formation and black hole accretion, are deeply involved in galaxy formation, and it provides general background for the studies focusing on these processes. Despite its importance, galaxy formation is currently still poorly understood, mainly because of its complexity.

Currently there are two ways to study galaxy formation, namely through observations or theoretical modelling. Observing distant galaxies reveals the properties of galaxy populations in the past, because it takes time for the light from these galaxies to cross these huge distances and reach observers. Observations are directly related to the galaxies in the real universe, but indirectly to the physical processes driving galaxy formation and evolution. On the other hand, theoretical modelling of galaxy formation is based on more direct considerations of physical principles, but, if the modelling is not physical enough then it does not provide information about galaxy formation in the real universe. This work focuses on the theoretical side and tries to improve certain aspects of the theoretical modelling of galaxy formation, and through it, some better understanding of galaxy formation can be reached.

In this chapter, a brief review of the background cosmology (ΛCDM) for galaxy formation is given in §1.1, and some more details about the role of galaxy formation in constraining cosmology are also given there. Then the general picture of galaxy formation in the ΛCDM cosmology is provided in §1.2. After that, an overview of galaxy formation modelling methods is given in §1.3. The complexity of galaxy formation and difficulties of modelling it are also discussed in more detail in this section. Finally, the outline of this thesis is provided in §1.4.

1.1 ΛCDM Cosmology

The ΛCDM cosmology is currently our standard cosmology. According to it, the evolution of the Universe starts from a Big-Bang singularity. After this, it is believed that there was a stage called inflation during which the Universe expanded exponentially. At the end of inflation, the observable universe becomes very close to spatially flat if not exactly flat, and the material in it is almost homogeneous and isotropic. The quantum fluctuations on very small scales before inflation were stretched to macroscopic scales during inflation and this left very small but non-zero perturbations in the energy density.

After inflation the Universe continues to expand, and there are three important kinds of energy densities in the Universe, namely, radiation (relativistic), matter (non-relativistic) and cosmological constant, or vacuum energy (Λ). In the matter budget, the normal matter, or baryonic matter, only takes about one fifth; the rest is thought to be non-baryonic cold dark matter (CDM). Originally the energy density is dominated by the radiation. But the energy density of radiation decreases with cosmic expansion faster than the non-relativistic matter, so later the Universe transfered from radiation dominated to matter dominated. This matter dominated stage covers a large fraction of cosmic history. The energy density of matter also goes down as the Universe expands, while the energy density of Λ stays constant, so at late times, the Universe goes into a Λ dominant stage. Today, the total energy density of the Universe has a 70% contribution from Λ and about 30% contribution from matter; the energy density of radiation is negligible.

This cosmology has observational support. The observed cosmic microwave background (CMB) is very isotropic and with a spectrum very close to a black body with temperature about 3 K (e.g. Penzias & Wilson, 1965). This is in good agreement with the hot Big Bang cosmology model. Further, observations of the small temperature fluctuations in CMB can determine cosmological parameters such as Ω_Λ, Ω_M and Ω_b, which are the energy densities in Λ, total matter and baryonic matter respectively. Current observations generally suggest that $\Omega_\Lambda > 0$ and $\Omega_\mathrm{b} \ll \Omega_\mathrm{M}$, which supports the ΛCDM cosmology (e.g. Komatsu et al., 2011; Planck Collaboration et al., 2014). Apart from the CMB observations, there are other independent observations that support ΛCDM cosmology. The measurement of cosmic expansion rate through Type Ia supernovae indicates $\Omega_\Lambda > 0$ (e.g. Riess et al., 1998; Perlmutter et al., 1999). The observation of rotation curves of giant disk galaxies indicates there are huge amounts of dark matter surrounding these galaxies (e.g. Rubin et al., 1980). The abundances of the light elements generated during the Big-Bang nucleosynthesis supports the non-baryonic nature of the dark matter (e.g. Alpher et al., 1948; Wagoner, 1973; Cyburt et al., 2008). The galaxy cluster abundance (e.g. Vikhlinin et al., 2009; Rozo et al., 2010), gravitational weak lensing observations (e.g. Massey et al., 2007) and large-scale structure (LSS) data (e.g. Cole et al., 2005; Hamann et al., 2010) also support this cosmology.

The weak lensing and LSS observations constrain cosmology mainly through the measurement of matter power spectrum. This power spectrum concerns the distribution of all matter, but only the luminous baryonic matter (galaxies) can be observed. Thus, galaxies become the tracers of the underlying matter distribution. However, this tracer is biased (e.g. Kaiser, 1984; Davis et al., 1985). Also redshift surveys typically provide magnitude limited samples, and the redshift distribution of the sample galaxies also affects the estimation of matter power spectrum. Modern surveys usually use large mock galaxy catalogs to estimate this effect (e.g. Laureijs et al., 2011). Both of these two effects involve galaxy formation at some level, and a better galaxy formation model helps to separate them from the cosmology, focusing the constraining power on the latter.

1.2 Galaxy Formation within ΛCDM Cosmology

The small perturbations generated during inflation seed structure formation. An initially slightly overdense region would continuously accrete matter under gravity and thus magnifies the overdensity. This growth finally leads to the formation of highly nonlinear structure.

In the nonlinear regime, both dark matter and gas undergo violent relaxation processes and turn the bulk flow velocity into random velocity. This is called virialization. The collisional gas is virialized through an accretion shock, while the collisionless dark matter may be virialized through rapid changes in the gravitational potential (e.g. White, 1996). The final result of the virialization is a dark matter halo with a hot gaseous halo, both of them in quasi-equilibrium, and roughly spherical.

These halos would continuously accrete new material and grow. It is possible that a smaller halo falls into a larger halo, and then this small halo becomes a substructure, or subhalo of the large halo; this is called a halo merger. As the subhalo moves in its host halo, it is gradually dissolved by the tidal forces from its host halo; its hot gas may additionally be swept by the pressure from the hot gas in the host halo. The accretion of new material also delivers angular momentums to the dark matter and gas. This angular momentum may originate from the tidal torques induced by the large scale matter distribution.

The above picture covers most parts of the nonlinear growth of the dark matter component. As time goes by, the dark matter halo becomes gradually more massive with more subhalos. For the gas, however, there are additional complexities because of its electromagnetic interactions. This kind of interaction allows gas to give away its thermal energy gained during virialization through radiation, or, in short, to cool down. The reduction of thermal energy also reduces the gas pressure, and the cooled gas falls towards the bottom of the gravitational potential well created by the dark matter halo. The gas' angular momentum finally stops the infall and lead to the formation of a rotationally supported gaseous disk in the central region of a dark matter halo. Fragmentation can happen in the disk, leading to the formation of giant molecular clouds and eventually stars. In this way, a galaxy is formed in the halo centre. If the angular momentum of the gaseous disk is not very high, then it may

be unstable. In this case, huge amounts of low angular momentum gas flows towards the disk central region, and triggers violent star formation (starbursts) and/or black hole accretion there. Disk instability also leads to the formation of pseudo-bulges (e.g. Kormendy & Kennicutt, 2004).

Note that once the gaseous halo begins to cool, which happens when the gas is hotter than 10^4 K (or 10^3 K if molecular hydrogen cooling is allowed, see e.g. Benson (2010)), it may change the nature of further gas accretion. The extended hot gas halo expected with ineffective cooling can continue to exist only if the cooling is not very fast. In the fast cooling regime, the dark matter halo is mainly filled by the cold gas generated by this rapid cooling (White & Frenk, 1991).

It is possible that a accreted small halo also forms a galaxy in its centre. After infall, this galaxy becomes a satellite of the central galaxy of the host halo. The satellite galaxy loses its orbital angular momentum through dynamical friction (e.g. Chandrasekhar, 1943). If this loss is efficient enough, then the satellite galaxy would fall onto the central galaxy and merge with it. The result of a galaxy merger depends on the mass ratio of the two galaxies. If the mass ratio is very small, namely the accreted galaxy is much less massive than the central galaxy, then the merger simply increases the mass of the central galaxy, while if the ratio > 0.1, and the galaxies are gas rich, then a starburst and/or black hole accretion can be triggered (e.g. Cox et al., 2008), and for even larger ratio (> 0.3), the merger can also strongly change galaxy morphology, turning the rotationally supported disk into a random motion supported ellipsoid. These mergers with mass ratio > 0.1 lead to the formation of classical bulges or elliptical galaxies (e.g. Toomre & Toomre, 1973; Bournaud et al., 2005).

The massive stars formed evolve into supernovae. The supernova explosion is very energetic and can expel gas out of the galactic disk, or even out of the dark matter halo. Through this, it regulates further star formation and thus is crucial for the abundance of faint galaxies (e.g. Larson, 1974; White & Rees, 1978; Dekel & Silk, 1986; White & Frenk, 1991). This process is called supernova feedback (SN feedback).

Apart from energy, supernovae also eject heavy elements (metals) formed during

the explosion into the interstellar medium (ISM) or intergalactic medium (IGM). These metals provide material for dust formation and they also significantly enhance the cooling rate of halo gas. This process is called as metal enrichment. Stellar winds may also eject some heavy elements formed in stars, and thus contribute to metal enrichment.

The accreting black hole may also release huge amounts of energy, which can quench star formation and gas cooling. This process is called AGN feedback. It is thought to be responsible for the formation of giant red galaxies. Because it suppresses gas cooling in massive halos, it regulates the abundance of bright galaxies, and solves the overcooling problem in galaxy cluster halos (e.g. Croton et al., 2006; Bower et al., 2006).

The UV photons from massive stars and/or accreting black holes can ionize the originally neutral IGM. This process is called reionization. Together with this ionization, the IGM is heated up to about 10^4 K. This heating suppresses the gas accretion and cooling in small dark matter halos, thus suppressing galaxy formation. It is another kind of feedback, called photonionization feedback (e.g. Couchman & Rees, 1986; Efstathiou, 1992; Thoul & Weinberg, 1996).

1.3 Modelling Galaxy Formation

The quantitative modelling of galaxy formation outlined in §1.2 is very challenging.

Firstly, this is because it involves many complex physical processes. The shock during gas accretion onto dark matter halos, the subsequent gas infall induced by cooling and the formation and fragmentation of a gaseous disk are complex hydrodynamical processes. Take a further look into details: turbulence may play a non-negligible role in the gaseous disk and giant molecular clouds, the magnetic field may also play a role in giant molecular clouds. All these details further complex the hydrodynamical problem. Apart from this, the calculation of reionization and cooling radiation in principle involves radiation transfer, another complicate physical process. Also, as the gas changes its temperature and density, it may transfer from ionized to neutral, atomic to molecular, or vice versa, and to accurately capture each

different phase, a complex chemical network is required. Finally, the stars and accreting black holes are all complex systems by themselves, and they have important roles in galaxy formation through various feedback and metal enrichment channels.

Secondly, the above mentioned processes happen on very different scales. Dark matter halos typically have sizes from several tens of Kpc to a few Mpc, while galactic disks typically have sizes about few Kpc; giant molecular clouds are typically about $5 - 200\,\mathrm{pc}$ (e.g. Murray, 2011), and the associated star formation happens on even smaller scales. Supernova explosions happen on stellar scales. For a $30\,\mathrm{M}_\odot$ star, the radius is about 8 solar radii or about $10^{-7}\,\mathrm{pc}$. Black hole accretion can also reach very small scales, i.e. of order of the Schwarzschild radius. For a $10^8\,\mathrm{M}_\odot$ black hole, the Schwarzschild radius is about $10^{-5}\,\mathrm{pc}$. In all, these processes cover more than ten orders of magnitudes on scale. The variety of spatial scales also leads to very different dynamical timescales. The dynamical timescale for dark matter halos is typically about few Gyrs, while the protostar contraction happens within $10^4\,\mathrm{yrs}$. Supernovae show significant brightness change in timescale about a month (e.g. Doggett & Branch, 1985), AGN have variabilities from few days to months (e.g. Peterson et al., 1999), and these can be a rough proxy for dynamical times associated with supernova explosions and black hole accretion. Detailed calculation across these huge spacial and temporal scale ranges is extremely difficult.

Because of these difficulties, modern galaxy formation models have to use various approximations. Different levels of approximations allow the model produce different amounts of detail on galaxy formation, also with different computational requirements. Below I review two major modelling methods.

1.3.1 Hydrodynamical Simulations

This kind of method tries to follow the hydrodynamical state of the gas by numerically solving the equations governing gas flows. The accuracy in capturing these states depends on the resolution of these simulations. Currently simulations of individual galaxies can reach spatial resolution of about 20 pc (e.g. Rosdahl et al., 2017), while simulations in a representative cosmological volume usually have rougher resolution, i.e. about 1 Kpc (e.g. Schaye et al., 2015; Vogelsberger et al., 2014). With

these resolutions, the gas accretion onto dark matter halos, as well as the infall onto central disks can be modelled in detail, while the fragmentation in the disk can only be marginally resolved. The star formation in giant molecular clouds cannot be properly resolved even in the simulations of individual galaxies, nor can supernovae explosions and black hole accretion. These poorly resolved processes cannot be captured by directly solving hydrodynamical equations; instead, they are modelled through some simple and approximate recipes called sub-grid recipes. These recipes can be inspired by observations, or by simple physical models, and usually contain free parameters, which are fixed through the calibration of certain simulation predictions against corresponding observations. Processes other than hydrodynamics, such as radiation transfer, ionization, molecule formation are also modelled by sub-grid recipes. All these sub-grid recipes appear as additional source terms in the hydrodynamical equations. The evolution of stellar population, as well as the chemical pattern of metal enrichment, are modelled using tables generated by stellar evolution models.

By numerically solving the hydrodynamical equations, this method can provide many details of the IGM (e.g. Nelson et al., 2016) and galactic internal structure (e.g. Bahé et al., 2016). Also, since the hydrodynamical equations directly include the interaction between gas (pressure), and the interaction between baryonic and dark matters (gravity), it is straightforward to study details of various environmental effects (e.g. Marasco et al., 2016) and baryon effects on the dark matter (e.g. Schaller et al., 2015a,b). But solving hydrodynamical equations is computationally expensive. With current computing power, hydrodynamical simulations cannot generate very large statistical samples, especially for very rare massive halos ($M_{\rm halo} > 10^{14}\,{\rm M}_{\odot}$). These samples are important for building mock catalogs for cosmological observations. Also, the high computational cost limits the ability to derive the best fit for the free parameters in sub-grid recipes or investigate the general behavior of these recipes in parameter space.

1.3.2 Semi-analytical Models

Semi-analytical models (SA models) first separate the whole galaxy formation process into several interrelated sections, and then develop simple, approximate recipes for each section. Just as the sub-grid recipes for hydrodynamical simulations, these recipes usually contain adjustable parameters. These parameters should be fixed through fitting to certain observations. After this calibration, SA models can make predictions for other independent observations.

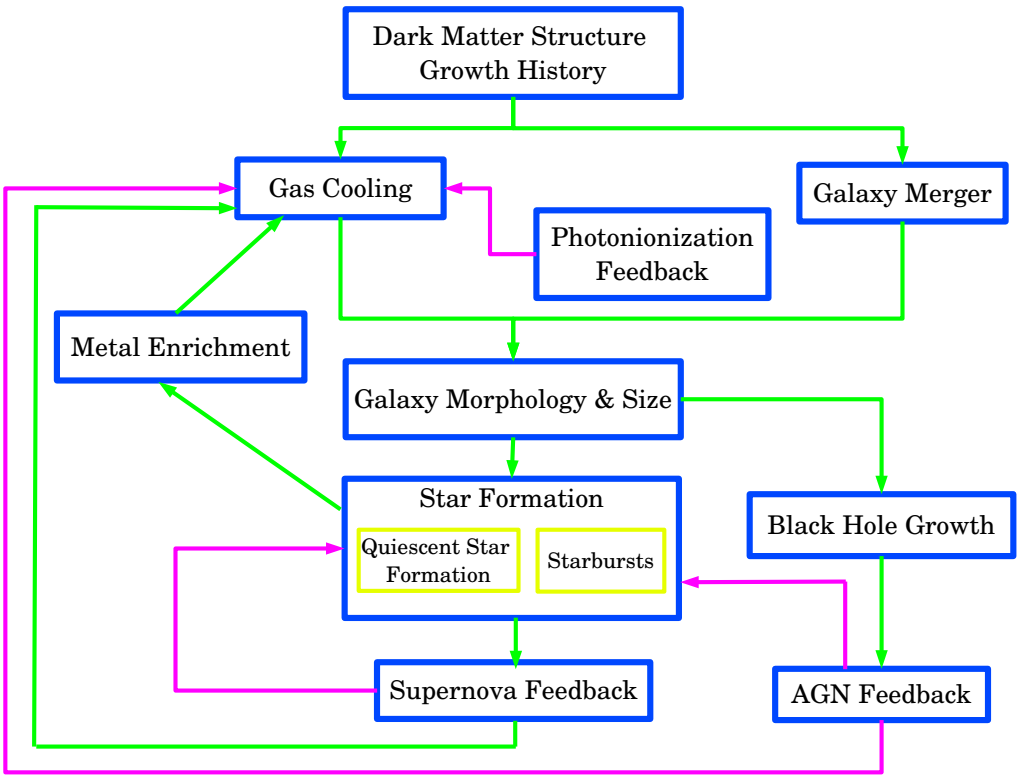

Figure 1.1: General structure of SA models. The boxes indicate sections representing different physical processes. The green arrows show the normal connections between processes, while the feedbacks are specially shown by magenta arrows.

Fig. 1.1 shows the general structure of an SA model. The backbone of an SA model is the growth history of dark matter structures, or halo merger trees. Based on this, halo properties such as virial radius, virial temperature and halo spin can be derived. By assuming the baryons follow dark matter when outside of a halo, this

also gives the rate of accretion of gas onto a halo. This information is important for determining halo hot gas properties, which is the first step of cooling calculation. The merger trees also provide a list of subhalos of each halo, based on which the galaxy merger is calculated.

The gas cooling section usually determines the amount of gas and angular momentum delivered to the central galaxy. These are the basis of disk size determination and further, for the disk instability test. These tasks are done by the galaxy size and morphology section. It also includes the modelling of effects of galaxy mergers.

Once the galaxy sizes and morphologies are determined, further processes happening inside a galaxy can be considered. These are mainly covered by two categories of sections. One of them is about star formation and related effects. This includes sections about star formation, SN feedback and metal enrichment. Typically, the star formation section can be further divided into quiescent star formation and starbursts. The metal enrichment section may affect gas cooling through the enhancement of metal concentration in the hot gas. The SN feedback section directly suppresses star formation, and it would also be entangled with the gas cooling section if the ejected gas is accreted by the halo hot gas. The other one involves black holes. This typically includes black hole growth section and AGN feedback section. AGN feedback can directly suppress gas cooling, and may also suppress star formation. The photonionization (induced by reionization) section also has negative feedback on gas cooling.

The halo merger trees are usually derived from N-body simulations or the Monte Carlo method based on the extend Press-Schechter theory (e.g. Lacey & Cole, 1993). The detailed recipes for other sections vary from one SA model to another. Generally speaking, there are two ways to construct these recipes. One is more phenomenological, including the constraints to a recipe from relevant observations. For example, an effective AGN feedback recipe can be constructed by requiring that it reproduces the correct bright end of the galaxy luminosity function. The advantage of this method is that it is relatively simple, not involving many details of a physical process. The disadvantage is that it is indirectly related to the physics of a process, and has a relatively strong model dependence. To see the latter point, consider

the example of AGN feedback. The prediction for the galaxy luminosity function involves the whole SA model, and any change in the recipes for other sections may affect the AGN feedback recipe. The other one is more physical, which builds the recipe directly from the physical picture and approximation. This method is less model dependent, but of course the construction is more complex and difficult.

The interrelated sections and associated recipes adopted by SA models are only a coarse grain representation of the hydrodynamical equations solved by numerical simulations. Thus, SA models cannot provide many details of galaxies, but only give global properties such as the total stellar mass, total cold gas mass etc. But this rough representation significantly reduces the computational cost, and thus allows it to build large statistical samples. This property makes SA models a good tool for building mock catalogs and investigating model parameter space (e.g. Bower et al., 2010). Also, SA models are more flexible than hydrodynamical simulations, so they can be used for modelling tests for hydrodynamical sub-grid recipes. It is important to make these SA model recipes as physical as possible to guarantee the reliability of the applications.

The original idea of SA models dates back to White & Rees (1978), and by now there have been many SA models, which share the abovementioned general structure but differ in details. Among them, there are several major SA models, namely the Durham SA model GALFORM (e.g. Cole et al., 2000; Baugh et al., 2005; Bower et al., 2006; Gonzalez-Perez et al., 2014; Lacey et al., 2016), the Munich SA model L-GALAXIES (e.g. Springel et al., 2001; Croton et al., 2006; Guo et al., 2011; Henriques et al., 2015) and the MORGANA SA model (Monaco et al., 2007; Viola et al., 2008). There are also other SA models, for example, the GALACTICUS SA model (Benson, 2012), the GALICS model (Hatton et al., 2003; Cattaneo et al., 2006, 2017) and the SANTACRUZ SA model (e.g. Somerville & Primack, 1999; Somerville et al., 2008; Hirschmann et al., 2012; Porter et al., 2014).

1.4 Thesis Outline

In this thesis we try to improve several recipes for semi-analytical galaxy formation models.

Chapter 2 gives an overview of the Lacey16 model (Lacey et al., 2016), which is the newest version of the Durham semi-analytic galaxy formation model GALFORM. This forms the basis of the improvements made in this thesis.

Chapter 3 presents improvements to the SN feedback recipe through the phenomenological method introduced in §1.3.2, and the application of GALFORM with this new recipe to investigate reionization.

Chapter 4 describes the construction of a new gas cooling recipe. This is done through the more physical way introduced in §1.3.2. This also provides a comparison of this new recipe with all other recipes currently used in major semi-analytical models.

Chapter 5 further compares the new gas cooling recipe with hydrodynamical simulations. Through this comparison, we can obtain not only an assessment of the accuracy of this recipe, but also some detailed physical insight about gas cooling.

Finally, a summary and a discussion of possible future work are given in Chapter 6.

Chapter 2

GALFORM Model

The Durham semi-analytical galaxy formation model GALFORM is one of the main SA models. It has experienced several major developments, e.g. Cole et al. (2000); Baugh et al. (2005); Bower et al. (2006) and Gonzalez-Perez et al. (2014). The Lacey16 model (Lacey et al., 2016) is the newest version of GALFORM. It serves as the starting point for the work in this thesis. This chapter gives a brief overview of its recipes of the sections introduced in §1.3.2. Some further details of this model will be mentioned in later chapters when relevant. This thesis also involves some other SA models, and they also will be introduced at appropriate time in later chapters.

2.1 Halo Merger Trees and Dark Matter Halo Structure

GALFORM can use two kinds of merger trees, namely trees from N-body simulations and Monte Carlo trees.

The N-body merger trees used by GALFORM are constructed by first identifying dark matter structures using the SUBFIND code (Springel et al., 2001) for each output snapshot, then matching the most bound particles to link the identified structures at different snapshots to form subhalo merger trees, and finally grouping subhalos into Dhalos to form the Dhalo merger trees (e.g. Helly et al., 2003a; Jiang et al., 2014). From N-body simulations it is also straightforward to measure the angular momentum of dark matter halos, which is important for calculating disk

sizes. The disadvantage of N-body merger trees is that N-body simulations are still computationally expensive, and so cannot be used to calculate merger trees for very large samples or to very high mass resolution.

The Monte Carlo merger trees are based on the conditional halo mass function derived through the extended Press-Schechter (EPS) theory (e.g. Bond et al., 1991; Lacey & Cole, 1993). Treating this function as a probability function, one can split a given base node halo at low redshift into many progenitors when going to higher redshifts, and thus build a merger tree for this base node. The conditional halo mass function derived from EPS theory deviates somewhat from that obtained from N-body simulations. Parkinson et al. (2008) introduced further modifications of the EPS conditional halo mass function to make the Monte Carlo merger trees be statistically closet matched to N-body merger trees, at least in the redshift range $z = 1-4$ and the mass range $M_{\rm halo}(z=0) = 10^{12}-10^{15}h^{-1}\,{\rm M}_{\odot}$. GALFORM typically uses this improved method to build Monte Carlo merger trees. This type of merger tree is much cheaper to compute than N-body merger trees, so one can easily build very large samples or go to very high mass resolution, but some other associated properties, such as halo angular momenta, have to be derived by using separate methods.

With the halo mass and redshift provided by merger trees, GALFORM further calculates halo internal structure, which is important for further calculations such as gas cooling and galaxy sizes. The first two quantities determined in GALFORM are the halo virial radius, $r_{\rm vir}$, and virial velocity, $v_{\rm vir}$, which are defined as

$$r_{\rm vir} = \left(\frac{M_{\rm halo}}{4\pi\Delta_{\rm vir}\bar{\rho}/3}\right)^{1/3} \tag{2.1.1}$$

$$v_{\rm vir} = \sqrt{\frac{GM_{\rm halo}}{r_{\rm vir}}}, \tag{2.1.2}$$

where $M_{\rm halo}$ is the halo mass, $\Delta_{\rm vir}$ is halo over density that calculated from the spherical top-hat collapse model (e.g. Eke et al., 1996), $\bar{\rho}$ is the mean matter density and G is the gravitational constant. Note that $\Delta_{\rm vir}$ and $\bar{\rho}$ generally depend on redshift. Also note that GALFORM typically uses these two equations to first derive $v_{\rm vir}$ for a given $M_{\rm halo}$, and then derives $r_{\rm vir}$ through $v_{\rm vir}$ by using Eq(2.1.2).

GALFORM assumes that the dark matter density profile is NFW (Navarro et al.,

1997), which is described as

$$\rho_{\rm DM}(r) \propto \frac{1}{(r/r_{\rm s})(1+r/r_{\rm s})^2}, \tag{2.1.3}$$

where $\rho_{\rm DM}(r)$ is the dark matter density at radius r from the halo center, and $r_{\rm s}$ is the scale radius of this profile. A related quantity is the so-called halo concentration, $c_{\rm NFW} = r_{\rm vir}/r_{\rm s}$. GALFORM calculates $c_{\rm NFW}$ from the halo mass-concentration relation given in Navarro et al. (1996). With $c_{\rm NFW}$ known, the normalization of $\rho_{\rm DM}(r)$ is fixed by $M_{\rm halo}$.

In Lacey16 model, and almost all other versions, $v_{\rm vir}$ and $c_{\rm NFW}$ are only updated at so-called halo formation events. These events are defined as follows. The appearance of a halo without progenitors is a halo formation event, and the instant a halo becomes twice or more massive than its progenitor at the previous most recent halo formation event is a new halo formation event.

The halo angular momentum is calculated from the halo spin $\lambda_{\rm halo}$, which is defined as

$$\lambda_{\rm halo} = \frac{J_{\rm halo}|E_{\rm halo}|^{1/2}}{GM_{\rm halo}^{5/2}}, \tag{2.1.4}$$

where $J_{\rm halo}$ is the total angular momentum of the halo and $E_{\rm halo}$ its binding energy. N-body simulations show that statistically $\lambda_{\rm halo}$ obeys a log-normal distribution, which is insensitive to halo mass and redshift (e.g. Warren et al., 1992; Gardner, 2001). GALFORM adopts a log-normal distribution to generate halo spins. The spin is randomly picked according to this log-normal distribution for each halo at halo formation events, and is propagated from progenitors to descendants between halo formation events. The log-normal distribution in GALFORM has median $\lambda_{\rm halo,med} = 0.039$ and dispersion in $\ln(\lambda_{\rm halo})$, $\sigma_{\lambda_{\rm halo}} = 0.53$ (Cole & Lacey, 1996).

In Chapter 4 we will describe an improvement in the assignment of $v_{\rm vir}$, $c_{\rm NFW}$ and $\lambda_{\rm halo}$, which removes the dependence on the artificial halo formation events.

2.2 Gas Cooling

The cooling model used in the Lacey16 model is the GFC1 cooling model introduced in Bower et al. (2006) and discussed in Chapter 4. This model assumes that the gas

in a dark matter halo is initially hot, settles in a hot gaseous halo, and any newly accreted gas also joins this hot halo. The later cooling depends on the properties of this gas halo.

GALFORM assumes that this hot halo has a single temperature and metallicity at each moment. The temperature is always the halo virial temperature $T_{\rm vir}$, which is calculated in GALFORM as

$$T_{\rm vir} = \frac{1}{2}\frac{\mu_{\rm m} v_{\rm vir}^2}{k_{\rm B}}, \tag{2.2.1}$$

where $\mu_{\rm m}$ is the mean molecular mass of the hot gas and $k_{\rm B}$ is the Boltzmann constant. While the metallicity $Z_{\rm hot}$ is given by

$$Z_{\rm hot} = \frac{M_{\rm z,hot}}{M_{\rm hot}}, \tag{2.2.2}$$

where $M_{\rm hot}$ is the mass of gas that is still hot by a given moment, while $M_{\rm z,hot}$ is the mass of metals in this hot gas. Initially $M_{\rm z,hot} = 0$ and is updated by the metal enrichment process described in §2.7.

The density profile of this gas halo is assumed to be the so-called β-profile

$$\rho_{\rm gas}(r) \propto \frac{1}{r^2 + r_{\rm core}^2}, \tag{2.2.3}$$

where $\rho_{\rm gas}(r)$ is the density of gas at radius r from halo centre and $r_{\rm core}$ is a parameter. Following Benson et al. (2003), $r_{\rm core}$ is set to be a fixed fraction of $r_{\rm vir}$, and in the Lacey16 model, $r_{\rm core} = 0.1 r_{\rm vir}$. This profile is assumed to include two kinds of gas, namely, the cold gas (including that has been turned into stars) that has cooled down since the last halo formation event, and the hot gas. The gas cooled down, then ejected by the SN feedback and reaccreted onto the hot gas halo is included in the hot gas. Thus, the total mass in this gas halo at a given moment can be expressed as $M_{\rm gas,tot} = M_{\rm hot} + M_{\rm cooled}$, where $M_{\rm cooled}$ is the mass of gas cooled down since the latest halo formation event, including the gas ejected by SN feedback but not yet be reaccreted onto the hot gas halo (see §2.6.1 for more details of SN feedback and this reincorporation). The value of $M_{\rm gas,tot}$ fixes the normalization of $\rho_{\rm gas}(r)$. So far, this profile is fully determined. Note that this determination means that the gas profile is largely fixed between halo formation events and reset at each halo formation event.

The angular momentum distribution of the gaseous halo is also needed by the cooling model. This cooling model assumes that the specific angular momentum distribution $j_{\rm gas}(r) \propto r$, and the normalization of $j_{\rm gas}(r)$ is determined by requiring the mean specific angular momentum of the gaseous halo is the same as that of the dark matter halo, with the latter determined through the halo spin $\lambda_{\rm halo}$.

Then the cooling model determines the mass and angular momentum delivered to the central galaxy by cooling. There are two factors affecting this delivery. One is the gas cooling rate, and the other one is the gravitational infall rate.

The progressing of cooling is represented by the so-called cooling radius $r_{\rm cool}(t)$, which is defined through

$$t_{\rm cool}(r_{\rm cool}(t)) = t - t_{\rm form}. \tag{2.2.4}$$

In this equation, $t - t_{\rm form}$ is the time since the last halo formation event. Since the gas halo is reset at each halo formation event, this is the time available for cooling by time t. $t_{\rm cool}(r)$ is the cooling time scale for gas at radius r, and is defined as

$$t_{\rm cool}(r) = \frac{3k_{\rm B}}{2\mu_{\rm m}} \frac{T_{\rm vir}}{\tilde{\Lambda}(T_{\rm vir}, Z_{\rm hot})\rho_{\rm gas}(r)}, \tag{2.2.5}$$

where $\tilde{\Lambda}(T_{\rm vir}, Z_{\rm hot})\rho^2_{\rm gas}$ gives cooling radiation rate per unit volume. The Lacey16 model, and other versions of GALFORM calculate $t_{\rm cool}$ based on the $\tilde{\Lambda}(T_{\rm vir}, Z_{\rm hot})$ provided in Sutherland & Dopita (1993).

The advancing of infall is described through another radius, i.e. the free-fall radius $r_{\rm ff}$, which is defined as

$$t_{\rm ff}(r_{\rm ff}) = t - t_{\rm form}, \tag{2.2.6}$$

where $t_{\rm ff}(r)$ is the free-fall time scale at radius r.

Then at a given moment t, the gas within $r_{\rm infall}(t) = \min[r_{\rm cool}(t), r_{\rm ff}(t)]$ would have enough time to both cool down and fall onto the central galaxy. This is the gas delivered by cooling. According to this consideration, for a time step $[t, t + \Delta t]$, the mass ($M_{\rm acc,gal}$) and angular momentum ($J_{\rm acc,gal}$) delivered by cooling to the central

galaxy are respectively given by

$$M_{\rm acc,gal} = 4\pi \int_{r_{\rm infall}(t)}^{r_{\rm infall}(t+\Delta t)} \rho_{\rm gas}(r) r^2 dr, \quad (2.2.7)$$

$$J_{\rm acc,gal} = 4\pi \int_{r_{\rm infall}(t)}^{r_{\rm infall}(t+\Delta t)} j_{\rm gas}(r) \times \rho_{\rm gas}(r) r^2 dr. \quad (2.2.8)$$

Note that here $r_{\rm infall}(t)$ can be directly used because the gas halo is assumed to be largely fixed between halo formation events. This gas halo would be more dynamical in some mode complex cooling models introduced in Chapter 4. In that case, $r_{\rm infall,pre}$ should be used instead. This radius corresponds to $r_{\rm infall}(t)$ but includes the effects induced by the dynamical adjustment of the gaseous halo.

2.3 Galaxy Mergers

Apart from cooling, galaxies can also gain matter through accreting other galaxies, a process known as galaxy mergers. Halo merger trees provide the subhalo list for each halo, and those galaxies contained in these subhalos are candidates for galaxy merger. Whether the merger happens by a given time t depends on the rate of orbital angular momentum loss. This loss is determined by the dynamical friction time scale $\tau_{\rm merger}$. If $\tau_{\rm merger} < t - t_{\rm infall}$, where $t_{\rm infall}$ is the time a galaxy becomes a satellite, then this satellite has merged with the central galaxy.

The Lacey16 model adopts the formula for $\tau_{\rm merger}$ from Jiang et al. (2008), which is a fit to cosmological gas simulations. Specifically

$$\tau_{\rm merger} = \frac{f(\epsilon)}{2C} \frac{M_{\rm cen}}{M_{\rm sat}} \frac{1}{\ln(1 + M_{\rm cen}/M_{\rm sat})} \left(\frac{r_{\rm circ}}{r_{\rm vir}}\right)^{1/2} \tau_{\rm dyn,halo}. \quad (2.3.1)$$

Here $M_{\rm cen}$ and $M_{\rm sat}$ are respectively the masses of the central and satellite galaxies, and both of them also include the mass of corresponding dark matter halo. $\tau_{\rm dyn,halo} = r_{\rm vir}/v_{\rm vir}$ is the halo dynamical time scale. $C = 0.43$, a constant, and $f(\epsilon) = 0.90\epsilon^{0.47} + 0.60$. ϵ is the circularity, which is defined as the ratio of the satellite orbital angular momentum at infall to the angular momentum of a circular orbit with the same energy as that of the satellite orbit. This circular orbit is derived in the same potential as for the satellite motion, and $r_{\rm circ}$ is its radius.

The satellite's orbital angular momentum and energy at infall can be calculated though its radial and tangential velocities ($v_{\rm r}$ and $v_{\rm t}$ respectively), assuming at infall the distance from satellite to the host halo centre is $r_{\rm vir}$. $v_{\rm r}$ and $v_{\rm t}$ are randomly picked for each satellite according to the probability distribution in Benson (2005), which is measured from N-body simulations.

The mass contained in the accreted galaxy is directly added to the central galaxy, while the contribution to angular momentum is more complex and entangled with the galaxy morphologies. This treatment will be described in §2.4.

2.4 Galaxy Morphology and Size

The cooling and galaxy merger calculations provide the mass accreted onto central galaxies. Further processes such as star formation, black hole growth and associated feedbacks happen inside galaxies. Modelling of these requires some information on galaxy internal structures, which are described in this section.

GALFORM assumes that a galaxy in general contains a bulge, which is random motion supported, containing stars and cold gas undergoing starbursts, and a disk, which is rotationally supported and generally also a mixture of stars and cold gas.

The cold gas received from cooling in this halo is assumed to be accreted onto the disk component. During a galaxy merger, the stars of the accreted galaxy are always assumed to be added to the bulge component, and the fate of the cold gas depends on the baryonic mass ratio of the merging galaxy pair. Letting $M_{\rm b,sat}$ and $M_{\rm b,cen}$ be the total baryon mass (stars plus cold gas) of the satellite and central galaxy respectively, then if $M_{\rm b,sat}/M_{\rm b,cen} < f_{\rm ellip}$, where $f_{\rm ellip}$ is a parameter, the merger is classified as minor merger. Further, if $M_{\rm b,sat}/M_{\rm b,cen} < f_{\rm burst}$, the cold gas is added to the disk component of the central galaxy. The contribution of this accreted gas to disk angular momentum is assumed to be such that after the merger, the central galaxy disk keeps its specific angular momentum unchanged. While if $M_{\rm b,sat}/M_{\rm b,cen} \geq f_{\rm burst}$, then the cold gas is added to the bulge of the central galaxy and triggers starbursts there. Otherwise, if $M_{\rm b,sat}/M_{\rm b,cen} \geq f_{\rm ellip}$, then the merger is a major merger, the central galaxy is assumed to be strongly disturbed and all

the mass in the accreted galaxy is added to the bulge of the central galaxy, and the mass in central galaxy's disk is also transfer to its bulge. Thus the major merger converts a disk galaxy to an elliptical galaxy. The Lacey16 model sets $f_{\rm ellip} = 0.3$ and $f_{\rm burst} = 0.1$.

Apart from mergers, disk instability can also affect morphology. A stable disk is required to satisfy

$$F_{\rm disk} \equiv \frac{V_{\rm c}(r_{\rm disk})}{(1.68GM_{\rm disk}/r_{\rm disk})^{1/2}} \geq F_{\rm stab} \tag{2.4.1}$$

where $V_{\rm c}(r_{\rm disk})$ is the circular velocity at the disk half-mass radius $r_{\rm disk}$, $M_{\rm disk}$ is the disk mass, and $F_{\rm stab}$ is a parameter. According to Efstathiou et al. (1982) and Christodoulou et al. (1995), $0.9 \lesssim F_{\rm stab} \lesssim 1.1$. The Lacey16 model adopts $F_{\rm stab} = 0.9$. Once a disk does not satisfy this condition, it is unstable. The unstable disk would form a bar, which subsequently evolves into a spheroid (Combes et al., 1990). GALFORM assumes that in this case, all disk materials are transfered to bulge component, and the relatively short time duration of this bar evolution is ignored.

The galaxy internal structure also includes the size of each component. The associated treatment was first introduced in Cole et al. (2000).

GALFORM assumes that the disk is thin and has an exponential 2D density distribution with half-mass radius $r_{\rm disk}$, while the bulge is spherical with an $r^{1/4}$ distribution as the projected density profile, and 3D half-mass radius $r_{\rm bulge}$.

The disk, bulge and dark matter halo interact through gravity. GALFORM calculates this using an adiabatic contraction model. For this calculation, the dark matter and baryons are assumed to be initially in the same profile, and then part of the baryons collapse to the halo centre to form the galaxy. This collapse also induces the contraction of halo matter (dark matter and baryons still in the halo), and it is assumed that this contraction is adiabatic, so that each halo shell conserves its specific pseudo angular momentum $\tilde{j}(r) = rV_{\rm c}(r)$, where r is the shell radius and $V_{\rm c}(r)$ is the gravitational circular velocity at r. For this calculation, the baryons left in the halo are assumed to keep the same density profile as the dark matter.

After the collapse, some halo shells move to the half-mass radii $r_{\rm disk}$ and $r_{\rm bulge}$ of disk and bulge respectively. Through $\tilde{j}(r)$, the radii of these shells after the collapse

can be related to their initial radii $r_{\rm disk,0}$ and $r_{\rm bulge,0}$, thus forming equations for $r_{\rm disk}$ and $r_{\rm bulge}$. Specifically, these equations are

$$\begin{aligned} r_{\rm disk,0}M_{\rm halo,0}(r_{\rm disk,0}) &= \frac{\tilde{j}^2(r_{\rm disk,0})}{G} \\ &= r_{\rm disk}\times[f_{\rm halo}M_{\rm halo,0}(r_{\rm disk,0})+ \\ &\quad \frac{1}{2}M_{\rm disk}+M_{\rm bulge}(r_{\rm disk})], \qquad (2.4.2) \\ r_{\rm bulge,0}M_{\rm halo,0}(r_{\rm bulge,0}) &= \frac{\tilde{j}^2(r_{\rm bulge,0})}{G} \\ &= r_{\rm bulge}\times[f_{\rm halo}M_{\rm halo,0}(r_{\rm bulge,0})+ \\ &\quad M_{\rm disk}(r_{\rm bulge})+\frac{1}{2}M_{\rm bulge}], \qquad (2.4.3) \end{aligned}$$

where $M_{\rm halo,0}(r)$ is the mass within r according to the initial halo density profile, which is assumed to be NFW, and $f_{\rm halo}$ is the mass fraction of materials that are still in the halo after collapse, while $M_{\rm disk}(r)$ and $M_{\rm bulge}(r)$ are derived through the assumed disk and bulge density distributions.

The above two equations are further supplemented with expressions for the disk specific angular momentum $j_{\rm disk}$ and the bulge specific pseudo angular momentum $\tilde{j}_{\rm bulge}$. For $j_{\rm disk}$, one has

$$\begin{aligned} j_{\rm disk} &= 1.19r_{\rm disk}V_{\rm c}(r_{\rm disk}) \\ &= 1.19r_{\rm disk}\sqrt{\frac{G[f_{\rm halo}M_{\rm halo,0}(r_{\rm disk,0})+k_{\rm h}M_{\rm disk}/2+M_{\rm bulge}(r_{\rm disk})]}{r_{\rm disk}}}, \qquad (2.4.4) \end{aligned}$$

where the factor 1.19 is derived for an exponential disk and a flat rotation curve, while $k_{\rm h}=1.25$ accounting for the disk geometry.

For $\tilde{j}_{\rm bulge}$, one has

$$\begin{aligned} \tilde{j}_{\rm bulge} &= r_{\rm bulge}V_{\rm c}(r_{\rm bulge}) \\ &= r_{\rm bulge}\sqrt{\frac{G[f_{\rm halo}M_{\rm halo,0}(r_{\rm bulge,0})+M_{\rm disk}(r_{\rm bulge})+M_{\rm bulge}/2]}{r_{\rm bulge}}} \\ &= \tilde{j}(r_{\rm bulge,0}). \qquad (2.4.5) \end{aligned}$$

Given $j_{\rm disk}$ and $\tilde{j}_{\rm bulge}$, Eq(2.4.2-2.4.5) can determine the disk half-mass radii $r_{\rm disk}$ and $r_{\rm bulge}$.

The angular momentum and mass of disk are straightforwardly derived through the gas cooling and galaxy merger recipes previously introduced. From them, it is

easy to derive $j_{\rm disk}$. For $\tilde{j}_{\rm bulge}$, GALFORM assumes that it is kept constant during the disk growth, but is reset when the bulge itself grows. So a new $\tilde{j}_{\rm bulge}$ should be assigned for each galaxy merger and disk instability event.

GALFORM estimates $\tilde{j}_{\rm bulge}$ by

$$\tilde{j}_{\rm bulge} = r'_{\rm bulge} V'_{\rm c} = \sqrt{\frac{G r'_{\rm bulge}[M_1 + M_2]}{2}}, \tag{2.4.6}$$

where M_1 and M_2 correspond to the masses of two merging galaxies or the disk and bulge of a galaxy, for galaxy merger and disk instability respectively, and they also include the associated dark matter mass in the case of galaxy merger, while $r'_{\rm bulge}$ is the estimated bulge half-mass radius, and is derived by assuming the internal energy (kinetic plus gravitational binding energy) conservation.

For a galaxy merger, the internal energy conservation has the form $E_{\rm int,remnant} = E_{\rm int,1} + E_{\rm int,2} + E_{\rm orbit}$, where $E_{\rm int,remnant}$ is the internal energy of the merger remnant, and the right hand side is the internal energy of the galaxy system just before merger, with $E_{\rm int,1}$ and $E_{\rm int,2}$ the internal energies of the two merging galaxies, and $E_{\rm orbit}$ the orbital energy. Here, for a galaxy, through the virial theorem one has

$$E_{\rm int} = -\frac{1}{2} E_{\rm bind} = -\frac{c_{\rm gal}}{2} \frac{G M_{\rm gal}^2}{r_{\rm gal}}, \tag{2.4.7}$$

where $M_{\rm gal}$ is the mass of a galaxy or a galaxy component, $r_{\rm gal}$ is the corresponding half-mass radius, and $c_{\rm gal}$ is a parameter that weakly depends on the galaxy geometry. For a purely exponential disk, $c_{\rm gal} = 0.49$, while for a $r^{1/4}$ law spherical bulge, $c_{\rm gal} = 0.45$, and Lacey16 model adopts $c_{\rm gal} = 0.5$ for simplicity. Also, here the galaxy mass $M_{\rm gal}$ does not only include the baryonic mass $M_{\rm gal,b}$ from cold gas and stars, but the dark matter mass $M_{\rm gal,dark}$ as well, and $M_{\rm gal,dark} = f_{\rm DM} M_{\rm halo}(r_{\rm gal})$. $f_{\rm DM}$ is a parameter. The Lacey16 model adopts $f_{\rm DM} = 2$. $M_{\rm gal,b}$ should include the masses of all baryons appearing in the final bulge.

The two galaxies just before merger are assumed to be separated by the distance equaling the sum of their half-mass radii. The orbital energy of this system is estimated as

$$E_{\rm orbit} = -\frac{f_{\rm orbit}}{2} \frac{G M_{\rm gal,1} M_{\rm gal,2}}{r_{\rm gal,1} + r_{\rm gal,2}}, \tag{2.4.8}$$

where $M_{\rm gal,1}$ and $M_{\rm gal,2}$ are the masses of the two galaxies, and $r_{\rm gal,1}$ and $r_{\rm gal,2}$ are their half-mass radii respectively, while $f_{\rm orbit}$ is a parameter. $0 \leq f_{\rm orbit} \lesssim 1$, with

$f_{\rm orbit} = 1$ corresponding to the energy of two point masses on a circular orbit with separation $r_{\rm gal,1} + r_{\rm gal,2}$, while $f_{\rm orbit} = 0$ means orbital energy is negligible. The Lacey16 model sets $f_{\rm orbit} = 0$.

Together these gives an equation for $r'_{\rm bulge}$ after the merger,

$$\frac{(M_{\rm gal,1} + M_{\rm gal,2})^2}{r'_{\rm bulge}} = \frac{M_{\rm gal,1}^2}{r_{\rm gal,1}} + \frac{M_{\rm gal,2}^2}{r_{\rm gal,2}} + \frac{f_{\rm orbit}}{c_{\rm gal}} \frac{M_{\rm gal,1} M_{\rm gal,2}}{r_{\rm gal,1} + r_{\rm gal,2}}. \tag{2.4.9}$$

For a disk instability, an equation similar to Eq(2.4.9) is adopted for $r'_{\rm bulge}$ estimation, but here the masses of two merging galaxies are replaced by the masses of the disk and bulge of one galaxy. Unlike for a galaxy merger, a disk instability induces mass transfer within one galaxy embedded in a roughly spherical dark matter component. The dark matter component is largely unaffected by this transfer, and only affects the potential energy zero point, so here the effect of dark matter can be neglected. Also, instead of the orbital energy, here the interaction between disk and bulge should be considered. Thus, finally one has

$$c_{\rm bulge} \frac{(M_{\rm disk} + M_{\rm bulge})^2}{r'_{\rm bulge}} = c_{\rm bulge} \frac{M_{\rm bulge}^2}{r_{\rm bulge}} + c_{\rm disk} \frac{M_{\rm disk}^2}{r_{\rm disk}} + f_{\rm int} \frac{M_{\rm disk} M_{\rm bulge}}{r_{\rm disk} + r_{\rm bulge}}, \tag{2.4.10}$$

where the last term on the right hand side is for the interaction between disk and bulge, with $f_{\rm int} = 2$ giving good approximations for a range of $r_{\rm bulge}/r_{\rm disk}$, and $c_{\rm bulge}$ and $c_{\rm disk}$ have the same meaning as $c_{\rm gal}$ in Eq(2.4.7), but this time more exact values are used, namely $c_{\rm bulge} = 0.45$ and $c_{\rm disk} = 0.49$.

2.5 Star Formation and Black Hole growth

Once the internal structure of a galaxy is known, the processes happening inside the galaxy can be calculated.

2.5.1 Quiescent Star Formation

This mode of star formation happens in disks. The treatment here was first introduced in Lagos et al. (2011). The cold gas in the disk is further divided into atomic and molecular. The star formation is assumed to be directly related to the molecular gas, namely

$$\psi_{\rm disk} = \nu_{\rm SF,disk} M_{\rm mol,disk} = \nu_{\rm SF,disk} f_{\rm mol} M_{\rm cold,disk}, \tag{2.5.1}$$

where $\psi_{\rm disk}$ is the quiescent star formation rate, $M_{\rm mol,disk}$ the molecular cold gas mass in the disk, $M_{\rm cold,disk}$ the total cold gas mass in the disk, $f_{\rm mol}$ the molecular fraction and $M_{\rm mol,disk} = f_{\rm mol} M_{\rm cold,disk}$, while $\nu_{\rm SF,disk}$ is a parameter. The Lacey16 model adopts $\nu_{\rm SF,disk} = 0.74\,{\rm Gyr}^{-1}$, which is consistent with the observations in Bigiel et al. (2011) within 1σ.

$M_{\rm cold,disk}$ is easy to derive through the cooling and galaxy merger calculation. $f_{\rm mol}$ is calculated based on the empirical scaling proposed in Blitz & Rosolowsky (2006). This law states that

$$f_{\rm mol} = \left(\frac{P}{P_0}\right)^{\alpha_{\rm P}}, \tag{2.5.2}$$

where P is the internal hydrostatic pressure of the disk, and P_0 and $\alpha_{\rm P}$ are two parameters. The Lacey16 model adopts $P_0 = 1700\,{\rm cm}^{-3}{\rm K}$ and $\alpha_{\rm P} = 0.8$ based on the observations of Leroy et al. (2008). P is calculated by the following equation (Elmegreen, 1993; Lagos et al., 2011)

$$P = \frac{\pi}{2} G \Sigma_{\rm gas} \left[\Sigma_{\rm gas} + \left(\frac{\sigma_{\rm gas}}{\sigma_{\rm star}}\right) \Sigma_{\rm star}\right], \tag{2.5.3}$$

where $\Sigma_{\rm gas}$ and $\Sigma_{\rm star}$ are the total surface densities of gas and stars, $\sigma_{\rm gas}$ is the gas velocity dispersion perpendicular to the disk, and is set to a constant $10\,{\rm kms}^{-1}$, and $\sigma_{\rm star}$ is the vertical velocity dispersion of stars, which is given by

$$\sigma_{\rm star} = \max[\sqrt{\pi G h_{\rm star} \Sigma_{\rm star}}, \sigma_{\rm gas}] \tag{2.5.4}$$

with $h_{\rm star}$ the scale height of stellar disk. It is assumed $h_{\rm star}$ is proportional to the disk scale radius based on observations of Kregel et al. (2002), specifically, $h_{\rm star} = 0.14 l_{\rm exp}$, where $l_{\rm exp}$ is the radial scale length of the exponential density distribution, and is related to the disk half-mass radius $r_{\rm disk}$ as $r_{\rm disk} = 1.68 l_{\rm exp}$. Note that the first term in the maximum function in Eq(2.5.4) is valid when star dominates the surface density, while the second term is for the case of gas dominant.

2.5.2 Starbursts

The Lacey16 model assumes that mergers with mass ratio $M_{\rm b,sat}/M_{\rm b,cen} > f_{\rm burst} = 0.1$ trigger starbursts, and that disk instabilities also induce starbursts. It is assumed that $f_{\rm mol} \sim 1$ for starbursts. Thus the star formation rate $\psi_{\rm burst}$ is directly

proportional to the cold gas mass available for bursts, $M_{\rm cold,burst}$, namely

$$\psi_{\rm burst} = \nu_{\rm SF,burst} M_{\rm cold,burst}, \tag{2.5.5}$$

where $\nu_{\rm SF,burst} = 1/\tau_{\rm burst}$ and

$$\tau_{\rm burst} = \max[f_{\rm dyn}\tau_{\rm dyn,bulge}, \tau_{\rm burst,min}], \tag{2.5.6}$$

with $f_{\rm dyn}$, $\tau_{\rm burst,min}$ two parameters, and $\tau_{\rm dyn,burst} = r_{\rm bulge}/V_{\rm c}(r_{\rm bulge})$ is the dynamical time scale of bulge. The Lacey16 model adopts $f_{\rm dyn} = 20$ and $\tau_{\rm burst,min} = 0.1\,{\rm Gyr}$.

2.5.3 Initial Mass Function

Knowing the star formation rate only tells us the total mass of a stellar population formed in a given duration, while further calculations about chemical enrichment and stellar luminosities depend on the abundance of each kind of stars, since different kinds of stars have different contributions to these process. These abundances are expressed by the initial mass function (IMF), which gives the number of stars of each mass in a stellar population of unit mass.

The Lacey16 model assumes different IMFs for quiescent star formation and starbursts. For the former one, the Lacey16 model assumes the IMF, $\Phi_{\rm disk}$, is a broken power-law, namely

$$\Phi_{\rm disk}(m) = \frac{dN}{d\ln m} \propto \begin{cases} m^{-0.4}, & 0.1\,{\rm M}_\odot \leq m < 1\,{\rm M}_\odot \\ m^{-1.5}, & 1\,{\rm M}_\odot \leq m \leq 100\,{\rm M}_\odot \end{cases}, \tag{2.5.7}$$

where m is the mass of individual stars, while dN is the number of stars in the mass interval $(m, m \times \exp[d\ln m])$. This IMF is based on the observations of nearby disk galaxies in Kennicutt (1983), and is also consistent with other observations of the solar neighborhood (e.g. Kroupa, 2002; Chabrier, 2003).

For starbursts, Lacey16 model assumes a single power-law IMF, $\Phi_{\rm burst}(m)$, where

$$\Phi_{\rm burst}(m) = \frac{dN}{d\ln m} \propto m^{-1}, \ 0.1\,{\rm M}_\odot \leq m \leq 100\,{\rm M}_\odot. \tag{2.5.8}$$

The massive end of $\Phi_{\rm burst}(m)$ is shallower than that of $\Phi_{\rm disk}(m)$, so the IMF for bursts is relatively top-heavy.

2.5.4 Black Hole Growth

GALFORM does not explicitly seed black holes in galaxies, this is because the seed mass is typically negligible compared to the final black hole mass (e.g. Malbon et al., 2007). GALFORM then assumes three growth channels for seed black holes.

Firstly, black holes can accrete through the massive cold gas flows towards the galaxy centre induced by a galaxy merger or disk instability. This black hole growth is accompanied by a starburst, thus GALFORM relates these two processes, and assumes that

$$\Delta M_{\rm BH} = f_{\rm BH} \Delta M_{\rm star,burst}, \tag{2.5.9}$$

where $\Delta M_{\rm BH}$ is the mass increase of the black hole, and $\Delta M_{\rm star,burst}$ is the mass of stars formed during the starburst, while $f_{\rm BH}$ is a parameter. The Lacey16 model adopts $f_{\rm BH} = 0.005$.

Secondly, black holes can also accrete diffuse matter from the hot gas halo. This typically leads to very low accretion rates, and is associated with the launch of a relativistic radio jet. The jet could suppress gas cooling, which is known as the radio mode AGN feedback (see §2.6.2). GALFORM assumes this mode of black hole growth is active only if the radio mode AGN feedback is effective. Under this assumption, the heating luminosity from the black hole $L_{\rm heat}$ must balance the gas cooling luminosity $L_{\rm cool}$, and the latter one can be derived based on the gas cooling calculation. Then GALFORM introduces a relation between the black hole mass accretion rate $\dot{M}_{\rm BH}$ and $L_{\rm heat}$, namely $L_{\rm heat} = \epsilon_{\rm heat} \dot{M}_{\rm BH} c^2$, where c is the speed of light and $\epsilon_{\rm heat}$ is a parameter. Thus, the black hole mass growth in a time interval Δt is

$$\Delta M_{\rm BH} = \dot{M}_{\rm BH} \Delta t = \frac{L_{\rm cool}}{\epsilon_{\rm heat} c^2} \Delta t. \tag{2.5.10}$$

The Lacey16 model adopts $\epsilon_{\rm heat} = 0.02$.

Thirdly, black holes increase mass during black hole mergers. GALFORM ignores the energy emitted in gravitational waves, and thus the merger conserves total black hole mass. GALFORM also ignores the time delay between the galaxy merger and the black hole merger.

2.6 Feedback Processes

2.6.1 Supernova Feedback

Massive stars evolve into supernovae (SN). The massive explosion of SN ejects huge amount of energy into the ISM, and may expel cold gas out of galaxies and thus regulate star formation. GALFORM typically ignores the relatively short time duration of massive star evolution, and assumes that the massive stars instantaneously become SNe. In this case, the mass outflow rate induced by SN feedback, $\dot{M}_{\rm eject}$, is proportional to the instantaneous star formation rate ψ, namely

$$\dot{M}_{\rm eject} = \beta\psi, \tag{2.6.1}$$

where ψ could be either $\psi_{\rm disk}$ or $\psi_{\rm burst}$. The Lacey16 model adopts a single power-law for the proportionality factor β, i.e.

$$\beta = \left(\frac{V_{\rm c}}{V_{\rm SN}}\right)^{-\gamma_{\rm SN}}, \tag{2.6.2}$$

where $\gamma_{\rm SN}$ and $V_{\rm SN}$ are two parameters, and $V_{\rm c} = V_{\rm c}(r_{\rm disk})$ for quiescent star formation, while $V_{\rm c} = V_{\rm c}(r_{\rm bulge})$ for starbursts. The Lacey16 model adopts the same parameter values for these two modes of star formation, that is, $\gamma_{\rm SN} = 3.2$ and $V_{\rm SN} = 320\,{\rm km\,s^{-1}}$.

The gas is assumed to be ejected out of the dark matter halo, and joins a gas reservoir called the reheated gas, which has mass $M_{\rm eject}$. This gas can later be reaccreted onto the hot gas halo, on a timescale comparable to the halo dynamical timescale, namely

$$\dot{M}_{\rm return} = \alpha_{\rm return}\frac{M_{\rm eject}}{\tau_{\rm dyn,halo}}, \tag{2.6.3}$$

where $\dot{M}_{\rm return}$ is the rate of reaccretion onto the hot gas halo, $\tau_{\rm dyn,halo} = r_{\rm vir}/v_{\rm vir}$ is the halo dynamical time scale and $\alpha_{\rm return} \sim 1$ is a parameter. The Lacey16 model adopts $\alpha_{\rm return} = 0.64$.

2.6.2 AGN Feedback

There are thought to be two modes of AGN feedback. One is called the quasar mode (Croton et al., 2006), in which the black hole accretion rate is high, and it is

thought that the energy from AGN may blow out the cold gas in the host galaxy, thus quenching star formation (e.g. Monaco et al., 2007). The high accretion rate is typically triggered by a disk instability or galaxy merger, which also lead to starbursts. Phenomenologically, this feedback mode is similar to the SN feedback in starbursts. In GALFORM, this feedback mode is not modelled explicitly.

The other mode is the so-called radio mode, in which the black hole accretion rate is low and it is believed that a relativistic radio jet is launched. The jet delivers energy to the hot gas halo and suppresses its cooling. There are naturally two conditions for this feedback to be effective. Firstly, the gas cooling is mainly through a hot, nearly hydrostatic gas halo, which is the slow cooling regime. Only in this case can the gas keep high enough pressure to effectively interact with the jet and gain energy from it. Secondly, the black hole accretion rate should be low, so that a jet can be efficiently produced (Fanidakis et al., 2011). And also for the feedback to be effective, its heating should balance the cooling luminosity, namely $L_{\rm heat} \geq L_{\rm cool}$.

Accordingly, GALFORM imposes two conditions for an effective radio mode AGN feedback. The first one is

$$\frac{t_{\rm cool}(r_{\rm cool})}{t_{\rm ff}(r_{\rm cool})} > \frac{1}{\alpha_{\rm cool}}, \tag{2.6.4}$$

where $t_{\rm cool}(r_{\rm cool})$ is the cooling timescale at the cooling radius $r_{\rm cool}$, while $t_{\rm ff}(r_{\rm cool})$ is the free-fall timescale at $r_{\rm cool}$, and $\alpha_{\rm cool} \sim 1$ is a parameter. When this condition is satisfied, the cooling gas can be treated as being quasi-hydrostatic. The Lacey16 model sets $\alpha_{\rm cool} = 0.8$.

The second condition is

$$L_{\rm cool} < f_{\rm Edd} L_{\rm Edd}(M_{\rm BH}), \tag{2.6.5}$$

where $L_{\rm Edd}(M_{\rm BH})$ is the Eddington luminosity of a black hole with mass $M_{\rm BH}$, and $f_{\rm Edd}$ is a parameter that should be small. The Lacey16 model adopts $f_{\rm Edd} = 0.01$. $f_{\rm Edd} L_{\rm Edd}(M_{\rm BH})$ gives the upper limit of the heating luminosity through accretion with a low enough rate to allow jet launch, while if this limit is higher than the cooling luminosity, then the radio mode feedback is effective, and the cooling calculation is stopped to model the suppression of cooling.

2.6.3 Photonionization Feedback

This feedback is modelled in GALFORM through the so-called $z_{\rm crit}$-$V_{\rm crit}$ approximation first introduced in Benson et al. (2003). This approximation treats reionization as an instantaneous phase transition of the IGM. Reionization happened at redshift $z_{\rm crit}$, and after that, any dark matter halo with $v_{\rm vir} < V_{\rm crit}$ would have neither gas accretion nor gas cooling, due to photonionization feedback. $z_{\rm crit}$ and $V_{\rm crit}$ are two parameters, which have values 10 and 30 $\rm km\,s^{-1}$ respectively in Lacey16 model. The value of $V_{\rm crit}$ is consistent with the hydrodynamical simulation results of Okamoto et al. (2008).

Benson et al. (2002) and Font et al. (2011) show that this method gives a good approximation to a more complex and self-consistent photonionization feedback calculation.

2.7 Metal Enrichment

The metals are produced in stars. Then through SN explosions and stellar winds, they are firstly ejected into the galaxy cold gas, and then some are blown out from the galaxy along with the gas outflows induced by SN feedback. This process brings metals into the reheated gas. Then, when the reheated gas is reaccreted onto the hot gas halo, metals are transported to the hot halo gas, and finally, through cooling, metals go back into galaxies' cold gas, and then go into the newly formed stars through star formation.

GALFORM assumes that the amount of metals in each flow mentioned above is the product of the mass in that flow and the metallicity of the corresponding gas component.

Among all these flows, those involving stars deserve further attention, for stars are the source of new metals. For a time interval dt, star formation converts ψdt mass of cold gas into stars, where ψ is the star formation rate. According to the above assumption, in this process, $Z_{\rm cold}\psi dt$ mass of metals also go into stars, where $Z_{\rm cold}$ is the cold gas metallicity and $Z_{\rm cold} = M_{\rm z,cold}/M_{\rm cold}$, with $M_{\rm z,cold}$ and $M_{\rm cold}$ respectively the total metal mass and total mass of the cold gas.

For the subsequent stellar evolution, GALFORM adopts the instantaneous recycling approximation, in which the stars (mainly massive stars) instantaneously return part of their mass and metals back to the cold gas, together with some newly formed metals, and lock the rest of the mass and metals in the stellar remnants forever. According to this, the returned mass in dt is $R\psi dt$, where R is the return mass fraction, and should be calculated according to the assumed IMF (§2.5.3). The mass of returned metals is $RZ_{\rm cold}\psi dt + p\psi dt$. Here the first term represents the metals taken from the cold gas during star formation, while the second term is the metals newly generated by stars, and p, the yield, is the ratio of the mass of new metals to the total mass of stars formed, and is also calculated based on the IMF.

Knowing the metal flows through stars, the other metal flows are relatively straightforward. The whole metal transport can be summarized as a set of equations below.

$$\Delta M_{\rm z,hot} = -Z_{\rm hot} M_{\rm acc,gal} + \alpha_{\rm return} \frac{M_{\rm z,eject}}{\tau_{\rm dyn,halo}} \Delta t \quad (2.7.1)$$

$$\Delta M_{\rm z,cold} = Z_{\rm hot} M_{\rm acc,gal} + [p - (1 - R + \beta) Z_{\rm cold}] \psi \Delta t \quad (2.7.2)$$

$$\Delta M_{\rm z,star} = (1 - R) Z_{\rm cold} \psi \Delta t \quad (2.7.3)$$

$$\Delta M_{\rm z,eject} = \beta Z_{\rm cold} \psi \Delta t - \alpha_{\rm return} \frac{M_{\rm z,eject}}{\tau_{\rm dyn,halo}} \Delta t \quad (2.7.4)$$

In the above Δt is the length of a time step, $M_{\rm hot}$, $M_{\rm cold}$, $M_{\rm eject}$ and $M_{\rm star}$ are the masses of halo hot gas, cold gas in galaxies, SN feedback ejected gas and stars respectively, while $\Delta M_{\rm z,hot}$, $\Delta M_{\rm z,cold}$, $\Delta M_{\rm z,eject}$ and $\Delta M_{\rm z,star}$ are the corresponding metal mass changes. $Z_{\rm cold}$ is the cold gas metallicity defined above, $Z_{\rm hot}$ is the hot gas metallicity defined similarly through Eq(2.2.2), and $M_{\rm z,eject}$ is the total metal mass of the ejected gas. $M_{\rm acc,gal}$ is provided by the gas cooling calculation, while ψ is given by the star formation calculation.

2.8 Calculating Galaxy Luminosities

Combining all the previous calculations gives the stellar mass and metallicity as a function of time for each galaxy. This information, combined with the luminosity of a single stellar population (SSP), can be converted to the spectral energy distribution

(SED) of each galaxy. Specifically, the SED of a galaxy at time t, $L_\lambda(t)$, is given by

$$L_\lambda(t) = \int_0^t dt' \int_0^\infty dZ' \Psi(t', Z') L_\lambda^{\rm SSP}(t - t', Z'; \Phi), \tag{2.8.1}$$

where $\Psi(t', Z')dt'dZ'$ is the mass of stars formed between $[t', t' + dt']$ and with metallicity within the range $[Z', Z' + dZ']$, while $L_\lambda^{\rm SSP}(t - t', Z'; \Phi)$ is the SED of a SSP with unit mass, age $t - t'$, metallicity Z' and IMF Φ. $L_\lambda^{\rm SSP}(t, Z; \Phi)$ can be calculated through the SED of a single star (with age t, metallicity Z and mass m), $L_\lambda^{\rm star}(t, Z, m)$, and a given IMF $\Phi(m)$ as

$$L_\lambda^{\rm SSP}(t, Z; \Phi) = \int_{m_{\rm L}}^{m_{\rm U}} L_\lambda^{\rm star}(t, Z, m)\Phi(m) d\ln m, \tag{2.8.2}$$

with $m_{\rm L}$ and $m_{\rm U}$ respectively the lower and upper mass limit of the IMF. The Lacey16 model uses the stellar population library provided in Maraston (2005) to build SEDs for galaxies.

To further derive the observable luminosities, dust extinction needs to be modelled. The Lacey16 model assumes the following dust extinction picture.

The dust is distributed in the cold gas, and its total amount is assumed to be linearly proportional to the metallicity of cold gas $Z_{\rm cold}$. There are two phases of the cold gas, i.e. diffuse cold gas and molecular clouds. These clouds have the typical mass and radius $m_{\rm cloud}$ and $r_{\rm cloud}$ respectively. It is then further assumed that there is a fraction $f_{\rm cloud}$ of dust in the molecular clouds, and the rest is in the diffuse cold gas.

Stars are formed in the centres of molecular clouds, and then gradually leave their birth clouds on the timescale $t_{\rm esc}$. For the stars inside the clouds, these clouds exert extinction according to the optical depth $\tau_{\rm cloud} \propto Z_{\rm cold} m_{\rm cloud}/r_{\rm cloud}^2$. Here $\tau_{\rm cloud}$ is assumed to be proportional to the amount of dust (thus proportional to $Z_{\rm cloud}$) and the projected density, which is estimated by $m_{\rm cloud}/r_{\rm cloud}^2$.

For the stars outside the molecular clouds as well as the light emitted from these clouds, the dust in the diffuse cold gas causes further extinction. This is calculated using the tabulated radiative transfer models of Ferrara et al. (1999). These tables provide the dust attenuations of the disk and bulge luminosities as functions of wavelength, ratio of disk to bulge half-light radii, disk inclination and central dust

optical depth. The ratio of half-light radii is directly estimated from the galaxy sizes provided by GALFORM, while the disk inclination is randomly picked for each galaxy. The central dust optical depth is estimated as $\tau_{\rm cen} \propto (1-f_{\rm cloud}) Z_{\rm cold} M_{\rm cold}/r_{\rm eff}^2$, where $r_{\rm eff} = r_{\rm disk}$ or $r_{\rm bulge}$ for quiescent star formation and starbursts respectively.

The Lacey16 model adopts $f_{\rm cloud} = 0.5$ and $t_{\rm esc} = 1\,{\rm Myr}$, while $M_{\rm cloud}$ and $r_{\rm cloud}$ are fixed based on observations of nearby galaxies following Granato et al. (2000).

Chapter 3

Constraining SN feedback: a tug of war between reionization and the Milky Way satellites

3.1 Introduction

Supernova feedback (SN feedback hereafter) is a very important physical process for regulating the star formation in galaxies (Larson, 1974; Dekel & Silk, 1986; White & Frenk, 1991). Despite its importance, SN feedback is not well understood. Perhaps the best way to improve our understanding of this process is by investigating its physical properties using hydrodynamical simulations. This, however, is very difficult to achieve with current computational power: cosmological hydrodynamical simulations (e.g. Davé et al., 2013; Vogelsberger et al., 2014; Schaye et al., 2015) can provide large galaxy samples and can follow galaxy evolution spanning the history of the Universe, but do not have high enough resolution to follow individual star forming regions, which is needed to understand the details of SN feedback; conversely high resolution hydrodynamical simulations (e.g. Bate, 2012; Hopkins et al., 2012) can resolve many more details of individual star forming regions, but do not provide a large sample and cannot follow a long period of evolution. Because of these limitations, it is worth trying to improve our understanding of SN feedback in alternative ways. One promising approach is to extract constraints on SN feedback from

theoretical models of galaxy formation combined with observational constraints.

Among all relevant observations, a combination of four observables may be particularly effective because they constrain the strength of feedback in opposite directions. These are the abundance of faint galaxies, including both the faint ends of the $z = 0$ field galaxy luminosity function (hereafter field LF) and the Milky Way satellite luminosity function (hereafter MW sat LF), the Milky Way satellite stellar metallicity vs. stellar mass correlation (hereafter MW sat $Z_* - M_*$ correlation) and the redshift, $z_{\rm re,half}$, at which the Universe was 50% reionized. The observed abundance of faint galaxies is very low compared to the abundance of low mass dark matter halos in the standard cold dark matter (CDM) model of cosmogony (e.g. Benson et al., 2003; Moore et al., 1999; Klypin et al., 1999), which cannot be reproduced by very weak SN feedback, and this puts a lower limit on the SN feedback strength. On the other hand, $z_{\rm re,half}$ and the MW sat $Z_* - M_*$ correlation put upper limits on the SN feedback strength, because too strong a SN feedback would cause too strong a metal loss and a suppression of star formation in galaxies, thus leading to too low Z_* at a given M_*, and too low $z_{\rm re,half}$. Also note that this combination of observations constrains SN feedback over a wide range of galaxy types and redshifts: the field LF mainly provides constraints on SN feedback in larger galaxies, with circular velocity $V_{\rm c} \gtrsim 80\,{\rm km\,s^{-1}}$, while $z_{\rm re,half}$ mainly constrains SN feedback at $z \gtrsim 8$, and the Milky Way satellite observations (MW sat LF and MW sat $Z_* - M_*$ correlation) provide constraints on the SN feedback in very small galaxies, i.e. $V_{\rm c} \lesssim 40\,{\rm km\,s^{-1}}$, and probably over a wide redshift range, from very high redshift to $z \sim 1$. (This is because recent observations (e.g. de Boer et al., 2012; Vargas et al., 2013) indicate that the Milky Way satellites have diverse star formation histories, with some of them forming all of their stars very early, and others having very extended star formation histories.)

In this chapter, we investigate the constraints placed by this combination of observations on SN feedback using the semi-analytical galaxy formation model GALFORM (Cole et al., 2000; Baugh et al., 2005; Bower et al., 2006; Lacey et al., 2016). A semi-analytical galaxy formation model is ideal for this aim, because with it one can generate large samples of galaxies with high mass resolution, which is important

for simulating both Milky Way satellites and star formation at high redshift, and it is also computationally feasible to explore various physical models and parameterizations.

This chapter is organized as follows. Section 3.2 describes the starting point of this work, the Lacey et al. (2016) (hereafter Lacey16) GALFORM model, as well as extensions of this model and details of the simulation runs. Section 3.3 presents the results from the Lacey16 and modified models. Section 3.4 discusses the physical motivation for some of the modifications, and also which galaxies drive cosmic reionization and what their $z = 0$ descendants are. Finally a summary and conclusions are given in Section 3.5.

3.2 Methods

3.2.1 Starting point: Lacey16 model

The basic model used in this work is the Lacey16 (Lacey et al., 2016) model, a recent version of GALFORM. This model, and the variants of it that we consider in this chapter, all assume a flat ΛCDM universe with cosmological parameters based on the WMAP-7 data (Komatsu et al., 2011): $\Omega_{\rm m0} = 0.272$, $\Omega_{\rm v0} = 0.728$, $\Omega_{\rm b0} = 0.0455$ and $H_0 = 70.4\ \mathrm{km\,s^{-1}Mpc^{-1}}$, and an initial power spectrum with slope $n_{\rm s} = 0.967$ and normalization $\sigma_8 = 0.810$. The Lacey16 model implements sophisticated modeling of disk star formation, improved treatments of dynamical friction on satellite galaxies and of starbursts triggered by disk instabilities and an improved stellar population synthesis model; it reproduces a wide range of observations, including field galaxy luminosity functions from $z = 0$ to $z = 3$, galaxy morphological types at $z = 0$, and the number counts and redshift distribution of submillimetre galaxies. An important feature of this model is that it assumes a top-heavy IMF for stars formed in starbursts, which is required to fit the submillimeter data, while stars formed by quiescent star formation in disks have a Solar neighbourhood IMF. Stellar luminosities of galaxies at different wavelengths, and the production of heavy elements by supernovae, are predicted self-consistently, allowing for the varying IMF.

SN feedback is modeled in this and earlier versions of GALFORM as follows. SN

feedback ejects gas out of galaxies, and thus reduces the amount of cold gas in galaxies, regulating the star formation. The gas ejection rate is formulated as:

$$\dot{M}_{\rm eject} = \beta\psi, \tag{3.2.1}$$

where $\dot{M}_{\rm eject}$ is the mass ejection rate, ψ is the star formation rate and the mass-loading factor, β, encodes the details of SN feedback models. In the approximation of instantaneous recycling that we use here, in which we neglect the time delay between the birth and death of a star, the supernova rate, and hence also the supernova energy injection rate, are proportional to the instantaneous star formation rate ψ.

In the Lacey16 model, β is set to be a single power law in galaxy circular velocity, $V_{\rm c}$, specifically,

$$\beta = \left(\frac{V_{\rm c}}{V_{\rm SN}}\right)^{-\gamma_{\rm SN}}, \tag{3.2.2}$$

where $V_{\rm SN}$ and $\gamma_{\rm SN}$ are two free parameters. In the Lacey16 model, $V_{\rm SN} = 320\,{\rm km\,s^{-1}}$ and $\gamma_{\rm SN} = 3.2$. β as a function of $V_{\rm c}$ for the Lacey16 model is illustrated in the left panel of Fig. 3.1.

As shown in Figs. 3.3 and 3.5, the above single power-law SN feedback model is disfavored by the combination of the four observational constraints mentioned in §3.1. We therefore investigate some modified SN feedback models and test them against the same set of observations. These modified models are described next.

3.2.2 Modified SN feedback models

In the modified SN feedback models we assume a broken power law for β, with a change in slope below a circular velocity, $V_{\rm thresh}$:

$$\beta = \begin{cases} (V_{\rm c}/V_{\rm SN})^{-\gamma_{\rm SN}} & V_{\rm c} \geq V_{\rm thresh} \\ (V_{\rm c}/V'_{\rm SN})^{-\gamma'_{\rm SN}} & V_{\rm c} < V_{\rm thresh} \end{cases}. \tag{3.2.3}$$

Here $V_{\rm SN}$, $\gamma_{\rm SN}$, $V_{\rm thresh}$ and $\gamma'_{\rm SN}$ are free parameters, while $V'_{\rm SN}$ is fixed by the condition that the two power laws should join at $V_{\rm c} = V_{\rm thresh}$.

3.2.2.1 Saturated feedback model

In this class of models we set $\gamma'_{\rm SN} < \gamma_{\rm SN}$, so that the mass-loading factor, β, for $V_{\rm c} < V_{\rm thresh}$ is lower than in the single power-law model. Note that we require

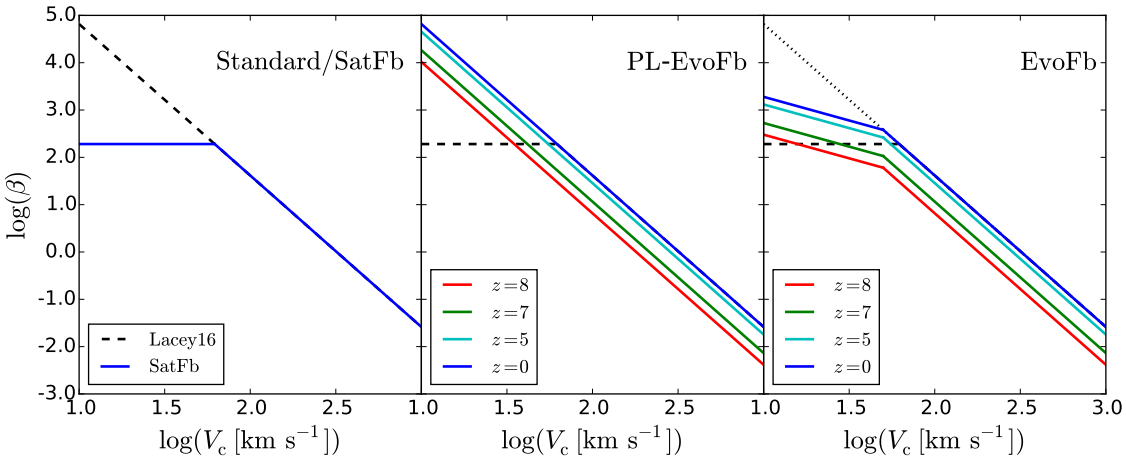

Figure 3.1: Mass-loading factor, β, as a function of circular velocity, V_c, and redshift, z, for the different supernova feedback models used in this work. *Left panel:* The dashed black line shows β in the Lacey16 model, while the solid blue line shows β for the SatFb model. *Middle panel:* β for the PL-EvoFb model. Different colours indicate different redshifts (from top to bottom, redshift increases from 0 to 8). This model is identical to the Lacey16 model for $z \leq 4$ (solid blue line). The SatFb model is also plotted as a dashed line for reference. *Right panel:* β for the EvoFb model. Different colours indicate different redshifts (from top to bottom, redshift increases from 0 to 8). β for the Lacey16 and SatFb models are also plotted for reference, and are shown by the black dotted and dashed lines respectively.

$\gamma'_{\rm SN} \geq 0$, because a negative $\gamma'_{\rm SN}$ would predict an anti-correlation between galaxy stellar metallicity and stellar mass, in contradiction with observations of the MW satellites.

A similar feedback model, with $\gamma'_{\rm SN} = 0$, was previously used by Font et al. (2011), which showed that it improved the agreement of GALFORM model predictions with Milky Way observations. However, in the present work, the observational constraints are more stringent than in Font et al. (2011), because here not only are Milky Way observations considered, but also the field LF and the reionization redshift.

In this work, we investigate a specific saturated feedback model, with $V_{\rm thresh} = 62\,{\rm km\,s^{-1}}$ and $\gamma'_{\rm SN} = 0$, which implies that β is a constant for galaxies with $V_{\rm c} < 62\,{\rm km\,s^{-1}}$ but reduces to the standard Lacey16 form for $V_{\rm c} > V_{\rm thresh}$. We call this specific saturated feedback model SatFb. The mass-loading factor for this model is also illustrated in the left panel of Fig. 3.1.

3.2.2.2 Evolving feedback model

This class of model has weaker SN feedback strength at high redshift. Here we investigate two specific models. For the first one, called PL-EvoFb, the feedback strength is a single power law in $V_{\rm c}$ at any redshift, but the normalization changes with redshift, being identical to the Lacey16 model for $z \leq 4$, but lower at high redshifts, Specifically, this model has $\gamma'_{\rm SN} = \gamma_{\rm SN} = 3.2$ (as in Lacey16) and

$$V_{\rm SN}\ [{\rm km\,s^{-1}}] = \begin{cases} 180 & z > 8 \\ -35z + 460 & 4 \leq z \leq 8 \\ 320 & z < 4 \end{cases} . \tag{3.2.4}$$

The general behaviour of this model is motivated by the results of Lagos et al. (2013), who predicted mass-loading factors from a detailed model of SN-driven superbubbles expanding in the ISM. (The Lagos et al. model was however incomplete, in that it considered only gas ejection out of the galaxy disk, but not out of the halo.) The mass loading factor, β, for this model is illustrated in the middle panel of Fig. 3.1.

The second model that we try, called EvoFb, has a normalization that evolves with redshift as in the PL-EvoFb model, but also has a shallower $V_{\rm c}$-dependence

at low V_c. Specifically, this model has $\gamma_{SN} = 3.2$ (as in Lacey16), $\gamma'_{SN} = 1.0$, $V_{thresh} = 50\,km\,s^{-1}$ and $V_{SN}(z)$ as given in Eq (3.2.4). For $V_c > 50\,km\,s^{-1}$, this model is identical to the PL-EvoFb model, but it has weaker feedback for $V_c \leq 50\,km\,s^{-1}$. The saturation in β at low V_c is therefore weaker than in the SatFb model. The mass loading factor for this model is illustrated in the right panel of Fig. 3.1.

The physical motivation for introducing the redshift evolution in the SN feedback will be discussed further in §3.4.

3.2.3 The redshift of reionization and photoionization feedback

We estimate the redshift of reionization predicted by a GALFORM model by calculating the ratio, $\mathcal{R}(z)$, of the number density of ionizing photons produced up to that redshift to the number density of hydrogen nuclei:

$$\mathcal{R}(z) = \frac{\int_z^\infty \epsilon(z')dz'}{n_H}, \tag{3.2.5}$$

where $\epsilon(z')$ is the number of hydrogen-ionizing photons produced per unit comoving volume per unit redshift at redshift z', and n_H is the comoving number density of hydrogen nuclei.

The Universe is assumed to be fully ionized at a redshift, $z_{re,full}$, for which,

$$\mathcal{R}(z_{re,full}) = \frac{1 + N_{rec}}{f_{esc}}, \tag{3.2.6}$$

where N_{rec} is the mean number of recombinations per hydrogen atom up to reionization, and f_{esc} is the fraction of ionizing photons that can escape from the galaxies producing them into the IGM. In this work we adopt $N_{rec} = 0.25$ and $f_{esc} = 0.2$, and thus the threshold for reionization is $\mathcal{R}(z_{re,full}) = 6.25$. Below we justify these choices.

Our estimation of the reionization redshift using $\mathcal{R}(z)$ (Eqs (3.2.5) and (3.2.6)) appears to be different from another commonly used estimator based on Q_{HII}, defined as the volume fraction of ionized hydrogen, with reionization being complete when $Q_{HII} = 1$, but in fact they are essentially equivalent. The evolution equation for Q_{HII} is given in Madau et al. (1999) as $\dot{Q}_{HII} = \dot{n}_{ion}/n_H - Q_{HII}/\bar{t}_{rec}$, where n_{ion} is

the comoving number density of ionizing photons escaping into IGM and $\bar{t}_{\rm rec}$ is the mean recombination time scale. Integrating both sides of this equation from $t = 0$ to the time $t = t_{\rm re,full}$ when reionization completes, one obtains

$$\begin{aligned} & Q_{\rm HII}(t_{\rm re,full}) \\ &= \frac{\int_0^{t_{\rm re,full}} \dot{n}_{\rm ion}\, dt}{n_{\rm H}} - \frac{\int_0^{t_{\rm re,full}} [n_{\rm H} Q_{\rm HII}/\bar{t}_{\rm rec}]\, dt}{n_{\rm H}} \\ &= \frac{f_{\rm esc} \int_{z_{\rm re,full}}^{\infty} \epsilon(z') dz'}{n_{\rm H}} - \frac{n_{\rm rec,tot}}{n_{\rm H}} \end{aligned} \tag{3.2.7}$$

where $n_{\rm rec,tot}$ is the mean number of recombinations per comoving volume up to $z_{\rm re,full}$. Setting $Q_{\rm HII}(t_{\rm re,full}) = 1$ and defining $N_{\rm rec} = n_{\rm rec,tot}/n_{\rm H}$, one then obtains Eq (3.2.6) for $z_{\rm re,full}$.

With the expression for $\bar{t}_{\rm rec}$ given by Madau & Haardt (2015) (their Eqn 4), $N_{\rm rec}$ can be expressed as

$$N_{\rm rec} = \int_0^{t_{\rm re,full}} [Q_{\rm HII}(1+\chi)\alpha_{\rm B}(1+z)^3 C_{\rm RR}] dt \tag{3.2.8}$$

where $\chi = 0.083$, $\alpha_{\rm B}$ is the case-B recombination rate coefficient and $C_{\rm RR}$ the clumping factor. Using the clumping factor in Shull et al. (2012) and solving the equation for $Q_{\rm HII}$, Eqn (3.2.8) gives $N_{\rm rec}$ in the range $0.13-0.34$ for our four different SN feedback models and an IGM temperature, $T = 1 - 2 \times 10^4\,{\rm K}$. Our choice of $N_{\rm rec} = 0.25$ lies within this range; note that Eqn (3.2.6) is not very sensitive to $N_{\rm rec}$ when its value is much smaller than 1. Our choice for $N_{\rm rec}$ is lower than the values assumed in some previous works (e.g. Raičević et al., 2011) because recent simulations give lower clumping factors (see Finlator et al. 2012 and references therein).

The calculation of $\epsilon(z')$ requires a knowledge of the ionizing sources. The traditional assumption has been that these sources are mainly star-forming galaxies, but recently there have been some works (e.g. Fontanot et al., 2012; Madau & Haardt, 2015; Giallongo et al., 2015) suggesting that AGN could be important contributors to reionization of hydrogen in the IGM. Although AGN might be important for reionization, these current works rely on extrapolatiing the AGN luminosity function faintwards of the observed luminosity limit, and also extrapolating the observations at $z \leq 6$ to $z \sim 10$, in order to obtain a significant contribution to reionization

from AGN. These extrapolations are uncertain, therefore in this work we ignore any AGN contribution and assume that the ionizing photon budget for reionization is dominated by galaxies. We discuss how the AGN contribution affects our conclusion in more detail in §3.4.4.

The value of the escape fraction, $f_{\rm esc}$, is also uncertain. Numerical simulations including gas dynamics and radiative transfer have given conflicting results: Kimm & Cen (2014) estimated $f_{\rm esc} \sim 0.1$, with $f_{\rm esc} \sim 0.2$ for starbursts, while Paardekooper et al. (2015) found much lower values. These differences between simulations may result from differences in the modelling of the ISM or in how well it is resolved, both of which are challenging problems. Observationally, it is impossible to measure $f_{\rm esc}$ directly for galaxies at the reionization epoch, because escaping ionizing photons would, in any case, be absorbed by the partially neutral IGM. Thus, one has to rely on observations of lower redshift galaxies for clues to its value.

Observations of Lyman-break galaxies at $z = 3 - 4$ suggest a relatively low value, $f_{\rm esc} \sim 0.05$ (Vanzella et al., 2010), while observations of local compact starburst galaxies show indirect evidence for higher $f_{\rm esc}$ (e.g. Alexandroff et al., 2015); Borthakur et al. (2014) estimated $f_{\rm esc} = 0.21$ for one local example. It is therefore important to determine what class of currently observed galaxies are the best analogues of galaxies at the reionization epoch. In our simulations, galaxies at high redshift tend to be compact and, in addition, the galaxies dominating the ionizing photon budget are starbursts (see Fig. 3.8), so, as argued by Sharma et al. (2016), they may well have similar escape fractions to local compact starburst galaxies. Sharma et al. (2016) provide further arguments that support our choice of $f_{\rm esc} = 0.2$. We discuss how the uncertainties in $f_{\rm esc}$ affect our conclusions in more detail in §3.4.4.

Note that, as advocated by Sharma et al. (2016) we only assume $f_{\rm esc} = 0.2$ for $z \geq 5$; for lower redshifts, $f_{\rm esc}$ may drop to low values, consistent with recent studies which argue that $f_{\rm esc}$ evolves with redshift and increases sharply for $z > 4$ (e.g. Haardt & Madau, 2012; Kuhlen & Faucher-Giguère, 2012).

Observations of the CMB directly constrain the electron scattering optical depth to recombination, which is then converted to a reionization redshift by assuming a simple model for the redshift dependence of the ionized fraction. Papers by the

WMAP and *Planck* collaborations (e.g. Planck Collaboration et al., 2014) typically express the reionization epoch in terms of the redshift, $z_{\rm re,half}$, at which the IGM is 50% ionized, by using the simple model for non-instantaneous reionization described in Appendix B of Lewis (2008). For comparing with such observational estimates, we therefore calculate $z_{\rm re,half}$ from GALFORM by assuming $\mathcal{R}(z_{\rm re,half}) = \frac{1}{2}\mathcal{R}(z_{\rm re,full})$. For the above mentioned choices of $N_{\rm rec}$ and $f_{\rm esc}$, this is equivalent to $\mathcal{R}(z_{\rm re,half}) = 3.125$.

Reionization may suppress galaxy formation in small halos, an effect called photoionization feedback (Couchman & Rees, 1986; Efstathiou, 1992; Thoul & Weinberg, 1996). In this work, the photoionization feedback is modeled using a simple approximation (Benson et al., 2003), in which dark matter halos with circular velocity at the virial radius $V_{\rm vir} < V_{\rm crit}$ have no gas accretion or gas cooling for $z < z_{\rm crit}$. As shown by Benson et al. (2002) and Font et al. (2011), this method provides a good approximation to a more complex, self-consistent photoionization feedback model. Here, $V_{\rm crit}$ and $z_{\rm crit}$ are two free parameters. In this chapter, unless otherwise specified, we adopt $z_{\rm crit} = z_{\rm re,full}$ and $V_{\rm crit} = 30\,{\rm km\,s^{-1}}$. This value of $V_{\rm crit}$ is consistent with the hydrodynamical simulation results of Okamoto et al. (2008). Note that this method does not necessarily imply that star formation in galaxies in halos with $V_{\rm vir} < V_{\rm crit}$ is turned off immediately after $z = z_{\rm re,full}$. The star formation in these galaxies can continue as long as the galaxy cold gas reservoir is not empty.

3.2.4 Simulation runs

Studying reionization requires resolving galaxy formation in low mass halos ($M_{\rm vir} \sim 10^8 - 10^{10}\,{\rm M_\odot}$) at high redshifts ($z \sim 7 - 15$), and thus very high mass resolution for the dark matter halo merger trees. The easiest way to achieve this high resolution is to use Monte Carlo (MC) merger trees.

Studying the properties of the Milky Way satellites also requires very high mass resolution because the host halos of these small satellites are small. This too is easily achieved using MC merger trees. Furthermore, because building MC merger trees is computationally inexpensive, it is possible to build a large statistical sample of Milky Way-like halos to study their satellites.

In this work we generate MC merger trees using the method of Parkinson et al.

(2008). To study reionization, we ran simulations starting at $z_{\rm start} = 20$ down to different final redshifts, $z_{\rm end}$, to derive $\epsilon(z)$ defined in Eq (3.2.5) at $z = 5 - 15$ and the $z = 0$ field LF. We scale the minimum progenitor mass in the merger trees as $(1 + z_{\rm end})^{-3}$, with a minimum resolved mass, $M_{\rm res} = 7 \times 10^9\,{\rm M}_\odot$ for $z_{\rm end} = 0$. We have tested that these choices are sufficient to derive converged results. For the Milky Way satellite study, the present-day host halo mass is chosen to be in the range $5 \times 10^{11} - 2 \times 10^{12}\,{\rm M}_\odot$, which represents the current observational constraints on the halo mass of the Milky Way, and we sample this range with five halo masses evenly spaced in log(mass). For each of these halo masses, GALFORM is run on 100 MC merger trees, with minimum progenitor mass $M_{\rm res} = 1.4 \times 10^6\,{\rm M}_\odot$, which is small enough for modeling the Milky Way satellites, and $z_{\rm start} = 20$ and $z_{\rm end} = 0$. We do not attempt to select Milky Way-like host galaxies, because we found that the satellite properties correlate better with the host halo mass than with the host galaxy properties.

3.3 Results

In this section, we show how the results from the different models compare with the key observational constraints that we have identified, namely: the field galaxy luminosity functions at $z = 0$; the redshift of reionization; the MW satellite galaxy luminosity function; and the stellar metallicity vs stellar mass relation for MW satellites.

3.3.1 Lacey16 model

We begin by showing the results for the default Lacey16 model, since this then motivates considering models with modified SN feedback. Fig. 3.2 shows the b_J- and K-band field LFs of different models at $z = 0$ (left and right panels respectively). The dotted blue lines show the LFs calculated using N-body merger trees, as used in the original Lacey et al. (2016) paper to calibrate the model parameters. The fit to the observed LFs is seen to be very good. The solid blue lines show the predictions with identical model parameters but instead using MC merger trees, as used in

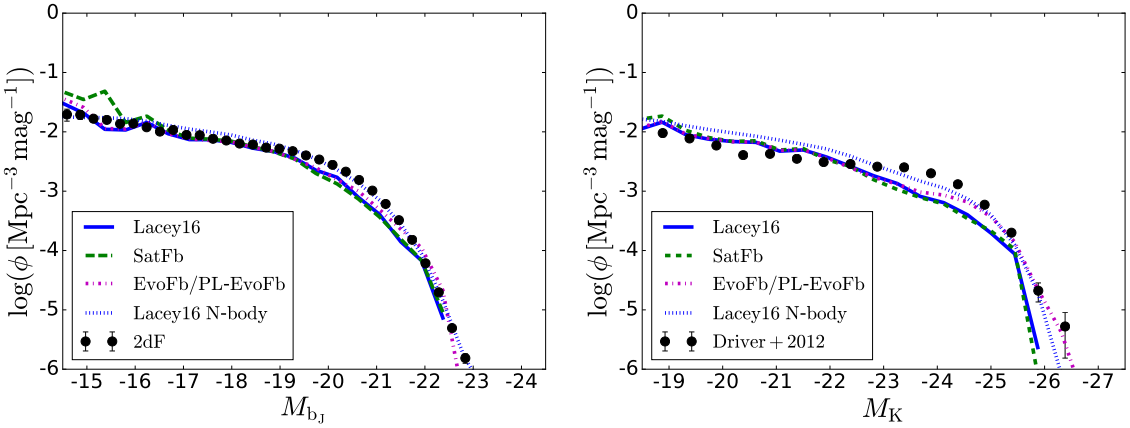

Figure 3.2: $z = 0$ field luminosity functions. The left panel shows the b_J-band luminosity function and the right panel the K-band luminosity function. In both panels the solid, dashed and dashed-dotted lines with different colours show the predictions using Monte Carlo merger trees for different SN feedback models, as indicated in the line key, while the blue dotted lines are for the Lacey16 model run with N-body merger trees. The magenta lines show the results of the EvoFb model, but the results for the PL-EvoFb model are almost identical. The black points with errorbars are observational data, from Norberg et al. (2002) for the b_J-band and from Driver et al. (2012) for the K-band.

the remainder of this chapter. The run with MC merger trees gives slightly lower LFs than the run with the N-body trees around the knee of the LF, but at lower luminosities, the results predicted using MC and N-body merger trees are in good agreement. We remind the reader that we use MC merger trees in the main part of this chapter in order to achieve the higher halo mass resolution that we need for the other observational comparisons. Since the differences in the LFs between the two types of merger tree are small, and barely affect the faint end of the field LFs which are the main focus of interest here, we do not consider them important for this chapter.

Fig. 3.3 shows the predicted $\mathcal{R}(z)$ (defined in Eq 3.2.5) for different SN feedback models. In each panel, the horizontal black dashed line indicates the criterion for 50% reionization, i.e. $\mathcal{R}(z_{\rm re,half}) = 3.125$, the vertical black dashed line indicates $z_{\rm re,half}$ of the corresponding model, and the corresponding value of $z_{\rm re,half}$ is given in the panel. The gray shaded area in each of these panels indicates the current observational constraint from *Planck*, namely $z_{\rm re} = 8.8^{+1.7}_{-1.4}$ (68% confidence region, Planck Collaboration et al., 2015). The redshift $z_{\rm re,full}$ for full reionization (given by $\mathcal{R}(z_{\rm re,full}) = 6.25$) for each model is also given in the corresponding panel. The results for the Lacey16 model are shown in the upper left panel. With the above mentioned criterion, this model predicts $z_{\rm re,half} = 6.3$, which is too low compared to the observational estimate. This indicates that in the Lacey16 model, star formation at high redshift, $z \gtrsim 8$, is suppressed too much. There are two possible reasons for this oversuppression: one is the SN feedback at high redshift is too strong, and the other is that the SN feedback in low-mass galaxies is too strong (since the typical galaxy mass is lower at higher redshift).

Fig. 3.4 shows the cumulative luminosity function of satellite galaxies in Milky Way-like host halos. In each panel, the red solid and dashed lines show the simulation results for the corresponding model. For each model, the simulations were run on 100 separate merger trees for each of 5 host halo masses, evenly spaced in the logarithm of the mass in the range $5 \times 10^{11} - 2 \times 10^{12}\,{\rm M}_\odot$. This simulated sample of MW-like halos contains 500 halos in total, and the red solid line shows the median satellite LF for this sample, while the red dashed lines indicate the 5−95% range. The black solid

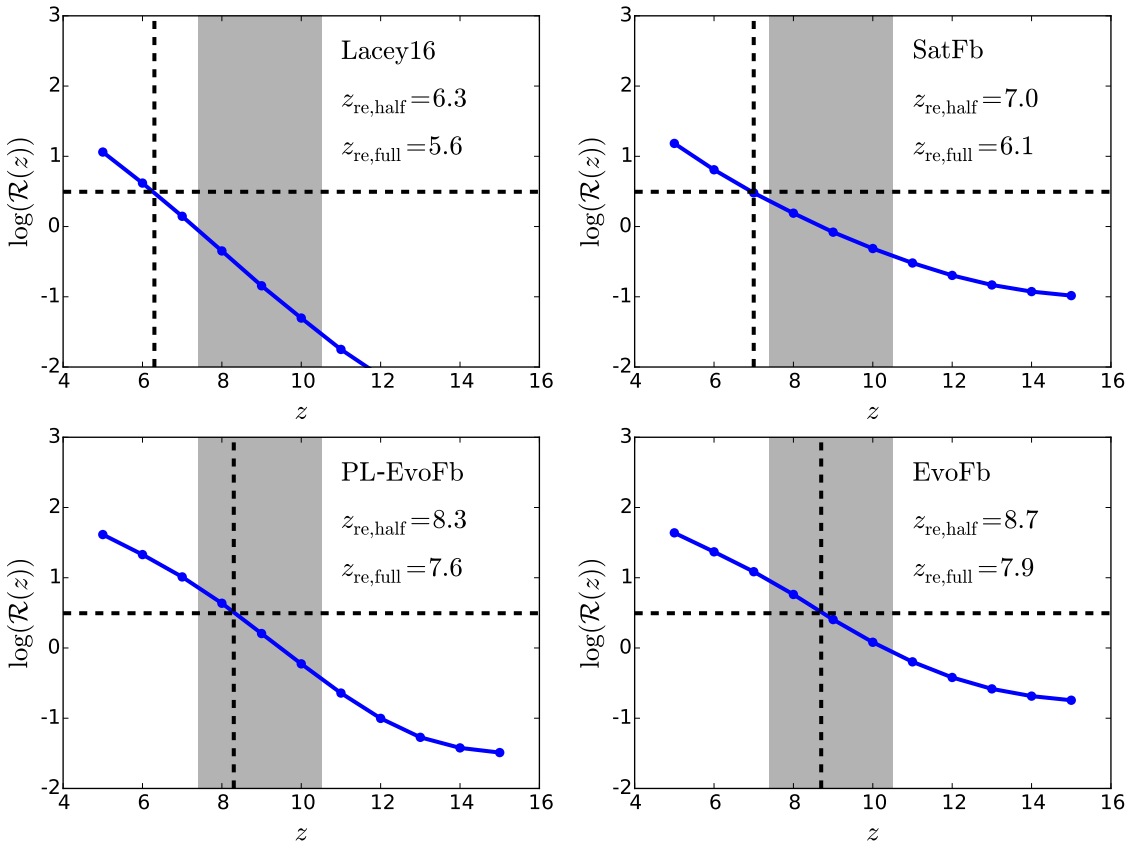

Figure 3.3: $\mathcal{R}(z)$, which is the ratio of the total number of ionizing photons produced up to redshift z to the total number of hydrogen nuclei, for different SN feedback models. Each panel shows a different model, as labelled. The blue line shows the predicted $\mathcal{R}(z)$, while the horizontal dashed line shows the threshold $\mathcal{R}(z_{\rm re,half}) = 5$ for 50% reionization, and the vertical dashed line the corresponding redshift $z_{\rm re,half}$. The grey shaded region shows the observational constraint on $z_{\rm re,half}$ from *Planck*, namely $z_{\rm re} = 8.8^{+1.7}_{-1.4}$ (68% confidence region, Planck Collaboration et al., 2015). The predicted value of the redshift $z_{\rm re,full}$ for 100% reionization is also given in each panel.

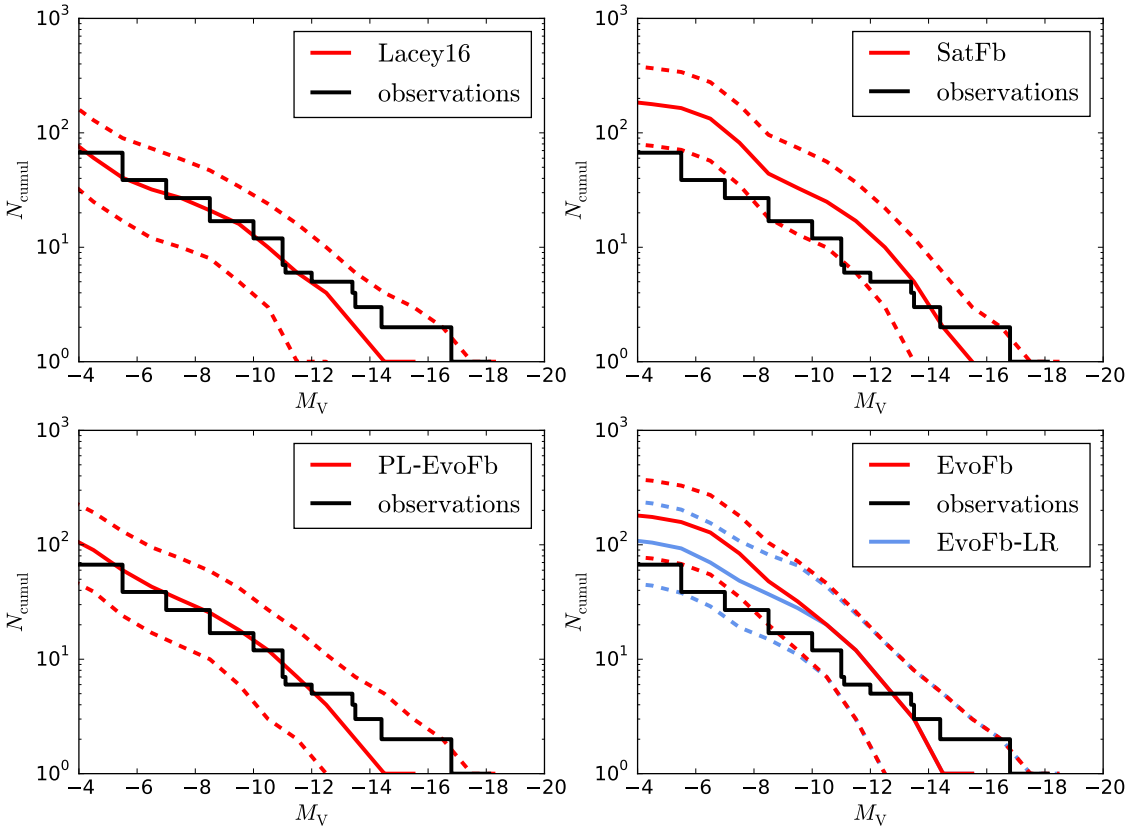

Figure 3.4: Cumulative luminosity function of satellite galaxies in Milky Way-like host halos at $z = 0$. The solid black line in each panel is the observed Milky Way satellite cumulative luminosity function. For $M_{\rm V} < -11$, this shows the direct observational results from McConnachie (2012), while for $M_{\rm V} \geq -11$, it shows the results from Koposov et al. (2008), who applies some corrections for incompleteness in the observations. The other lines in each panel show the model predictions, with the solid red line showing the median for a sample of MW-like halos, and the dashed lines indicating the $5-95\%$ range. The corresponding model names are given in the line key in each panel.

line in each panel shows the observed Milky Way satellite luminosity function. For $M_V < -11$, we plot the direct observational measurement from McConnachie (2012). For these brighter magnitudes, current surveys for MW satellites are thought to be complete over the whole sky. For $M_V \geq -11$ we plot the observational estimate from Koposov et al. (2008) based on SDSS, which includes corrections for incompleteness due to both partial sky coverage and in detecting satellites in imaging data. The predictions for the Lacey16 model are shown in the upper left panel, and are in very good agreement with the observations.

Fig. 3.5 shows the $Z_* - M_*$ correlation for satellite galaxies in Milky Way-like host halos. The sample is the same as that for Fig. 3.4. In each panel, the red solid line shows the median of the sample, while the red dashed lines indicate the $5-95\%$ range. The black filled circles in each panel show observational data. We have converted the observed [Fe/H] values into the total stellar metallicities, $Z_*/Z_\odot$, by assuming that the chemical abundance patterns in the observed satellites are the same as in the Sun. This assumption may lead to an underestimation of the metallicities of low mass satellites, which may not have had enough enrichment by Type Ia supernova to reach the Solar pattern. For these satellites, the observed Z_* values shown in the figure are therefore effectively lower limits. The results of the Lacey16 model are again shown in the upper left panel. The $Z_* - M_*$ relation predicted by this model is about an order of magnitude below the observations. Because the discrepancy in metallicity is about one order of magnitude, it cannot be caused by inaccuracies in the theoretical stellar yields of metals in this model or by the variation of these yields with stellar metallicity. These yields are obtained by integrating the yields predicted by stellar evolution models over the IMFs assumed for stars formed either quiescently or in starbursts. Assuming that the true metal yields are similar to what is assumed in the model, then for a given stellar mass, the total metals produced are fixed, so the low metallicities seen in the Lacey16 model imply that the loss of metals from satellite galaxies is excessive. Since the metal loss is caused by the outflows induced by SN feedback, this indicates that the SN feedback in these small galaxies is too strong.

In summary, the Lacey16 model motivates two types of modification to the SN

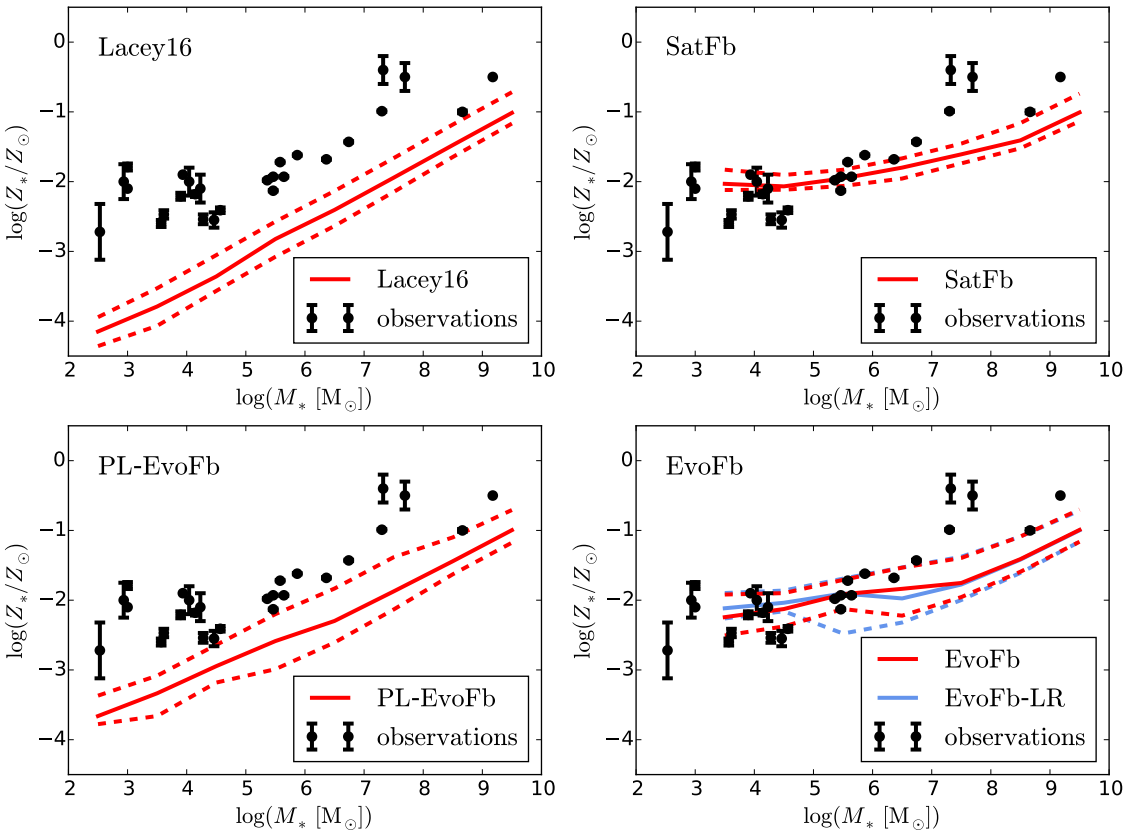

Figure 3.5: The stellar metallicity (Z_*) vs stellar mass (M_*) relation for satellite galaxies in Milky Way-like host halos at $z = 0$. The simulated sample for each model is the same as in Fig. 3.4. In each panel, the solid line shows the median of the simulated sample, while the dashed lines indicate the $5 - 95\%$ range, and the corresponding model name is given in the line key. The black filled circles show the observational results compiled by McConnachie (2012). The observed [Fe/H] values in McConnachie (2012) are converted into the total stellar metallicities, $Z_*/Z_\odot$, in Solar units by assuming the chemical patterns of the observed satellites are Solar. The total metallicities, Z_*, predicted by the model, which are absolute values, are converted into Solar units assuming $Z_\odot = 0.0142$ (Asplund et al., 2009).

feedback. One is suppressing SN feedback in small galaxies, which is the saturated feedback model. The other one is suppressing SN feedback at high redshift, i.e. $z \geq 8$, but keeping strong feedback at $z < 4$ in order to reproduce the $z = 0$ field LFs. This corresponds to the evolving feedback model. Below, these two kinds of modification will be tested one at a time.

3.3.2 Saturated feedback model (SatFb)

The dashed green lines in Fig. 3.2 show the b_J-band and K-band LFs for the SatFb model. These predictions are still roughly consistent with the observations, but a small excess of galaxies begins to appear at the very faint end of the b_J-band LF ($M_{b_J} > -17$). Reducing the SN feedback strength further in this model would exacerbate this discrepancy.

The upper right panel of Fig. 3.3 shows $\mathcal{R}(z)$ for our SatFb model; the predicted $z_{\rm re,half}$ is 7.0, outside the 1-σ region allowed by the *Planck* observations. Thus, SN feedback in the SatFb model is much too strong to allow production of enough ionizing photons to reionize the Universe early enough. The upper right panel of Fig. 3.4 shows the satellite LF of Milky Way-like galaxies in the SatFb model. The relatively weak SN feedback in this model leads to an overprediction of faint ($M_V \geq -8$) satellites. Bearing in mind the significant uncertainties in the numbers of faint satellites, this model prediction may be deemed to be roughly acceptable. Furthermore, these very faint Milky Way satellites are very small, and so their abundance could be further supressed by adjusting the strength of photoionization feedback. However, this would not help reduce the excess at the faint end of the field LFs, because these galaxies are larger and thus not strongly affected by photoionization feedback. The upper right panel of Fig. 3.5 shows the satellite $Z_* - M_*$ correlation for Milky Way-like hosts. This model prediction agrees with observations only roughly. The correlation is shallow because most of these satellites have $V_c < V_{\rm thresh} = 62\,{\rm km\,s^{-1}}$, and thus similar values of β.

If the SN feedback strength in the SatFb model were further reduced, the excess in the satellite LF would shift to brighter luminosities, $M_V < -8$, where there are fewer uncertainties in the data and where photoionization feedback is ineffective.

The stellar metallicity of satellites of a given stellar mass would become even higher, spoiling the already marginal agreement with observations. Together, these results from the Milky Way satellites suggest that the strength of SN feedback in model SatFb is a lower limit to the acceptable value.

This SatFb model therefore does not provide a solution to the problems identified in the Lacey16 model. Further adjustments within the framework of the saturated feedback model would involve changing the saturated power-law slope $\gamma'_{\rm SN}$ and/or the threshold velocity, $V_{\rm thresh}$. In the present SatFb model, as mentioned above, $\gamma'_{\rm SN}$ is already at its lower limit, namely 0, and introducing a positive $\gamma'_{\rm SN}$ only leads to a stronger SN feedback in small galaxies than in the current SatFb model, and this would not predict a high enough $z_{\rm re,half}$. Reducing $V_{\rm thresh}$ would also lead to a stronger SN feedback in small galaxies than in the current SatFb model, so would not improve the prediction for $z_{\rm re,half}$ either, while enhancing $V_{\rm thresh}$ would lead to a saturation of the SN feedback in even larger galaxies and a stronger saturation in small galaxies than in the current SatFb model. Since the feedback strength in the SatFb model is already as low as allowed by observations of the field LFs, the MW sat LF and the MW sat $Z_* - M_*$ relation, this adjustment would only worsen these discrepancies. Thus, the saturated feedback model is disfavoured by this combination of observational constraints.

3.3.3 Evolving feedback model

3.3.3.1 PL-EvoFb model

The magenta lines in Fig. 3.2 show the b_J-band and K-band field LFs for the PL-EvoFb model. The results are very close to those in the Lacey16 model, and the observed faint ends are well reproduced. This is because in the PL-EvoFb model, the SN feedback at $z \leq 4$ is the same as in the Lacey16 model. The lower left panel in Fig. 3.3 shows $\mathcal{R}(z)$ for this model; the corresponding $z_{\rm re,half}$ is 8.3, which is in agreement with observations. This shows that the evolving feedback model is more successful at generating early reionization than the saturated feedback model.

The lower left panel of Fig. 3.4 shows the satellite luminosity function of Milky

Way-like host halos in the PL-EvoFb model, which is in very good agreement with the observations. The lower left panel of Fig. 3.5 shows the $Z_* - M_*$ relation for satellite galaxies in Milky Way-like host halos in this model. This model predicts stellar metallicities of satellites with $M_* \leq 10^6\,\mathrm{M}_\odot$ several times to one order of magnitude lower than observations, with the discrepancy increasing with decreasing stellar mass. Although weakening the SN feedback at high redshifts does improve the result compared to the Lacey16 model, it is still inconsistent with observations. Thus this model is disfavoured by observations of MW satellite metallicities. The discrepancy again suggests that the SN feedback in small galaxies is too strong, but since at the same time this model successfully reproduces the faint ends of the field LFs, it suggests that this problem of too strong feedback is restricted to very small galaxies. This then motivates our next model, in which we preferentially suppress the SN feedback strength in very small galaxies, while retaining the same evolution of feedback strength with redshift as in the PL-EvoFb model.

3.3.3.2 EvoFb model

The field LFs predicted by the EvoFb model are almost identical to those given by the PL-EvoFb model, so this model likewise successfully reproduces the faint ends of the field LFs. The reason for the similarity between the field LFs predicted by the two models is that the saturation introduced in the EvoFb model is only effective for $V_\mathrm{c} \leq 50\,\mathrm{km\,s^{-1}}$, and would not significantly affect the galaxies in the observed faint ends of the field LFs, which typically have higher V_c.

The lower right panel in Fig. 3.3 shows $\mathcal{R}(z)$ for the EvoFb model; the corresponding $z_\mathrm{re,half}$ is 8.7, which is in agreement with the observations. Compared to the result of the PL-EvoFb model, $z_\mathrm{re,half}$ only increases slightly, so the saturation in the feedback has only a small effect, and the main factor leading to the agreement with observations is still the redshift evolving behavior of the SN feedback strength.

The lower rigth panel of Fig. 3.4 shows the satellite luminosity function of Milky Way-like host halos in the EvoFb model. The model predictions are roughly consistent with the observations, although the very faint end ($M_\mathrm{V} \geq -8$) of the MW sat LF is somewhat too high. However, as mentioned in connection with the SatFB

model, the observations of this very faint end have significant uncertainties, so this model is still acceptable. The lower right panel of Fig. 3.5 shows the $Z_* - M_*$ relation for the satellite galaxies in Milky Way-like host halos in the EvoFb model. The model predictions are now roughly consistent with the observations. This improvement is achieved by adopting both an evolving SN feedback strength and a saturation of the feedback in galaxies with $V_c \leq 50\,\mathrm{km\,s^{-1}}$.

Because the predictions for Milky Way satellites are sensitive to the photoionization feedback, it is possible to further improve the agreement with observations for these galaxies by adjusting the photoionization feedback. One possible adjustment is to adopt the so-called local reionization model (see Font et al. (2011) and references therein), in which higher density regions reionize earlier, so that $z_{\mathrm{re,full}}$ for the Local Group region is earlier than the global average $z_{\mathrm{re,full}}$ constrained by the *Planck* data. Earlier reionization means earlier photoionization feedback, so that for the Milky Way satellites one has $z_{\mathrm{crit}} > z_{\mathrm{re,full}}$. Font et al. (2011) adopted a detailed model to study this local reionization effect, and suggested that using $z_{\mathrm{crit}} = 10$ gives a good approximation to the results of the more detailed model. Here we also adopt $z_{\mathrm{crit}} = 10$, and we label the model with evolving SN feedback and $z_{\mathrm{crit}} = 10$ as EvoFb-LR.

We tested that the predictions for global properties like $z_{\mathrm{re,full}}$, $z_{\mathrm{re,half}}$ and the field LFs are not very sensitive to the value of z_{crit}. It is therefore justified to ignore the variation of z_{crit} with local density when calculating these global properties, and adopt a single $z_{\mathrm{crit}} = z_{\mathrm{re,full}}$ when predicting these. This also means that introducing such a local reionization model does not allow one to bring the standard Lacey16 or SatFb models into agreement with all of our observational constraints, since some of the discrepancies described above involve these global properties.

The satellite luminosity function of the Milky Way-like host halos in the EvoFb-LR model is also shown in the lower right panel of Fig. 3.4. The model predictions agree with observations better than the EvoFb model, because the abundance of the very faint satellites is reduced by the enhanced photoionization feedback. The $Z_* - M_*$ relation for satellite galaxies in Milky Way-like host halos for the EvoFb-LR model is very similar to that of the EvoFb model, shown in Fig. 3.5.

3.4 discussion

3.4.1 Why should the SN feedback strength evolve with redshift?

The physical idea behind formulating the mass loading factor, β, of SN-driven outflows (Eq 3.2.1) as a function of V_c is that the strength of the SN feedback driven outflows (for a given star formation rate, ψ) depends on the gravitational potential well, and V_c is a proxy for the depth of the gravitational potential well. However, in reality the strength of outflows does not only depend on the gravitational potential well, but may also depend on the galaxy gas density, gas metallicity and molecular gas fraction. This is because the gas density and metallicity determine the local gas cooling rate in the ISM, which determines the fraction of the injected SN energy that can finally be used to launch outflows, while the dense molecular gas in galaxies may not be affected by the SN explosions, and thus may not be ejected as outflows. These additional factors may evolve with redshift, and V_c may not be a good proxy for them, so if the outflow mass loading factor, β, is still formulated as a function of V_c only, a single function may not be valid for all redshifts and some redshift evolution of β may need to be introduced.

The detailed dependence of β on the galaxy gas density, gas metallicity and molecular gas fraction can only be derived by using a model which considers the details of the ISM. The model of Lagos et al. (2013) is an effort towards this direction, and the dependence of β on V_c predicted by that model is shown in Fig. 15 of that paper. But since the model in Lagos et al. (2013) only considers ejecting gas out of galaxies, but does not predict what fraction of this escapes from the halo, the model is incomplete. We therefore only use very general and rough features of the dependence of β on V_c and z predicted by Lagos et al. (2013) to motivate our PL-EvoFb and EvoFb models, which assume a redshift-dependent β.

Lagos et al. (2013) suggest that the mass loading, β, is weaker in starbursts than for quiescent star formation in galaxy disks, because starbursts have higher gas density and molecular gas fraction. While this feature is not included in our model, as it may be too complex for a phenomenological SN feedback models, it

has the potential to enhance the reionization redshift and the stellar metallicities of galaxies, so it might be worth investigating it in future work.

3.4.2 What kind of galaxies reionized the Universe?

Fig. 3.6 shows some simple statistics of the galaxies producing the ionizing photons. The first row shows the statistics of the stellar mass, M_*, of the galaxies producing ionizing photons, the second row shows the statistics of the dust-extincted rest-frame far-UV absolute magnitude, $M_{\rm AB}(1500\text{Å})$, of these galaxies, while the third row shows the statistics of the halo masses, $M_{\rm halo}$, and the fourth row the statistics of the galaxy circular velocity, $V_{\rm c}$. For each quantity, the dots in each panel indicate the medians of the corresponding quantity, and the error bars indicate the $5-95\%$ range, with the medians and percentiles determined not by the number of galaxies but by their contributions to the ionizing emissivity at that redshift. The median means that galaxies below it contribute 50% of the ionizing photons at a given redshift, while the $5-95\%$ range indicates that the galaxies within it contribute 90% of the ionizing photons at a given redshift. Each column corresponds to a different SN feedback model. The vertical dashed lines in each panel indicate $z_{\rm re,full}$, the redshift at which the Universe is fully ionized, for that model, with the numerical values of $z_{\rm re,full}$ given in the panels in the first row.

From Fig. 3.6 it is clear that the median of M_* at $z\sim z_{\rm re,full}$ for each SN feedback model is around $10^8-10^9\,{\rm M}_\odot$, the median of $M_{\rm AB}(1500\text{Å})$ is around $-17--19$, and the median of $V_{\rm c}$ is around $100-200\,{\rm km\,s^{-1}}$. These values indicate that the corresponding galaxies are progenitors of large massive galaxies at $z=0$. This means in these models, the progenitors of large galaxies make significant contributions to the cosmic reionization. It is also true that the progenitors of large galaxies have already made contributions to the ionizing photons when the Universe was half ionized, i.e. by $z=z_{\rm re,half}$. This means that a preferential suppression of the SN feedback in very small galaxies is not very effective in boosting $z_{\rm re,half}$, and to predict a high enough $z_{\rm re,half}$ by these means usually requires heavy suppression of the SN feedback in very small galaxies, which spoils the agreement with observations of faint galaxies at $z=0$. This is the reason for the failure of the SatFb model to

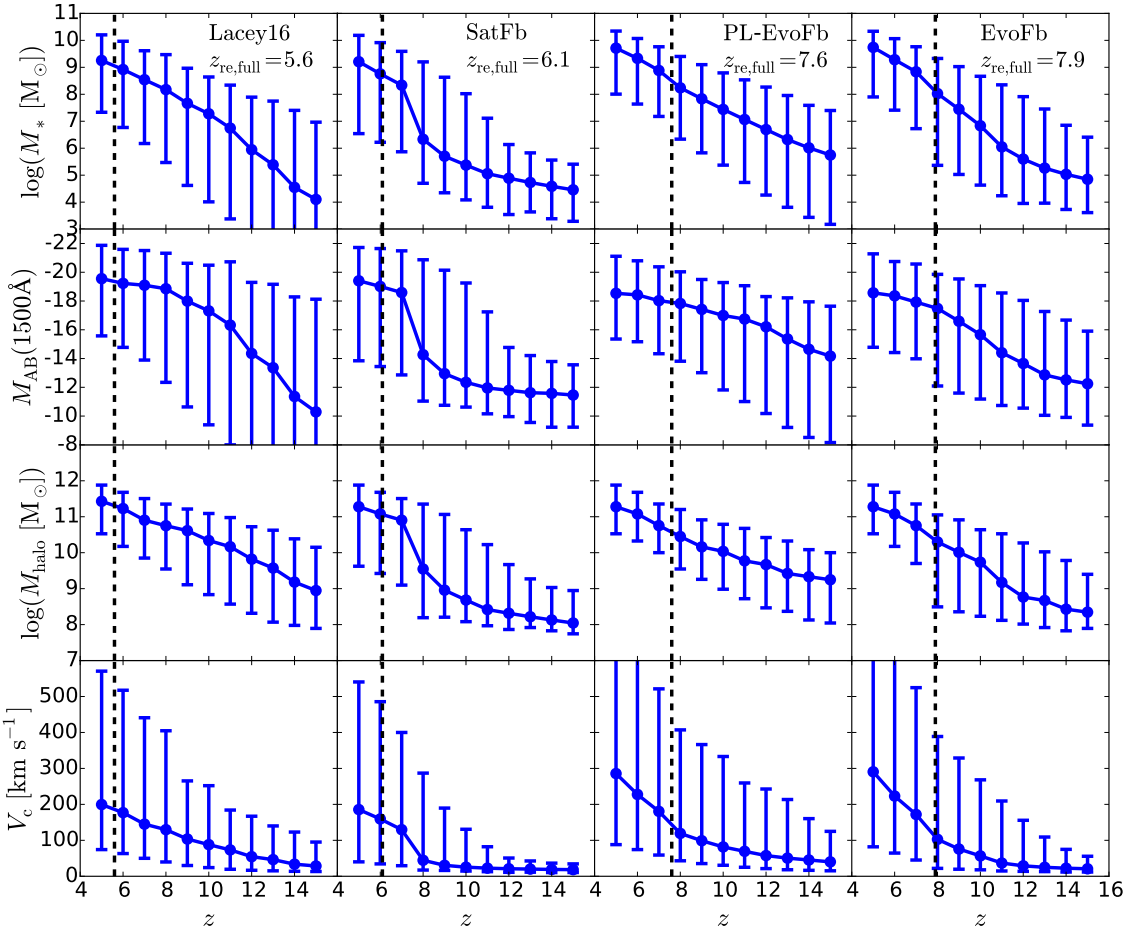

Figure 3.6: Simple statistics of the galaxies producing ionizing photons. Each column corresponds to a different SN feedback model, with the corresponding model name given in the top of each panel in the first row, along with the value of $z_{\rm re,full}$, the redshift at which the Universe is fully ionized. The vertical dashed lines in each panel also indicate $z_{\rm re,full}$. The first row shows the statistics of the stellar mass, M_*, of the galaxies producing ionizing photons, the second row shows the statistics of the dust-extincted rest frame UV magnitude, $M_{\rm AB}(1500{\rm \AA})$, of these galaxies, while the third row shows the statistics of the halo masses, $M_{\rm halo}$ and the fourth row shows the statistics of the galaxy circular velocity, $V_{\rm c}$. For each quantity shown in these rows, the dots indicate the medians of the corresponding quantity, and the errorbars the $5-95\%$ range, with both the medians and the $5-95\%$ ranges being determined by their contributions to the ionizing photon emissivity at that redshift.

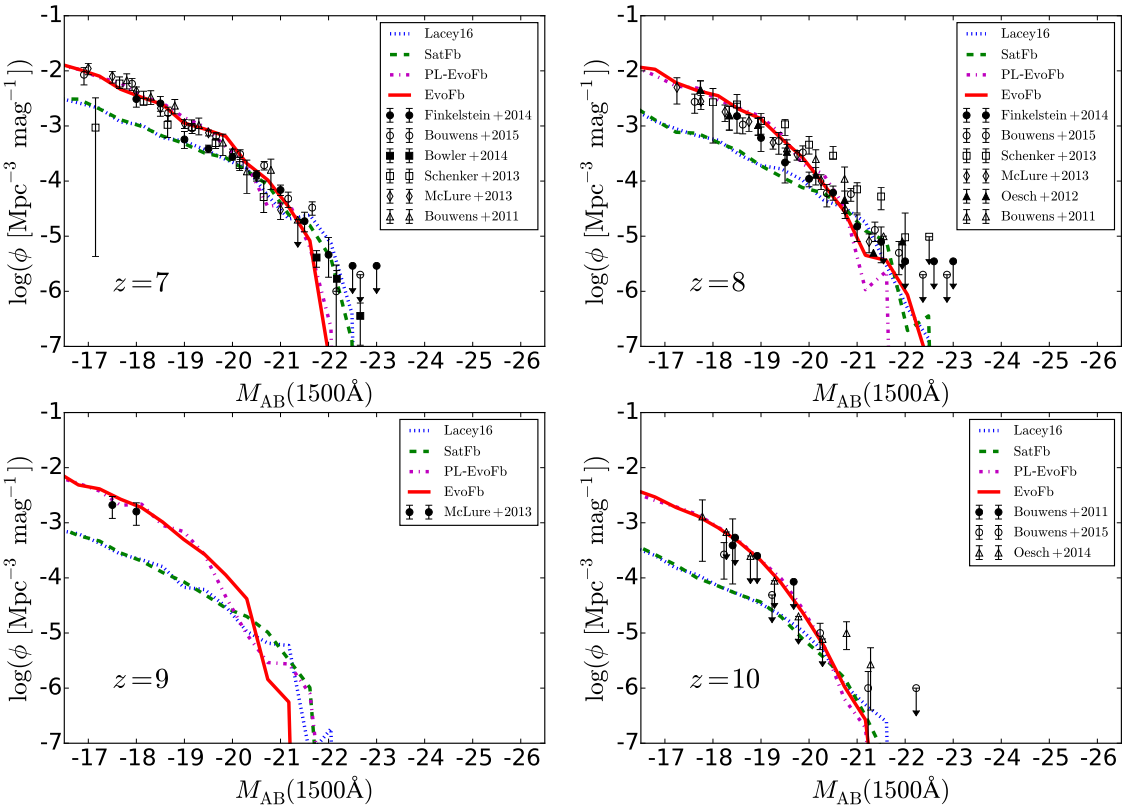

Figure 3.7: The rest frame far-UV luminosity functions at $z = 7$, 8, 9, 10 for the 4 different SN feedback models. In each panel, the blue dotted line shows the prediction for the Lacey16 model, the dashed green line that for the SatFb model, the magenta dotted dashed line that for the PL-EvoFb model and the red solid line that for the EvoFb model, while symbols with errorbars indicate observational measurements (Bouwens et al., 2011b,a; Oesch et al., 2012; Schenker et al., 2013; McLure et al., 2013; Finkelstein et al., 2014; Bowler et al., 2014; Oesch et al., 2014; Bouwens et al., 2015). The dust extinction is calculated self-consistently based on galaxy gas content, size and metallicity (see Lacey et al. (2016) for more details).

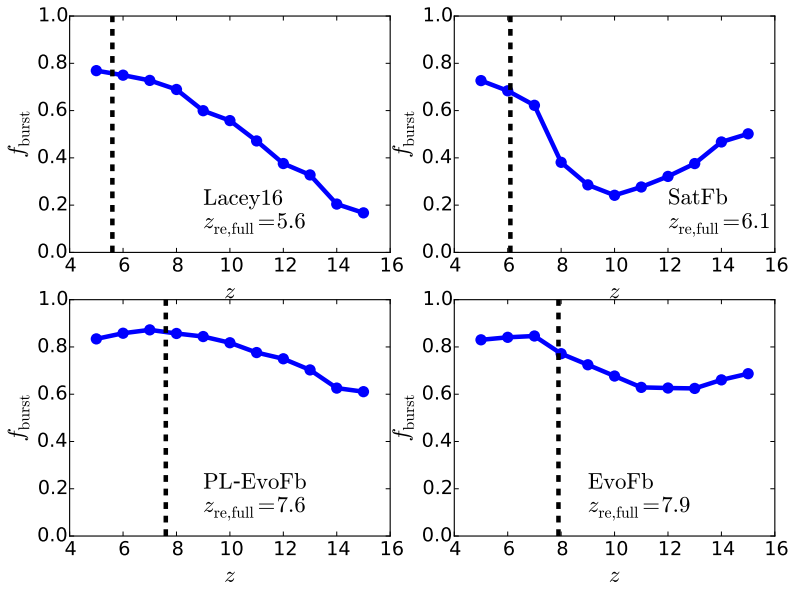

Figure 3.8: The fraction of the ionizing photon emissivity contributed by starbursts at a given redshift. Different panels are for different SN feedback models, as labelled, and the vertical dashed lines indicate $z_{\rm re,full}$.

satisfy all the observational constraints considered in this work.

Fig. 3.6 also shows that the median of $M_{\rm halo}$ in each SN feedback model is roughly in the range $10^{10}-10^{11}\,{\rm M}_\odot$ at $z\sim z_{\rm re,full}$, which means there are significant contributions to the ionizing photons from large atomic hydrogen cooling halos. This is consistent with the results from Boylan-Kolchin et al. (2014), who show that it is difficult to obtain reionization at $z\sim 8$ mainly from star formation in small atomic cooling halos with $M_{\rm halo}\sim 10^8\,{\rm M}_\odot$.

We also calculated the rest-frame far-UV luminosity functions at $z=7,\ 8,\ 9,\ 10$ for our 4 different SN feedback models. These predictions are shown in Fig. 3.7, and compared with recent observational data. The best fit (EvoFb) model is seen to agree quite well with the observations over the whole range $z=7-10$. The PL-EvoFb model, which adopts similar redshift evolving SN feedback, also reaches similar level of agreement with observations. On the other hand, the other 2 models, which generally have stronger SN feedback at high redshift than the EvoFb model, predict too few low UV luminosity galaxies at $z=7-10$. Note that the current observational limit is $M_{\rm AB}(1500\text{Å})\sim -17--18$ at these redshifts, which is close to

the median of $M_{\rm AB}(1500\text{Å})$ at reionization for the EvoFb model shown in Fig. 3.6 (for this model, at $z = 8$, the median is $M_{\rm AB}(1500\text{Å}) = -17.5$, and the $5-95\%$ range is $M_{\rm AB}(1500\text{Å}) = -12.1$ to $M_{\rm AB}(1500\text{Å}) = -19.8$.). Thus the best fit model suggests that the currently observed high redshift galaxy population should contribute about half of the ionizing photons that reionized the Universe. This is consistent with Kuhlen & Faucher-Giguère (2012), which suggests that the sources of reionization can not be too heavily dominated by very faint galaxies.

We also checked that the rest-frame far-UV luminosity functions predicted by all 4 models become very similar at $z \leq 6$, and thus the modifications to the SN feedback do not spoil the good agreement of these luminosity functions with observations at $3 \leq z \leq 6$ found in the original Lacey16 model.

Fig. 3.8 shows the fraction of the ionizing photons that are contributed by starbursts at a given redshift (as compared to stars formed quiescently in galaxy disks). Different panels are for different SN feedback models, and the vertical dashed lines indicate $z_{\rm re,full}$ for the corresponding models. It is clear that at $z \sim z_{\rm re,full}$, the starburst fractions are high, with $f_{\rm burst} \approx 0.8$ in all four models. This indicates that starbursts are a major source of the ionizing photons for cosmic reionization.

3.4.3 The descendants of the galaxies that ionized Universe

For the best fit model, i.e. the EvoFb model, we also identified the $z = 0$ descendants of the galaxies which ionized the Universe. To do this, we ran a simulation with fixed dark matter halo mass resolution $M_{\rm res} = 7.1 \times 10^7\,{\rm M}_\odot$ from $z = 20$ to $z = 0$. This $M_{\rm res}$ is low enough to ensure that we resolve all the atomic cooling halos up to $z = 11$. According to Fig. 3.3, most of the ionizing photons that reionized the Universe are produced near $z_{\rm re,full}$, and for the EvoFb model, $z_{\rm re,full} = 7.9$. Thus, resolving all the atomic cooling halos up to $z = 11$ ensures that all galaxies which are major sources of the ionizing photons and their star formation histories are well resolved.

In Fig. 3.9 we show the mass distributions of the $z = 0$ descendants of the objects which produced the photons which reionized the Universe, weighted by the number of ionizing photons produced. The top panel shows the stellar mass of the

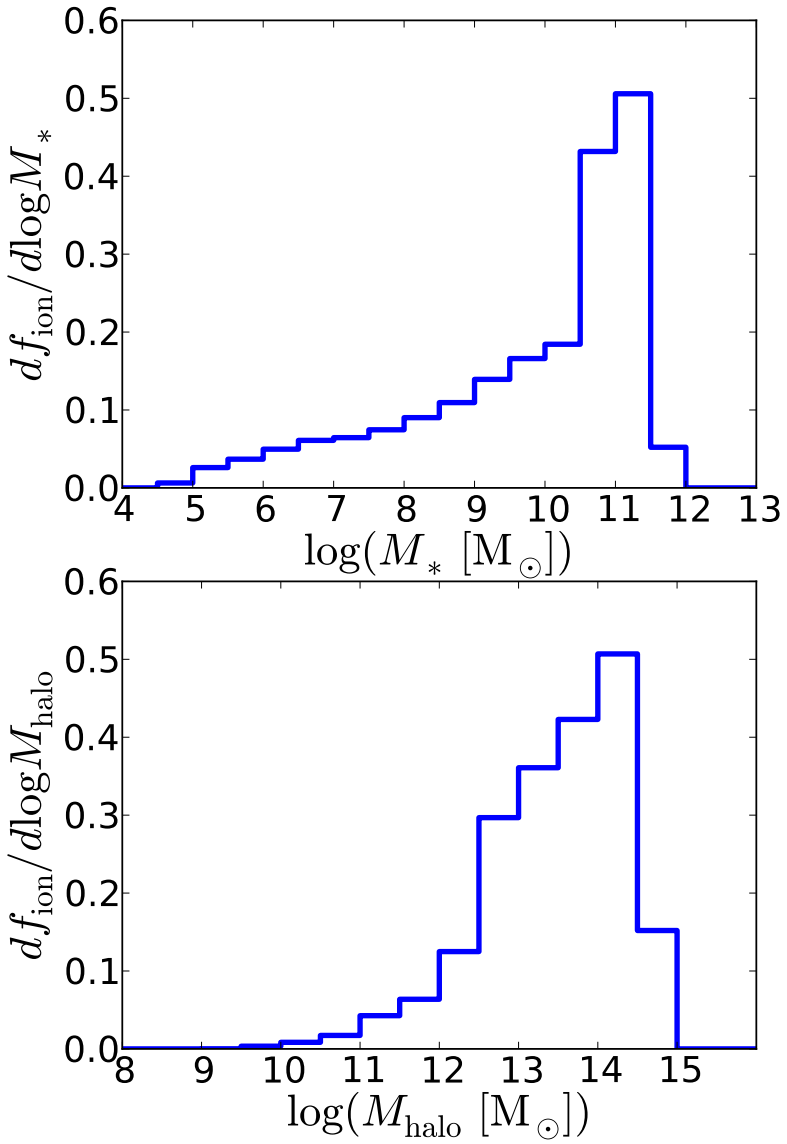

Figure 3.9: Probability distributions of masses of $z = 0$ descendants of objects which emit ionizing photons at $z \geq z_{\rm re,full}$, weighted by number of ionizing photons produced. *Upper panel:* Probability distribution of stellar mass of descendant at $z = 0$. *Lower panel:* Probability distribution of halo mass of descendant at $z = 0$.

descendant galaxy, while the bottom panel shows the mass of the descendant dark matter halo. To calculate these, we effectively identify each ionizing photon emitted at $z \geq z_{\rm re,full}$, then identify the $z = 0$ descendant (galaxy or halo) of the galaxy which emitted it, then construct the probability distribution of descendant mass, giving equal weight to each ionizing photon. The upper panel of Fig. 3.9 shows that over 50% of the ionizing photons are from the progenitors of large galaxies with $M_* > 3 \times 10^{10}\,{\rm M}_\odot$, or equivalently, the major ionizing sources have $z = 0$ large galaxies as their descendants. The lower panel of Fig. 3.9 shows that 50% of the ionizing photons are from the progenitors of high mass dark matter halos at $z = 0$ with $M_{\rm halo} > 3.7 \times 10^{13}\,{\rm M}_\odot$, which means that the reionization is driven mainly by sources at very rare density peaks. These results are consistent with the indications given by Fig. 3.6.

In Fig. 3.10, we show the fraction of stellar mass in galaxies at $z = 0$ that was formed before reionization, i.e. at $z \geq z_{\rm re,full}$, for the best fit model (the EvoFb model). The upper panel shows this for all galaxies, while the lower panel shows this quantity only for galaxies in Milky Way-like halos, defined as halos with $z = 0$ halo mass in the range $5 \times 10^{11}\,{\rm M}_\odot \leq M_{\rm halo} \leq 2 \times 10^{12}\,{\rm M}_\odot$. The upper panel shows that even though the progenitors of the $z = 0$ large galaxies provided about half of the ionizing photons, only a tiny fraction of their stars are formed before reionization, and while the $z = 0$ dwarf galaxies ($M_*(z = 0) < 10^6\,{\rm M}_\odot$) contributed only a small fraction of the photons for reionization, their stellar populations typically are dominated by the stars formed before reionization. This is consistent with the hierarchical structure formation picture, because smaller objects formed earlier, and also formation of galaxies in small halos is suppressed after reionization by photoionization feedback. Also note that the ratio of the mass of the stars formed at $z \geq z_{\rm re,full}$ to the $z = 0$ stellar mass shows considerable scatter for galaxies with $M_*(z = 0) < 10^7\,{\rm M}_\odot$, which means the star formation histories of these small galaxies are very diverse.

The lower panel of Fig. 3.10 shows galaxies in Milky Way-like halos only, but the predicted fraction of stars formed before reionization is in fact very similar to the average over all halos shown in the upper panel. For reference, the short vertical solid

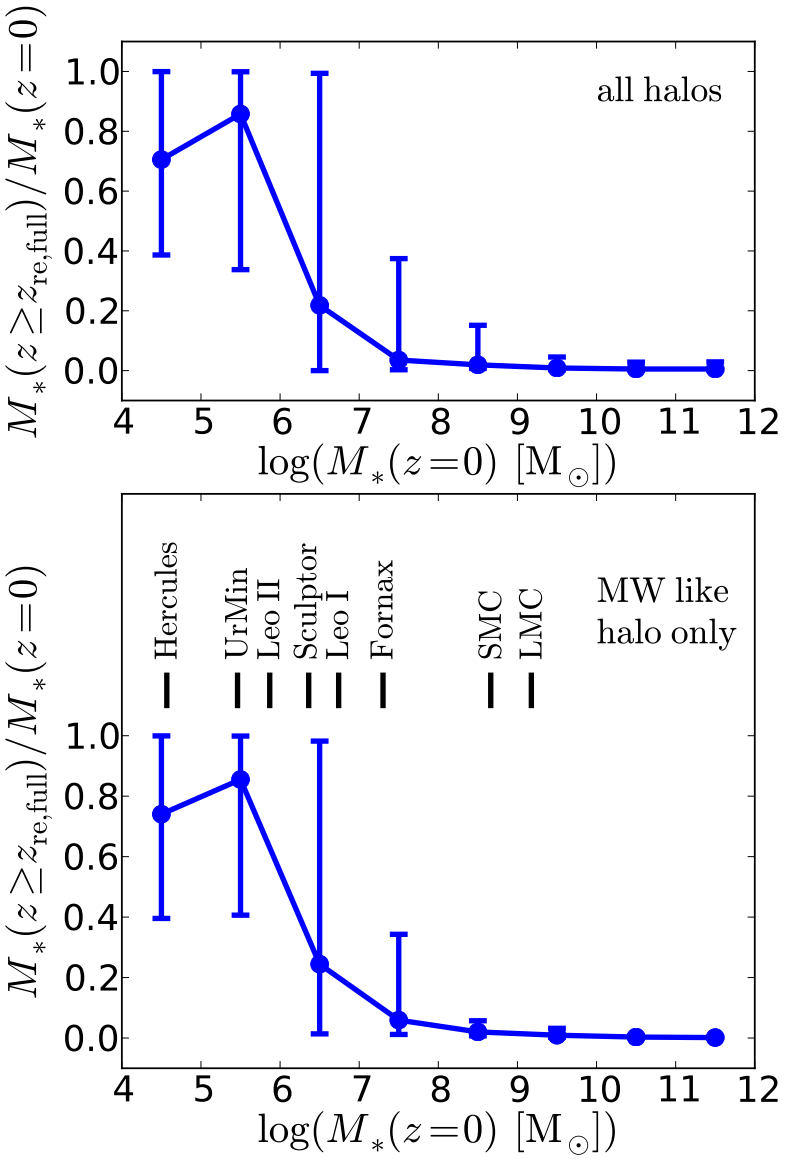

Figure 3.10: Fraction of stellar mass in galaxies at $z = 0$ which was formed before reionization (i.e. at $z \geq z_{\rm re,full}$). In both panels, each filled circle shows the median of the ratio in the corresponding $z = 0$ stellar mass bin, while the corresponding error bar indicates the $5-95\%$ range of this ratio. *Upper panel:* All galaxies. *Lower panel:* Galaxies in Milky Way-like halos only (defined as halos with $z = 0$ halo mass in the range $5 \times 10^{11}\,{\rm M}_\odot \leq M_{\rm halo} \leq 2 \times 10^{12}\,{\rm M}_\odot$). The short vertical solid black lines indicate the observed stellar masses of several Milky Way satellites, namely LMC, SMC, Fornax, Sculptor, Leo I, Leo II, Ursa Minor (UrMin) and Hercules, for reference (from McConnachie 2012).

black lines indicate the observed stellar masses of several Milky Way satellites (from McConnachie 2012), namely LMC, SMC, Fornax, Sculptor, Leo I, Leo II, Ursa Minor (UrMin) and Hercules. As shown by this panel, the best fit model implies that for the large satellites like the LMC, SMC and Fornax, only tiny fractions of their stellar mass, typically 5% or less, were formed before reionization. However, this fraction increases dramatically with decreasing satellite mass, as does the scatter around the median. For the lowest mass satellites, with stellar mass $M_* < 10^6\,\mathrm{M}_\odot$, including objects like Leo II, Ursa Minor and Hercules, the median fraction increases to around 80%, meaning that most of the satellites in this mass range form the bulk of their stars before reionization, with the 5–95% range in this fraction extending from 40% to 100%, indicating diverse star formation histories for different satellites of the same mass. Satellites in the intermediate mass range $10^6\,\mathrm{M}_\odot \leq M_* < 10^7\,\mathrm{M}_\odot$, like Leo I and Sculptor, have somewhat lower median fractions formed before reionization, around 20–50%, but with an even larger scatter around this median, with the 5–95% range extending nearly from 0% to 100%.

3.4.4 Modelling uncertainties

An important assumption in our study is that $f_{\rm esc}$ is constant and, in our default model, equal to 0.2. This choice is justified in Section 3.2.3; here we explore the effects of varying this parameter. We also explore the effect of including a contribution from AGN to the photoionizing budget, which in our standard model we assume to be negligible.

Madau & Haardt (2015) have recently revived the old idea that photons produced by AGN could be responsible for reionization. They take the observed AGN Lyman limit emissivity, ϵ_{912}, at $z \leq 6$ and extrapolate it to $z \approx 12$. Assuming that the AGN UV spectrum is a power law with index -1.7, they calculate the number of ionizing photons emitted by AGN per unit time per unit comoving volume, $\epsilon'_{\rm AGN}$. The redshift of reionization can then be obtained either by solving the equation for $Q_{\rm HII}$, or using the simpler method we introduced in §3.2.3. Madau & Haardt (2015) conclude that AGN alone could have been the dominant source of the photons respsonsible for reionization.

The estimate of ϵ_{912} at $z \approx 6$ has a large errorbar and so a major uncertainty in the model of Madau & Haardt (2015) is their extrapolation to higher redshifts. They extrapolate using a complex functional form that, however, is close to an exponential, $\epsilon_{912} \propto \exp(\mathrm{k_{AGN}\,z})$, at $z \geq 5$, the regime relevant to hydrogen reionization. To assess the plausibility of the Madau & Haardt (2015) model, we investigate other extrapolations of ϵ_{912}, which are consistent with the measured value at $z = 6$. We consider the same exponential form, constrained in all cases to lie within the errorbar of the measured value at $z = 6$ and to give the same value at $z = 5$ as the Madau & Haardt (2015) model. (Unlike Madau & Haardt, for simplicity, we extrapolate to $z = \infty$ rather than to $z \approx 12$, as they do, but this overestimate of the AGN contribution introduces only very small changes to the redshift of reionization.) These two requirements result in a family of extrapolated estimates, with $-1.92 \leq k_{\rm AGN} \leq -0.15$, illustrated by the grey shaded region in the upper left panel of Fig. 3.11. The emissivity assumed by Madau & Haardt (2015) lies at the upper boundary of this allowed region. Following Madau & Haardt we adopt an escape fraction of 100% for AGN. There is considerable uncertainty on this parameter as well (see Madau & Haardt (2015) for further discussion).

Once $\epsilon'_{\rm AGN}$ is known, the calculation in §3.2.3 can be extended to include AGN. Specifically, we have,

$$\epsilon_{\rm tot}(z) = f_{\rm esc}\epsilon_{\rm star}(z) + \epsilon_{\rm AGN}(z) \tag{3.4.1}$$

$$\mathcal{R}'(z) = \frac{\int_z^\infty \epsilon_{\rm tot}(z')\,dz'}{n_{\rm H}} \tag{3.4.2}$$

$$\mathcal{R}'(z_{\rm re,full}) = 1 + N_{\rm rec}, \tag{3.4.3}$$

where $\epsilon_{\rm star}$ is the emissivity of the stars, which is given by GALFORM, $f_{\rm esc}$ is the corresponding escape fraction, $n_{\rm H}$ is the comoving number density of hydrogen nuclei, $N_{\rm rec} = 0.25$ is the mean number of recombinations per hydrogen nucleus up to $z_{\rm re,full}$, and $\epsilon_{\rm AGN}$ is the AGN photon emissivity per unit redshift, which is related to $\epsilon'_{\rm AGN}$ by $\epsilon_{\rm AGN} = \epsilon'_{\rm AGN} dt/dz$. The redshift of at which reionization is 50% complete, $z_{\rm re,half}$, is calculated as in Eq(3.4.3), but for half the threshold. To explore the effect of different assumptions for $f_{\rm esc}$, we allow this parameter to vary in the range $0 - 0.25$.

Fig. 3.11 shows the effect of varying the AGN contribution (by varying $k_{\rm AGN}$) and $f_{\rm esc}$ on $z_{\rm re,half}$. We consider three models: Lacey16, SatFb and EvoFb, as indicated in the corresponding legends. The contour lines show the predicted values of $z_{\rm re,half}$ in each model and the shaded area shows the region consistent with the *Planck* data. The PL-EvoFb model is not considered here because it is disfavoured by the MW satellite metallicity data.

As we have seen, stars in the Lacey16 model do not produce enough ionizing photons to reionize the Universe sufficiently early; AGN can reionize the Universe in this model but only if their emissivity has a very flat slope, $-0.25 \leq k_{\rm AGN} \leq -0.15$; this extreme region is illustrated in the upper left panel of Fig. 3.11 as the red hatched area. The SatFb model also requires an AGN contribution in order to be consistent with the the values of $z_{\rm re,half}$ allowed by the *Planck* data, but this is generally less than required for the Lacey16 model. For our fiducial value of $f_{\rm esc} = 0.2$, the required AGN emissivity corresponds to $-0.49 \leq k_{\rm AGN} \leq -0.15$; this region is the blue hatched area in the upper left panel of Fig. 3.11. For lower values of $f_{\rm esc}$, the required range of $k_{\rm AGN}$ shrinks and comes close to the allowed upper limit. Finally, the EvoFb model is consistent with the *Planck* data in the case where all ionizing photons are produced by stars so long as $f_{\rm esc} \geq 0.07$; of course adding an AGN contribution makes it easier to reionize the Universe for even lower values of $f_{\rm esc} < 0.07$.

In summary, even if AGN make a contribution to the ionizing photon budget, as long as $k_{\rm AGN} < -0.25$, our original, single power-law SN feedback model is incompatible with the *Planck* data. If $k_{\rm AGN} < -0.49$ and $f_{\rm esc} \geq 0.07$, then the evolving feedback model is preferred to the saturated feedback model, and our major conclusions regarding SN feedback still apply. Note that when $k_{\rm AGN} > -0.49$, the reionization redshift alone cannot discriminate between the SatFb and EvoFb models, but the measured far-UV galaxy luminosity functions at $z = 7-10$ (Fig. 3.7) still prefer the evolving feedback model.

Our earlier conclusions regarding the sources of reionizing photons and their descendants are only valid when stars are the dominant source of reionizing photons. The contour lines in Fig. 3.12 show the fraction of the total ionizing photon budget

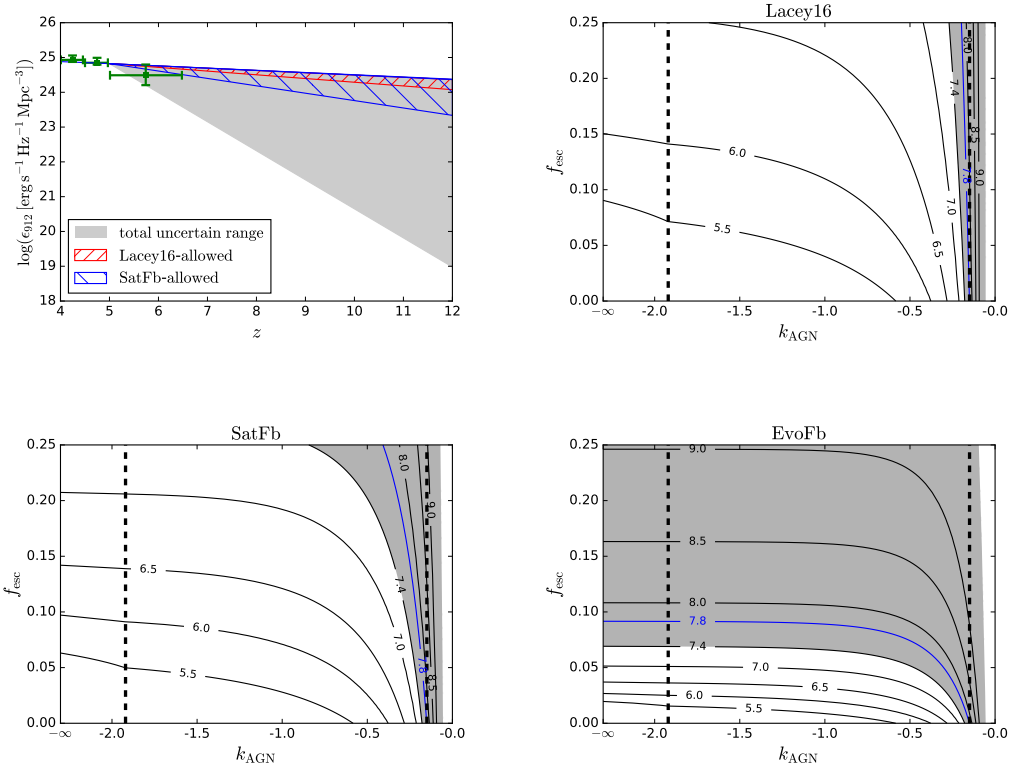

Figure 3.11: *Upper left panel:* extrapolations of the AGN emissivity at the Lyman limit, ϵ_{912}, allowed by the errorbar of the measurement at $z \simeq 6$. The three data points with errorbars are the observations taken from Fig. 1 of Madau & Haardt (2015), the grey shaded region shows the allowed extrapolations. The extrapolation adopted by Madau & Haardt (2015) lies on the upper boundary of the region; the red and blue hatched regions encompass the extrapolations required to bring the Lacey16 and SatFb models respectively into agreement with the *Planck* constraints, for $f_{\rm esc} = 0.2$. Reducing $f_{\rm esc}$ shrinks these regions towards the upper boundary. *Remaining three panels:* predicted $z_{\rm re,half}$ (contour lines) for different combinations of $f_{\rm esc}$ and $k_{\rm AGN}$, where $f_{\rm esc}$ is the escape fraction for stars and $k_{\rm AGN}$ is the slope of the AGN emissivity extrapolation shown in the upper left panel. The panels correspond to the Lacey16, SatFb and EvoFb models, as labelled. The grey shaded area in each panel represents the region allowed at 1-σ by the *Planck* data. The vertical dashed lines indicate the lower and upper limits of the extrapolation slope, i.e. $k_{\rm AGN} = -1.92$ and $k_{\rm AGN} = -0.15$. The line labelled $k_{\rm AGN} = -\infty$ corresponds to the case of no AGN contribution.

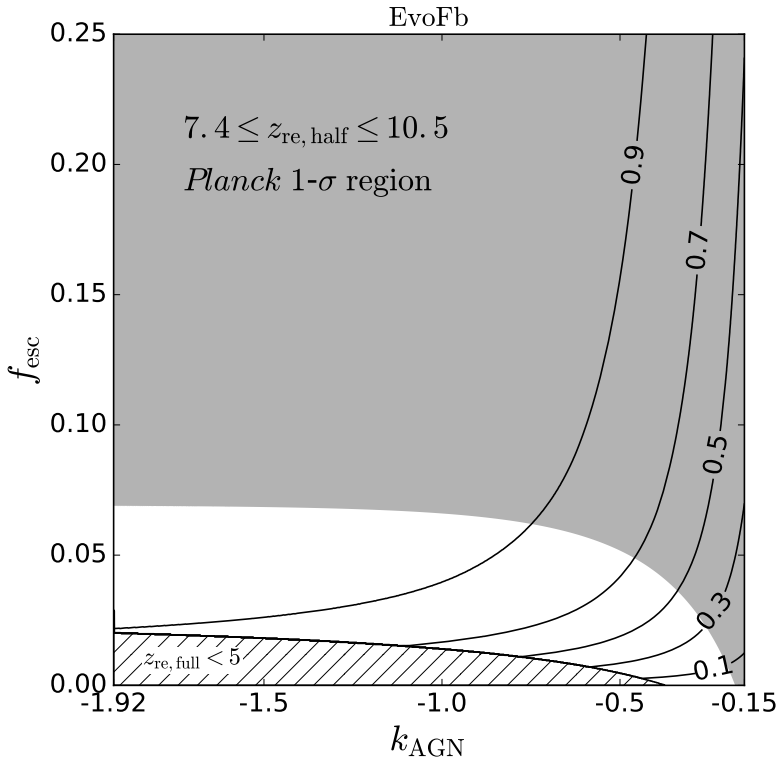

Figure 3.12: The fraction of ionizing photons from stars for different combinations of f_{esc} and k_{AGN} for our best-fit model, EvoFb. This photon budget includes all ionizing photons emitted from $z = \infty$ to $z_{\mathrm{re,full}}$. The fractions are shown as contour lines. The lower hatched region corresponds to $z_{\mathrm{re,full}} < 5$ and is strongly excluded by other observations. The upper grey shaded region is allowed by the *Planck* data.

produced by stars for different combinations of f_{esc} and k_{AGN}. This photon budget includes all ionizing photons emitted from $z = \infty$ to $z_{\mathrm{re,full}}$. This is only shown for the EvoFb model, because this is our best-fit model and thus the most relevant to a discussion of reionization sources and their descendants. As the figure shows, so long as $f_{\mathrm{esc}} > 0.07$ and $k_{\mathrm{AGN}} < -0.75$, over 90% of the ionizing photons required for reionization come from stars; this fraction drops to 70% if $k_{\mathrm{AGN}} = -0.49$, but is still dominant. Thus, our earlier conclusions regarding the reionization sources and their descendants remain valid so long as $f_{\mathrm{esc}} > 0.07$ and $k_{\mathrm{AGN}} < -0.49$.

3.5 summary

We have investigated what constraints can be placed on supernova (SN) feedback by combining a physical model of galaxy formation with critical observations which constrain the strength of feedback in opposite directions. The observational constraints are: the optical and near-IR field luminosity functions (LFs) at $z = 0$; the redshift $z_{\rm re,half}$, at which the Universe was half reionized; the Milky Way (MW) satellite LF; and the stellar metallicity vs. stellar mass ($Z_* - M_*$) relation for MW satellites. We use the GALFORM semi-analytical model of galaxy formation embedded in the ΛCDM model of structure formation, with 4 different formulations for the mass-loading factor, β, of galactic outflows driven by SN feedback: (a) in the Fiducial model, β is a simple power law in galaxy circular velocity, $V_{\rm c}$; (b) in the Saturated feedback model, β is a broken power law in $V_{\rm c}$, with a flat slope at low $V_{\rm c}$; (c) in the power law Evolving feedback model, β is a single power law in $V_{\rm c}$, but with a normalization that is lower at higher redshifts; (d) in the Evolving feedback model, β decreases at high redshift, as well as having a break to a shallower slope at low $V_{\rm c}$. The Fiducial model was previously tuned by Lacey et al. (2016) to fit a wide range of observational constraints, but not including reionization or the MW satellites. Our main conclusions are:

(i) The single power law formulation of β as used in the Fiducial model can reproduce the faint ends of the $z = 0$ field LFs and MW satellite LF, but leads to too low $z_{\rm re,half}$ and too low MW satellite metallicities. This indicates that in this model, the SN feedback is too strong in small galaxies and/or at $z > 8$.

(ii) Simply reducing the SN feedback in small galaxies, as in the Saturated model, does not provide an improvement relative to the single power law formulation of β.

(iii) The power law Evolving SN feedback model, with weaker SN feedback at high redshifts and stronger SN feedback at low redshifts, can successfully reproduce the faint ends of the $z = 0$ field LFs, $z_{\rm re,half}$ and the MW satellite LF, but still predicts MW satellite metallicities that are too low, indicating the necessity of weakening the SN feedback in low $V_{\rm c}$ galaxies.

(iv) The Evolving SN feedback model, with the SN feedback strength decreasing with increasing redshift and a saturation for $V_{\rm c} \leq 50\,{\rm km\,s^{-1}}$, seems to be preferred by the above mentioned observational constraints. Including the effects of local reionization may further improve the predictions for the MW satellite LF.

(v) The physical reasons for the redshift evolution in our phenomenological Evolving SN feedback models could be that a single function of galaxy $V_{\rm c}$ only captures the effects of the gravitational potential well on the SN feedback, but the SN feedback is likely also to depend on factors such as the cold gas density and metallicity and the molecular gas fraction, which evolve with redshift. However, a more detailed ISM model is required to test the conclusions from this work further.

(vi) In all of the SN feedback models analysed in this work, around 50% of the photons which reionize the IGM are emitted by galaxies with stellar masses $M_* \gtrsim 10^9\,{\rm M_\odot}$, rest-frame far-UV absolute magnitudes, $M_{\rm AB}(1500{\rm \AA}) \lesssim -18$, galaxy circular velocities $V_{\rm c} \gtrsim 100\,{\rm km\,s^{-1}}$ and halo masses $M_{\rm halo} \gtrsim 10^{11}\,{\rm M_\odot}$ at the redshift $z \sim z_{\rm re,full}$ at which the Universe is fully reionized. In addition, most of the ionizing photons are predicted to be emitted by galaxies undergoing starbursts, rather than forming stars quiescently. This implies that the currently observed high redshift galaxy population should contribute about half of the ionizing photons that reionized Universe.

(vii) For our best fit model, namely the Evolving feedback model, the $z = 0$ descendants of the major ionizing photon sources are relatively large galaxies with $M_* \gtrsim 10^{10}\,{\rm M_\odot}$, and are mainly in dark matter halos with $M_{\rm halo} \gtrsim 10^{13}\,{\rm M_\odot}$. However, for these galaxies, the fraction of stars formed before reionization is low, while this fraction is high for dwarf galaxies with $z = 0$ stellar masses $M_* < 10^6\,{\rm M_\odot}$, even though the progenitors of such dwarfs contribute little to reionizing the Universe. This fraction also shows considerable scatter for the dwarfs, indicating that the star formation histories of these dwarf galaxies are very diverse.

(viii) For satellite galaxies in Milky Way-like halos, our best fit model implies that the fraction of stars formed before reionization is very low for large satellites like the LMC, SMC and Fornax, but reaches very high values for very small satellites with stellar masses $M_* < 10^6\,\mathrm{M}_\odot$, like Leo II, Ursa Minor and Hercules, with median fractions around 80%, indicating that typically these small satellites formed most of their stars before reionization.

Chapter 4

A new gas cooling model for semi-analytical galaxy formation models

4.1 Introduction

In this chapter we focus on the modelling of gas cooling and accretion in halos in semi-analytical models. In hierarchical structure formation models, dark matter halos grow in mass through both accretion and mergers. Baryons in the form of gas are accreted into halos along with the dark matter. However, only some fraction of this gas is accreted onto the central galaxy in the halo, this being determined by the combined effects of gravity, pressure, shock heating and radiative cooling. This whole process of gas accretion onto galaxies in halos is what we mean by "halo gas cooling". This is a crucial process in galaxy formation, for, along with galaxy mergers, it determines the amount of mass and angular momentum delivered to a galaxy, and thus is a primary determinant of the properties and evolution of galaxies.

Currently, most semi-analytical models use treatments of halo gas cooling that are more or less based on the gas cooling picture set out in White & Frenk (1991) [also see Binney (1977); Rees & Ostriker (1977); Silk (1977), White & Rees (1978), Bertschinger (1989) and Abadi et al. (2000)], in which the gas in a dark matter halo initially settles in a spherical pressure-supported hot gas halo, and this gas gradually

cools down and contracts under gravity as it loses pressure support, while new gas joins the halo due to structure growth or to the reincorporation of the gas ejected by feedback from supernovae (SN) and AGN.

The above picture is challenged by the so-called cold accretion scenario (e.g. Birnboim & Dekel, 2003; Kereš et al., 2005), in which the accreted gas in low mass halos ($M_{\rm halo} \lesssim 3 \times 10^{11}\,{\rm M}_{\odot}$) does not build a hot gaseous halo, but rather stays cold and falls freely onto the central galaxy. However, in these small halos, the cooling time scale of the assumed hot gas halo in SA models is very short, and the gas accretion onto central galaxies is in practice limited by the free-fall time scale, both in the original White & Frenk (1991) model and in most current SA models. Therefore the use of the the White & Frenk cooling picture for these halos should not introduce large errors in the accreted gas masses (Benson & Bower, 2011). In the cold accretion picture, cold gas flows through the halo along filaments (Kereš et al., 2005), and it has been argued that even in more massive halos some gas from the filaments can penetrate the hot gas halo and deliver cold gas directly to the central galaxy (e.g. Kereš et al., 2009), or to a shock close to the central galaxy (e.g. Nelson et al., 2016). However, this only happens when the temperature of the hot gas halo is not very high and the filaments still narrow, and so only in a limited range of redshift and halo mass (e.g. Kereš et al., 2009). Furthermore, the effects of accretion along filaments within halos are expected to be reduced when the effects of gas heating by SN and AGN are included (e.g. Benson & Bower, 2011). Therefore the cooling picture of White & Frenk (1991) should remain a reasonable approximation for the cold gas accretion rate.

There are three main gas cooling models used in SA models, namely those used in the Durham model GALFORM (e.g. Cole et al., 2000; Baugh et al., 2005; Bower et al., 2006; Lacey et al., 2016), in the Munich model L-GALAXIES (Springel et al., 2001; Croton et al., 2006; De Lucia & Blaizot, 2007; Guo et al., 2011; Henriques et al., 2015) and in the MORGANA model (Monaco et al., 2007; Viola et al., 2008). Most other SA models (e.g. Somerville et al., 2008) use a variant of one of these. We outline the key differences between the three cooling models here, and give more details in §4.2.2.

The GALFORM cooling model calculates the evolution of a cooling front, integrating outwards from the centre. However, it introduces artificial 'halo formation' events, when the halo gas density profile is reset, and the radius of the cooling front is reset to zero. Between these formation events, there is no contraction in the profile of the gas that is yet to cool. An improved version of this model, in which the artificial halo formation events are removed, was introduced in Benson & Bower (2010), but the treatment of the cooling history and contraction of the hot gas halo is still fairly approximate.

The L-GALAXIES cooling model is simpler to calculate than that in GALFORM. However, it does not seem physically very self-consistent, as it assumes that the hot gas is always distributed in the same profile between the halo centre and virial radius, and that the gas at smaller radii cools faster, but only the gas near the cooling radius contributes to the mass cooling rate, rather than all of the gas within this radius.

The MORGANA cooling model incorporates a more detailed calculation of the contraction of the hot gas halo due to cooling compared to the above models, but instead of letting the gas at small radius cool first, it assumes that hot gas at different radii contributes to the mass cooling rate simultaneously. This is in tension with the fact that the gas at smaller radius has a shorter cooling time scale, and so should contribute to cooling earlier.

Furthermore, while the GALFORM cooling model accounts for an angular momentum profile in the halo gas when calculating the angular momentum of the cooled down gas, the L-GALAXIES and MORGANA models are much more simplified in this respect.

In summary, all of the main cooling models used in current semi-analytical models have important limitations. In this chapter, we introduce a new cooling model for semi-analytical models. This new model treats the evolution of the hot gas density profile and of the gas cooling more self-constently compared to the models mentioned above, while also incorporating a detailed treatment of the angular momentum of the cooled down gas. This new cooling model is still based on the cooling picture in White & Frenk (1991). In particular, it still assumes a spherical hot gas halo. As

argued above, this picture may be a good approximation, but it needs to be further checked by comparing with hydrodynamical simulations in which shock heating and filamentary accretion are considered in detail. We leave this comparison for a future work. Note that even if accretion of cold gas along filaments within halos is significant, this does not exclude the existence of a diffuse, roughly spherical hot gas halo, and our new model should provide a better modeling of this component than the previous models mentioned above, and thus will still constitute a step towards an even more accurate and complete model of halo gas cooling.

This chapter is organized as follows. §4.2 first describes our new cooling model, and then describes the other main cooling models used in semi-analytical modelling. Then §4.3 compares predictions from the new cooling model with those from other models, first in static halos and then in hierarchically growing halos. The effects of the new cooling model on a full galaxy formation model are also shown and briefly discussed in this section. Finally a summary is given in §4.4.

4.2 Models

4.2.1 The new cooling model

4.2.1.1 Overview of the new cooling model

The hot gas inside a dark matter halo is assumed to form a spherical pressure-supported hot gas halo in hydrostatic equilibrium. The gas accreted during halo growth and also the reincorporated gas that was previously ejected by SN feedback are shock heated and join this hot gas halo. The hot gas halo itself can cool down due to radiation, and this cooling removes gas from this gas halo. The cooled down gas, which lacks pressure support, falls into the central region of the dark matter halo and delivers mass and angular momentum to the central galaxy. We call this component of cold infalling gas the cold gas halo. Typically the gas at smaller radii cools faster, and this kind of cooling leads to the reduction of pressure support from center outwards. The hot gas halo then contracts under gravity.

The boundary between the cold gas halo and hot gas halo is the so-called cooling

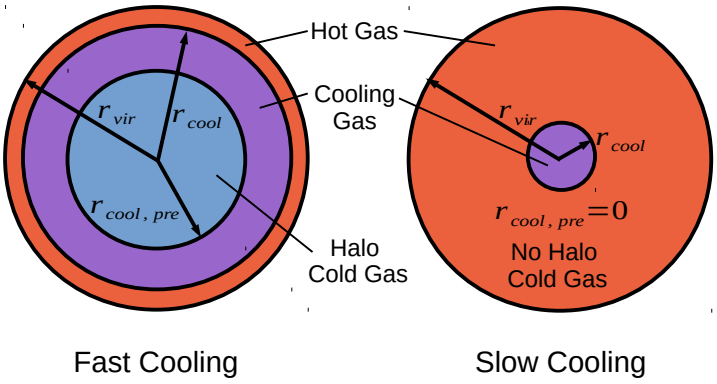

Figure 4.1: Sketch of the new cooling model.

radius $r_{\rm cool}$, at which the gas just has enough time to cool down. When discrete time steps are used, we introduce another quantity, $r_{\rm cool,pre}$, which is the boundary at the beginning of a time step. The hot gaseous halo is treated as fixed during a time step, $r_{\rm cool}$ is calculated based on this fixed halo, and the gas between $r_{\rm cool,pre}$ and $r_{\rm cool}$ cools down in this time step, and is called the cooling gas. Note that $r_{\rm cool,pre}$ is identical to $r_{\rm cool}$ calculated in the previous time step only if there is no contraction of the hot gas halo. This picture is sketched in Fig. 4.1.

The above scheme is similar to that in White & Frenk (1991) and to those in many other semi-analytical models, but most of these other models (apart from MORGANA) do not explicitly introduce the cold gas halo component or the contraction of the hot gas halo. Unlike the MORGANA model, in which the whole hot gas halo contributes to the cooled down gas in any timestep, here the hot gas cools gradually from halo center outwards. A more detailed discussion of the relation of the new cooling model to those in other semi-analytical models is given in §4.2.2.

4.2.1.2 Basic assumptions of the new cooling model

Based on the above picture, we impose our basic assumptions about the cooling as follows:

1. The hot gas in a dark matter halo is in a spherical hot gas halo, with a density

distribution described by the so-called β-distribution:

$$\rho_{\rm hot}(r) \propto \frac{1}{r^2 + r_{\rm core}^2}, \; r_{\rm cool,pre} \leq r \leq r_{\rm vir} \tag{4.2.1}$$

where $r_{\rm core}$ is called the core radius and is a parameter of this density distribution, while $r_{\rm vir}$ is the virial radius of the dark matter halo, defined as

$$r_{\rm vir} = \left(\frac{3M_{\rm halo}}{4\pi\Delta_{\rm vir}\overline{\rho}} \right)^{1/3}, \tag{4.2.2}$$

where $\overline{\rho}$ is the mean density of the universe at that redshift, and the overdensity $\Delta_{\rm vir}(\Omega_{\rm m}, \Omega_{\rm v})$ is calculated from the spherical collapse model (e.g. Eke et al., 1996). In GALFORM, typically $r_{\rm core}$ is set to be a fixed fraction of $r_{\rm vir}$ or of the NFW scale radius $r_{\rm NFW}$ (Navarro et al., 1997).

2. The hot gas has only one temperature at any time, and it is set to be the dark matter halo virial temperature $T_{\rm vir}$, where

$$T_{\rm vir} = \frac{\mu_{\rm m} v_{\rm vir}^2}{2k_{\rm B}}, \tag{4.2.3}$$

where $k_{\rm B}$ is the Boltzmann constant, $\mu_{\rm m}$ is the mean mass per particle, and $v_{\rm vir} = (GM_{\rm halo}/r_{\rm vir})^{1/2}$ is the circular velocity at $r_{\rm vir}$.

3. When new gas is added to the hot gas halo, it is assumed to mix with the existing hot gas halo homogeneously. This also means that the hot gas halo only has a single metallicity $Z_{\rm hot}$ at any given time.

4. In the absence of cooling, the specific angular momentum distribution of the hot gas $j_{\rm hot}(r) \propto r$, corresponding to a mean rotation velocity in spherical shells that is constant with radius. This applies to the initial moment when no cooling has happened and also to the gas newly added to the hot gas halo, which is newly heated up. When cooling induces contraction of the hot gas halo, the angular momentum of each Lagrangian hot gas shell is conserved during the contraction.

Our choices of $\rho_{\rm hot}(r)$ and of the initial $j_{\rm hot}(r)$ follow those of Cole et al. (2000), which are based on hydrodynamical simulations without cooling. This is reasonable, because here they only apply to the hot gas.

4.2.1.3 Cooling calculation

We describe the calculation for a single timestep, starting at time t and ending at time $t + \Delta t$. The timestep Δt should generally be chosen to be small compared to the halo dynamical timescale, so that the evolution in the halo mass and the contraction in the hot gas halo over a timestep are small. At the beginning of each step, $M_{\rm halo}$ is updated according to the halo merger tree, and , $r_{\rm vir}$ and $T_{\rm vir}$ are then updated according to the current values of $\Delta_{\rm vir}$ and $\overline{\rho}$. Next, the hot gas density profile $\rho_{\rm hot}(r,t)$ is updated, which involves two quantities, namely $r_{\rm core}$ and the density normalization. As mentioned above, $r_{\rm core}$ is calculated from the halo radius $r_{\rm vir}$ or $r_{\rm NFW}$. The normalization is fixed by the integral

$$4\pi \int_{r_{\rm cool,pre}(t)}^{r_{\rm vir}(t)} \rho_{\rm hot}(r,t)\, r^2 dr = M_{\rm hot}(t), \tag{4.2.4}$$

where $M_{\rm hot}$ is the total hot gas mass, and $r_{\rm cool,pre}$ the inner boundary of the hot gas halo at time t. Initially $r_{\rm cool,pre} = 0$ and is updated (see below) in each time step for the calculation of the next time step. For a static halo, $r_{\rm cool,pre}(t) = r_{\rm cool}(t)$, but this no longer applies if the halo grows or the hot gas distribution contracts.

With the density profile determined, the cooling radius $r_{\rm cool}(t + \Delta t)$ at the end of the timestep can be calculated. $r_{\rm cool}$ is defined by

$$t_{\rm cool}(r_{\rm cool}, t + \Delta t) = t_{\rm cool,avail}(r_{\rm cool}, t + \Delta t), \tag{4.2.5}$$

where $t_{\rm cool}(r,t)$ is the cooling timescale of a shell at radius r at time t, and $t_{\rm cool,avail}(r,t)$ is the time available for cooling for that shell. $t_{\rm cool}(r,t)$ is defined as

$$t_{\rm cool} = \frac{\delta U}{\delta L_{\rm cool}} = \frac{3k_{\rm B}}{2\mu_{\rm m}} \frac{T_{\rm vir}}{\tilde{\Lambda}(T_{\rm vir}, Z)\rho_{\rm hot}}, \tag{4.2.6}$$

where δU is the total thermal energy of this shell, while $\delta L_{\rm cool}$ is its current cooling luminosity. For gas with temperature $T_{\rm vir}$ and metallicity $Z_{\rm hot}$, we express the thermal energy density as $(3/2)(\rho_{\rm hot}/\mu_{\rm m})k_{\rm B}T_{\rm vir}$, and the radiative cooling rate per unit volume as $\tilde{\Lambda}(T_{\rm vir}, Z_{\rm hot})\rho_{\rm hot}^2$, assuming collisional ionization equilibrium. This then leads to the final expression on the RHS above.

The calculation of the time available for cooling, $t_{\rm cool,avail}(r,t)$, is more complicated. For a halo in which the hot gas density distribution, temperature and

metallicity are static, and in which the gas started cooling at some time $t_{\rm form}$, we would define $t_{\rm cool,avail} = t - t_{\rm form}$, as in Cole et al. (2000). However, this definition is not applicable to an evolving halo. Instead, we would like to define a gas shell as having cooled when $\delta U = \delta E_{\rm cool}$, where δU is defined as above, and $\delta E_{\rm cool}$ is the total energy radiated away by this hot gas shell over its past history when we track the shell in a Lagrangian sense. When we calculate δU and $\delta E_{\rm cool}$ for a gas shell, we include the effects of evolution in $\rho_{\rm hot}$, $T_{\rm vir}$ and $Z_{\rm hot}$ due to halo growth, reaccretion of ejected gas and contraction of the hot gas. However, in our approach, $\rho_{\rm hot}$ and T in a gas shell are assumed to be unaffected by radiative cooling within that shell, up until the the time when the cooling condition is met, when the hot gas shell is assumed to lose all of its thermal energy in a single instant, and be converted to cold gas. Combining the condition $\delta U = \delta E_{\rm cool}$ with Eq(4.2.6) then leads to a cooling condition of the form $t_{\rm cool}(r,t) = t_{\rm cool,avail}(r,t)$ if $t_{\rm cool,avail}$ for a shell is defined as

$$t_{\rm cool,avail} = \frac{\delta E_{\rm cool}}{\delta L_{\rm cool}} \tag{4.2.7}$$

This is just the time that it would take for the gas shell to radiate the energy actually radiated over its past history, if it were radiating at its current rate. Note that for a static halo cooling since time $t_{\rm form}$, $L_{\rm cool}$ is constant over the past history of a hot gas shell, so $\delta E_{\rm cool} = \delta L_{\rm cool}\,(t - t_{\rm form})$, and the above definition reduces to $t_{\rm cool,avail} = t - t_{\rm form}$.

The quantity $t_{\rm cool}$ is easy to calculate for each hot gas shell because it only involves quantities at time t. In contrast, the calculation of $t_{\rm cool,avail}$ is more difficult, because $\delta E_{\rm cool}$ involves the previous cooling history. To calculate $t_{\rm cool,avail}$ exactly, the cooling history of each Lagrangian hot gas shell would have to be stored, however, this is too computationally expensive for a semi-analytical model, and some further approximations are needed. We first note that for a discrete time step with length Δt and starting at t,

$$t_{\rm cool,avail}(r_{\rm cool}, t+\Delta t) = t_{\rm cool,avail}(r_{\rm cool}, t) + \Delta t \tag{4.2.8}$$

$$\approx t_{\rm cool,avail}(r_{\rm cool,pre}, t) + \Delta t. \tag{4.2.9}$$

The first line above comes from the assumption that the hot gas halo is fixed within a given time step, and thus the increase of $t_{\rm cool,avail}$ over the step is just

the increase of the physical time. To justify the above approximation in the second line, we consider two cases: (a) $r_{\rm cool} \sim r_{\rm cool,pre}$. In this case, which typically happens when the gas cools slowly compared to the halo dynamical timescale, $t_{\rm cool,avail}(r_{\rm cool}, t) \approx t_{\rm cool,avail}(r_{\rm cool,pre}, t)$. (b) $r_{\rm cool} \gg r_{\rm cool,pre}$. This typically happens when the gas cools fast compared to the halo dynamical timescale, but in that case, halo growth and hot gas halo contraction play only a weak role in cooling, which means that $t_{\rm cool,avail}$ is nearly the same for all gas shells (as in a completely static halo), so again $t_{\rm cool,avail}(r_{\rm cool}, t) \approx t_{\rm cool,avail}(r_{\rm cool,pre}, t)$.

Finally, we make the approximation

$$t_{\rm cool,avail}(r_{\rm cool,pre}, t) = \frac{\delta E_{\rm cool}(r_{\rm cool,pre}, t)}{\delta L_{\rm cool}(r_{\rm cool,pre}, t)} \approx \frac{E_{\rm cool}(t)}{L_{\rm cool}(t)} \tag{4.2.10}$$

Here, $L_{\rm cool}$ is the cooling luminosity of the whole hot gas halo at time t,

$$L_{\rm cool}(t) = 4\pi \int_{r_{\rm cool,pre}}^{r_{\rm vir}} \tilde{\Lambda}(T_{\rm vir}, Z_{\rm hot})\, \rho_{\rm hot}^2(r, t)\, r^2\, dr, \tag{4.2.11}$$

and $E_{\rm cool}(t)$ is the total energy radiated away over its past history by all of the hot gas that is within the halo at time t,

$$E_{\rm cool}(t) = 4\pi \int_{t_{\rm init}}^{t} \int_{r_{\rm p}(\tau)}^{r_{\rm vir}(\tau)} \tilde{\Lambda}(T_{\rm vir}, Z_{\rm hot})\, \rho_{\rm hot}^2(r, \tau)\, r^2 dr d\tau. \tag{4.2.12}$$

In the above integral, $t_{\rm init}$ is the starting time for the cooling calculation, and $r_{\rm p}(\tau)$ is the radius at time τ of the shell that has radius $r_{\rm cool,pre}$ at time t.

To justify the approximation made in Eq(4.2.10), we first note that, due to the integrals in Eqs.(4.2.11) and (4.2.12) involving $\rho_{\rm hot}^2$, both are dominated by the densest regions in the hot gas halo. We now need to consider two cases. (a) $r_{\rm cool,pre} \gtrsim r_{\rm core}$. In this case, the gas density decreases monotonically for $r > r_{\rm cool,pre}$, so that both integrals are dominated by the contributions from the gas shells near the lower limit of the integral, i.e. near $r_{\rm cool,pre}$. It follows that $E_{\rm cool}(t)/L_{\rm cool}(t) \approx \delta E_{\rm cool}(r_{\rm cool,pre}, t)/\delta L_{\rm cool}(r_{\rm cool,pre}, t)$. (b) $r_{\rm cool,pre} \lesssim r_{\rm core}$. In this case, $\delta E_{\rm cool}(r, t)/\delta L_{\rm cool}(r, t)$ is approximately independent of radius for $r \lesssim r_{\rm core}$ due to the approximately constant density, while the integrals for $E_{\rm cool}(t)$ and $L_{\rm cool}(t)$ are dominated by the region $r \lesssim r_{\rm core}$, so that we again have $E_{\rm cool}(t)/L_{\rm cool}(t) \approx \delta E_{\rm cool}(r_{\rm cool,pre}, t)/\delta L_{\rm cool}(r_{\rm cool,pre}, t)$.

By combining Eqs.(4.2.9) and (4.2.10), we obtain the expression for $t_{\rm cool,avail}$ that we actually use:

$$t_{\rm cool,avail}(r_{\rm cool}, t+\Delta t) = \frac{E_{\rm cool}(t)}{L_{\rm cool}(t)} + \Delta t, \tag{4.2.13}$$

In the above, the term $E_{\rm cool}(t)/L_{\rm cool}(t)$ represents the available time at the start of the step, calculated from the previous cooling history.

The calculation of $E_{\rm cool}$ from Eq(4.2.12) appears to require storing the histories of all of the shells of hot gas in order to evaluate the integral. However, from its definition, it is easy to derive an approximate recursive equation for it (see Appendix A)

$$\begin{aligned} E_{\rm cool}(t+\Delta t) &\approx E_{\rm cool}(t) + L_{\rm cool}(t) \times \Delta t \\ &- L'_{\rm cool}(t) \times t_{\rm cool,avail}(r_{\rm cool}, t+\Delta t), \end{aligned} \tag{4.2.14}$$

where

$$L'_{\rm cool}(t) = 4\pi \int_{r_{\rm cool,pre}}^{r_{\rm cool}} \tilde{\Lambda}(T_{\rm vir}, Z_{\rm hot})\, \rho_{\rm hot}^2\, r^2\, dr. \tag{4.2.15}$$

The second term in Eq(4.2.14) adds the energy radiated away in the current time step, while the third term removes the contribution from gas between $r_{\rm cool,pre}$ and $r_{\rm cool}$, because it cools down in the current time step and therefore is not a part of the hot gas halo at the next time step. Starting from the initial value $E_{\rm cool} = 0$, Eq(4.2.14) can be used to derive $E_{\rm cool}$ for the subsequent time steps, and then Eq(4.2.5), Eq(4.2.6) and Eq(4.2.13) can be used to calculate $r_{\rm cool}$. For a static halo, in which there is no accretion and no contraction of the hot gas, it can be shown that Eqs.(4.2.13)-(4.2.15) lead to $t_{\rm avail}(t+\Delta t) = t + \Delta t - t_{\rm init}$, the same as in Cole et al. (2000). Note that the calculation of $t_{\rm cool,avail}$ here is similar to that in the GFC2 cooling model introduced in §4.2.2.2, but is more accurate than the latter, because the calculations of $E_{\rm cool}$ and $L_{\rm cool}$ are more accurate here.

With $r_{\rm cool,pre}$ and $r_{\rm cool}$ determined, the mass and angular momentum of the gas cooled down over the time interval $(t, t+\Delta t)$ are calculated from

$$\Delta M_{\rm cool} = 4\pi \int_{r_{\rm cool,pre}}^{r_{\rm cool}} \rho_{\rm hot} r^2 dr \tag{4.2.16}$$

$$\Delta J_{\rm cool} = 4\pi \int_{r_{\rm cool,pre}}^{r_{\rm cool}} j_{\rm hot} \rho_{\rm hot} r^2 dr, \tag{4.2.17}$$

where $j_{\rm hot}(r)$ is the specific angular momentum distribution of the hot gas, which is calculated as described in §4.2.1.4. $\Delta M_{\rm cool}$ and $\Delta J_{\rm cool}$ are used to update the mass, $M_{\rm halo,cold}$, and angular momentum, $J_{\rm halo,cold}$, of the cold halo gas component.

Gas in the cold halo gas component is not pressure supported, and so is assumed to fall to the central galaxy in the halo on the freefall timescale. We therefore calculate the mass, $\Delta M_{\rm acc,gal}$, and angular momentum, $\Delta J_{\rm acc,gal}$, accreted onto the central galaxy over a timestep as

$$\Delta M_{\rm acc,gal} = M_{\rm halo,cold} \times \min[1, \Delta t/t_{\rm ff}(r_{\rm cool})] \quad (4.2.18)$$

$$\Delta J_{\rm acc,gal} = J_{\rm halo,cold} \times \min[1, \Delta t/t_{\rm ff}(r_{\rm cool})] \quad (4.2.19)$$

where $t_{\rm ff}(r_{\rm cool})$ is the free-fall time scale at cooling radius. Note that in the slow cooling regime, where $t_{\rm ff}(r_{\rm cool}) < t_{\rm cool}(r_{\rm cool})$, the mass of the cold halo gas component remains relatively small, since the timescale for draining it ($t_{\rm ff}$) is small compared to the timescale for feeding it ($t_{\rm cool}$).

Finally, we consider the contraction of the hot gas halo. The gas between the cooling radius and the virial radius is assumed to remain in approximate hydrostatic equilibrium, so for simplicity we assume that it always follows the β-profile. The hot gas at the cooling radius is not pressure-supported by the cold gas at smaller radii, so we assume that this gas contracts towards the halo centre on a timescale $t_{\rm ff}(r_{\rm cool})$. The new $r_{\rm cool,pre}$ at the next time step starting at $t + \Delta t$ is therefore estimated as

$$r_{\rm cool,pre}(t + \Delta t) = r_{\rm cool}(t + \Delta t) \times \max[0, 1 - \Delta t/t_{\rm ff}(r_{\rm cool})]. \quad (4.2.20)$$

The above equation only applies if the gravitational potential of the halo is fixed. When the halo grows in mass, and when the mean halo density within $r_{\rm vir}$ adjusts with the mean density of the universe, the gravitational potential also changes, and this affects the contraction of the hot gas halo. We estimate the effect of this on the inner boundary of the hot halo gas by requiring that the mass of dark matter contained inside $r_{\rm cool,pre}$ remains the same before and after the change in the halo potential, i.e.

$$M'_{\rm halo}[r'_{\rm cool,pre}(t + \Delta t)] = M_{\rm halo}[r_{\rm cool,pre}(t + \Delta t)], \quad (4.2.21)$$

where the quantities with apostrophes are after halo growth, while those without apostrophes are before halo growth. The reason for using the dark matter to trace this contraction is that the gas within $r_{\rm cool,pre}$ is cold with negligible pressure effects, so its dynamics should be similar to those of the collisionless dark matter. Note that the MORGANA cooling model introduced in §4.2.2.4 also includes a contraction of the halo hot gas component, but there the effects of dark matter growth on this contraction is not considered.

4.2.1.4 Calculating $j_{\rm hot}(r)$

The specific angular momentum of the hot gas averaged over spherical shells is assumed to follow $j_{\rm hot}(r) \propto r$ at the initial time, as stated in §4.2.1.2, with the normalization set by assumption that the mean specific angular momentum of the hot gas in the whole halo, $\bar{j}_{\rm hot}$, is initially equal to that of the dark matter, $J_{\rm halo}/M_{\rm halo}$ (see §4.2.3). Later on, the dark matter halo growth, the contraction of the hot gas halo and the addition of new gas all can change the angular momentum profile. In the new cooling model, at the beginning of each time step, we first consider the angular momentum profile change of the existing hot gas due to the dark matter halo growth and the hot gas halo contraction happened during the last time step, and then add the contribution from the newly added hot gas to this adjusted profile.

In deriving the change of angular momentum profile of the existing hot gas, we assume mass and angular momentum conservation for each Lagrangian shell. Consider a shell with mass dm, original radius r and specific angular momentum $j_{\rm hot}(r)$, which, after the dark matter growth and hot gas halo contraction, moves to radius r' with specific angular momentum $j'_{\rm hot}(r')$. The shell mass is unchanged because of mass conservation. Then angular momentum conservation implies $j'_{\rm hot}(r') = j_{\rm hot}(r)$. In other words, the angular momentum profile after these changes is $j_{\rm hot}[r(r')]$. Given $j_{\rm hot}(r)$ from the last time step, the major task for deriving $j'_{\rm hot}(r')$ is to derive $r(r')$. This can be done by considering shell mass conservation and the density profiles of the hot gas. Specifically, assuming $\rho_{\rm hot}(r)$ and $\rho'_{\rm hot}(r')$ are respectively the density profiles of the existing hot gas before and after the dark matter halo growth

and hot gas halo contraction, then one has

$$4\pi\rho_{\rm hot}(r)r^2dr = dm = 4\pi\rho'_{\rm hot}(r')r'^2dr'. \tag{4.2.22}$$

This together with the assumption that $\rho_{\rm hot}(r)$ and $\rho'_{\rm hot}(r')$ follow the β-distribution can then be solved for $r(r')$. Unfortunately, this equation can only provide an implicit form for $r'(r)$, and does not lead to an explicit analytical expression for $j'_{\rm hot}(r')$. A straightforward way to deal with this is to evaluate $j'_{\rm hot}(r')$ numerically for a grid of radii and then store this information, however, this is computationally expensive. Instead, we apply further approximations to reduce the computational cost of solving for $j'_{\rm hot}(r')$, as is described in detail in Appendix B.

To derive the final angular momentum distribution, $j''_{\rm hot}(r')$, one still needs to consider the contribution from the newly added hot gas. Assuming the gas newly added to a given shell with radius r' has mass $dm_{\rm new}$ and specific angular momentum $j_{\rm new}(r')$, then one has

$$j''_{\rm hot}(r')(dm + dm_{\rm new}) = j'_{\rm hot}(r')dm + j_{\rm new}(r')dm_{\rm new}. \tag{4.2.23}$$

Because the newly added gas is assumed to be mixed homogeneously with the hot gas halo, so all $dm_{\rm new}/(dm+dm_{\rm new})$ should be the same for all shells, and hence

$$\frac{dm_{\rm new}}{dm_{\rm new} + dm} = \frac{M_{\rm new}}{M_{\rm new} + M_{\rm old}}, \tag{4.2.24}$$

where $M_{\rm new}$ is the total mass added to the hot gas halo during the timestep, while $M_{\rm old}$ is the previous mass.

Further, according to the assumption in §4.2.1.2, $j_{\rm new}(r') \propto r'$. In general, there are two components to the newly added hot gas: (a) gas brought in through growth of the dark matter halo; and (b) gas that has been ejected from the galaxy by SN feedback, has joined the ejected gas reservoir, and then been reaccreted into the hot gas halo. Their contributions to the total angular momentum of the newly added gas are described in §4.2.1.5. With this, the normalization of $j_{\rm new}(r')$ can be determined.

Finally, with $j'_{\rm hot}(r')$ and $j_{\rm new}(r')$ known, the specific angular momentum distribution of the current time step, $j''_{\rm hot}(r')$, is determined as

$$j''_{\rm hot}(r') = j'_{\rm hot}(r')\frac{M_{\rm old}}{M_{\rm new} + M_{\rm old}} + j_{\rm new}(r')\frac{M_{\rm new}}{M_{\rm new} + M_{\rm old}}. \tag{4.2.25}$$

In this way, the specific angular momentum distribution for any given time step can be derived recursively from the initial distribution. Previous cooling models introduced in §4.2.2 have either less detailed or less self-consistent calculations of angular momentum compared to the new cooling model.

4.2.1.5 Treatments of gas ejected by feedback and halo mergers

The SN feedback can heat and eject gas in galaxies, and the ejected gas is added to the so-called ejected gas reservoir. This transfers mass and angular momentum from galaxies to that reservoir. The gas ejected from both the central galaxy and its satellites is added to the ejected gas reservoir of the central galaxy. The ejected mass is determined by the SN feedback recipe, and is typically proportional to the instantaneous star formation rate. The angular momentum of this ejected gas is calculated as follows. The total angular momentum of the ejected gas can be expressed as the product of its mass and its specific angular momentum. For the gas ejected from the central galaxy, its specific angular momentum is estimated as that of the central galaxy, while for the gas ejected from satellites, this specific angular momentum is estimated as the mean specific angular momentum of the central galaxy's host dark matter halo, i.e. $J_{\rm halo}/M_{\rm halo}$, in order to roughly include the contribution to the ejected angular momentum from the satellite orbital motion.

This ejected gas can later be reaccreted onto the hot gas halo, thus delivering mass and angular momentum to it. The reaccretion rates of mass and angular momentum are respectively estimated as

$$\dot{M}_{\rm return} = \alpha_{\rm return} \times M_{\rm eject}/t_{\rm dyn}(r_{\rm vir}) \quad (4.2.26)$$

$$\dot{J}_{\rm return} = \alpha_{\rm return} \times J_{\rm eject}/t_{\rm dyn}(r_{\rm vir})], \quad (4.2.27)$$

where $\dot{M}_{\rm return}$ and $\dot{J}_{\rm return}$ are respectively the mass and angular momentum reaccretion rates, $M_{\rm eject}$ and $J_{\rm eject}$ are respectively the total mass and angular momentum of the ejected gas reservoir, $t_{\rm dyn}(r_{\rm vir}) = r_{\rm vir}/v_{\rm vir}$ is the halo dynamical timescale and $\alpha_{\rm return} \sim 1$ a free parameter. For a time step with finite length Δt, the mass, $\Delta M_{\rm return}$, and angular momentum, $\Delta J_{\rm return}$, reaccreted within it is then calculated as the products of the corresponding rates and Δt.

When a halo falls into a larger halo, it becomes a subhalo, while the larger one becomes the host halo of this subhalo. The new cooling model assumes that the hot gas and ejected gas reservoir associated with this subhalo are instantaneously transferred to the corresponding gas components of the host halo at infall. This is used to mimic the effects of ram-pressure and tidal stripping. The masses of these transferred components can be simply added to the corresponding components of the host halo. However, the angular momentum cannot be directly added, because it is calculated before the infall, when the subhalo was still an isolated halo, and the reference point for this angular momentum is the center of the subhalo, while after the transition, the reference point becomes the center of the host halo. Here the angular momentum transferred is estimated as follows. The total angular momentum transferred is expressed as a product of the total transfered mass and the specific angular momentum. The latter one is estimated as $j_{\rm new,halo} = \Delta J_{\rm halo}/\Delta M_{\rm halo}$, where $\Delta J_{\rm halo}$ and $\Delta M_{\rm halo}$ are the angular momentum and mass changes in dark matter halo during the halo merger, and they can be determined when the mass and spin, $\lambda_{\rm halo}$, of each halo in a merger tree are given (see §4.2.3). The reason for this estimation is that the dark matter and baryon matter accreted by the host halo have roughly the same motion, and thus should gain similar specific angular momentum though the torque of the surrounding large scale structures. Then the mass and angular momentum transferred during the halo merger can be summarized as

$$\Delta M_{\rm hot,mrg} = \sum_{i=1}^{N_{\rm mrg}} M_{\rm hot,i} \tag{4.2.28}$$

$$\Delta J_{\rm hot,mrg} = j_{\rm new,halo} \times \Delta M_{\rm hot,mrg} \tag{4.2.29}$$

$$\Delta M_{\rm eject,mrg} = \sum_{i=1}^{N_{\rm mrg}} M_{\rm eject,i} \tag{4.2.30}$$

$$\Delta J_{\rm eject,mrg} = j_{\rm new,halo} \times \Delta M_{\rm eject,mrg} \tag{4.2.31}$$

where $\Delta M_{\rm hot,mrg}$ and $\Delta J_{\rm hot,mrg}$ are respectively the total mass and angular momentum transferred to the hot gas halo of the host halo during the halo merger, while $\Delta M_{\rm eject,mrg}$ and $\Delta J_{\rm eject,mrg}$ are the mass and angular momentum transferred to the ejected gas reservoir, and $N_{\rm mrg}$ is the total number of infalling halos over the time step, $M_{\rm hot,i}$ is the total mass of the hot gas halo of the ith infalling halo, and $M_{\rm eject,i}$

is the mass of its ejected gas reservoir.

In this cooling model, by default, the halo cold gas is not transferred during halo mergers, because it is cold and in the relatively central region of the infalling halo, and thus less affected by ram-pressure and tidal stripping. After infall, this component can still deliver cold gas to the satellite for a while. There are also options in the code to transfer the halo cold gas to the hot gas halo or halo cold gas of the host halo. In this work, we always adopt the default setting.

A dark matter halo may also accrete smoothly. The accreted gas is assumed to be shock heated and join the hot gas halo. In each time step, the mass of this gas, $\Delta M_{\rm hot,smooth}$, is given as $\Delta M_{\rm hot,smooth} = [\Omega_{\rm b}/\Omega_{\rm m}]\Delta M_{\rm halo,smooth}$, with $\Delta M_{\rm halo,smooth}$ the mass of smoothly accreted dark matter, which is provided by the merger tree, while the associated angular momentum is estimated as $\Delta J_{\rm hot,smooth} = j_{\rm new,halo} \times \Delta M_{\rm hot,smooth}$.

In each time step, $\Delta M_{\rm return}$, $\Delta M_{\rm hot,mrg}$ and $\Delta M_{\rm hot,smooth}$ increase the mass of the hot gas halo, but do not increase $E_{\rm cool}$. This means the newly added gas has no previous cooling history, consistent with the assumption that this gas is newly heated up by shocks. The total angular momentum of this newly added gas is $\Delta J_{\rm return}+\Delta J_{\rm hot,mrg}+\Delta J_{\rm hot,smooth}$, and together with the assumption that $j_{\rm new}(r) \propto r$, it completely determines the specific angular momentum distribution of the newly added gas.

4.2.2 Previous cooling models

4.2.2.1 galform cooling model GFC1

This cooling model is used in all recent versions of GALFORM (e.g. Gonzalez-Perez et al., 2014; Lacey et al., 2016), and is based on the cooling model introduced in Cole et al. (2000), and modified in Bower et al. (2006). The Cole et al. (2000) cooling model introduces so-called halo formation events. These are defined such that the appearance of a halo with no progenitor in a merger tree is a halo formation event, and the time when a halo first becomes at least twice as more massive as at the last halo formation event is a new halo formation event. The Cole et al. then assumes

that the hot gas halo is fixed between two adjacent halo formation events, and is reset at each formation event. Under this assumption, $t_{\rm cool,avail}$ is always the time elapsed since the latest halo formation event, which is straightforward to calculate. With $t_{\rm cool,avail}$ given, $r_{\rm cool}$ can be then calculated from Eq(4.2.5), and the mass and angular momentum cooled down can be calculated, as described below. The assumption of a fixed hot gas halo between two halo formation events means that changes of $r_{\rm vir}$ and $T_{\rm vir}$ induced by halo growth, and by the addition of new hot gas either by halo growth or by the reincorporation of gas ejected by feedback between halo formation events, are not considered until the coming of a halo formation event. While this may be reasonable for halo formation events induced by halo major mergers, in which the hot gas halo properties change fairly abruptly, it is not physical if the halo formation event is triggered through smooth halo growth, in which case the changes in the hot gas halo should also happen smoothly, instead of happening at a sudden jump at the halo formation event.

The GFC1 model (Bower et al., 2006) improves on this, by updating some hot gas halo properties at each time step instead of only at halo formation events. Specifically, in this model, the hot gas is still assumed to settle in a density profile described by the β-distribution, with the temperature equal to the current halo virial temperature $T_{\rm vir}$ and $r_{\rm core}$ set to be a fixed fraction of the current $r_{\rm vir}$. The halo mass is updated at each timestep, and the total hot gas mass and metallicity include the contributions from the hot gas newly added at each timestep. However, $v_{\rm vir}$ and $T_{\rm vir}$ are fixed at the values calculated at the last halo formation event. Unlike in the new cooling model, the normalization of the density profile is determined by requiring that

$$4\pi \int_0^{r_{\rm vir}} \rho_{\rm hot} r^2 dr = M_{\rm hot} + M_{\rm cooled}, \tag{4.2.32}$$

where $M_{\rm hot}$ is the total mass of the hot gas, while $M_{\rm cooled}$ is the total mass of the gas that has cooled down from this halo since the last halo formation event, and is either in the central galaxy or ejected by SN feedback but not yet reaccreted by the hot gas halo. Accordingly, $M_{\rm cooled}$ is reset to 0 at each halo formation event, while the ejected gas reservoir mass, $M_{\rm eject}$, evolves smoothly and is not affected by halo formation events.

This is not very physical, because the cooled down gas might have collapsed onto the central galaxy long ago, while the ejected gas is outside the halo. This also means that the contraction of the hot gas halo due to cooling is largely ignored in the determination of its density profile. This point is most obvious in the case of a static halo, when the dark matter halo does not grow. In this case, if there is no feedback and subsequent reaccretion, then the amount of hot gas gradually reduces due to cooling, and the hot gas halo should gradually contract in response to the reduction of pressure support caused by this cooling. However, in the GFC1 model, in this situation, the hot gas profile remains fixed, because $M_{\rm hot} + M_{\rm cooled}$ always equals the initial total hot gas mass. For a dynamical halo, $M_{\rm cooled}$ is reset to zero at each halo formation event, and thus the hot gas contracts to halo center at these events. In this way, the halo contraction due to cooling is included to some extent.

In the GFC1 model, $r_{\rm cool}$ is calculated in the same way as in Cole et al. (2000). For the estimation of $t_{\rm cool,avail}$, the GFC1 model retains the artificial halo formation events. This means that in both the GFC1 and Cole et al. (2000) cooling models, the hot gas cooling history is effectively reset at each halo formation event. While this might be physical when the halo grows through major mergers[1], it is artificial when a halo grows smoothly, in which case the cooling history is expected to evolve smoothly as well. Moreover, in principle $t_{\rm cool,avail}$ should change when the hot gas halo changes, which happens between halo formation events in the GFC1 model, so estimating $t_{\rm cool,avail}$ in the GFC1 model in the same way as in Cole et al. (2000) is not very self-consistent.

Unlike new cooling model that explicitly introduces a cold halo gas component that drains onto the central galaxy on the free-fall timescale, the GFC1 and Cole et al. (2000) cooling models introduce a so-called free-fall radius $r_{\rm ff}$ to allow for the fact that gas cannot accrete onto the central galaxy more rapidly than on a free-fall timescale, no matter how rapidly it cools. $r_{\rm ff}$ is calculated as

$$t_{\rm ff}(r_{\rm ff}) = t_{\rm ff,avail}, \tag{4.2.33}$$

where $t_{\rm ff}(r)$ is the free-fall timescale at radius r, defined as the time for a particle to

[1] although, Monaco et al. (2014) suggests that halo major mergers do not strongly affect cooling

fall to $r = 0$ starting at rest at radius r, and $t_{\rm ff,avail}$ is the time available for free-fall, which is set to be the same as $t_{\rm cool,avail}$ in these two cooling models. Then the mass accreted onto the cental galaxy over a timestep is given by

$$\Delta M_{\rm acc,gal} = 4\pi \int_{r_{\rm infall,pre}}^{r_{\rm infall}} \rho_{\rm hot}\, r^2 dr, \tag{4.2.34}$$

where $\rho_{\rm hot}$ is the current halo gas density distribution, while $r_{\rm infall} = \min(r_{\rm cool}, r_{\rm ff})$, and $r_{\rm infall,pre}$ is determined by $4\pi \int_0^{r_{\rm infall,pre}} \rho_{\rm hot} r^2 dr = M_{\rm cooled}$.

The introduction of $r_{\rm ff}$ and $r_{\rm infall}$ leaves part of the cooled down gas in the nominal hot gas halo when $r_{\rm cool} > r_{\rm ff}$, which is the case in the fast cooling regime. This gas is treated as hot gas in subsequent time steps. While in the fast cooling regime this should not strongly affect the final results for the amount of gas that cools, due to the cooling and accretion being rapid, this misclassification of cold gas as hot is still an unwanted physical feature of a cooling model.

The calculation of the angular momentum of the gas accreted onto the central galaxy is the same in the cooling model in Cole et al. (2000) and GFC1 model. The angular momentum is calculated as

$$\Delta J_{\rm acc,gal} = 4\pi \int_{r_{\rm infall,pre}}^{r_{\rm infall}} j_{\rm hot} \rho_{\rm hot} r^2 dr, \tag{4.2.35}$$

where $j_{\rm hot}$ is the specific angular momentum distribution of the hot gas halo, which is assumed to vary as $j_{\rm hot} \propto r$. As mentioned in §4.2.1.2, this assumption is based on the hydrodynamical simulations without cooling. Assuming it to apply unchanged in the presence of cooling means that the effect of contraction of the hot gas halo due to cooling is ignored.

This model adopts the treatments for the gas ejected by feedback and for halo mergers similar to those of the new cooling model. Because this model assumes that $t_{\rm cool,avail}$ is always the physical time since the last halo formation event, here the gas newly added through halo growth and reaccretion of the feedback ejected gas would share this $t_{\rm cool,avail}$ and thus implicitly gain some previous cooling history. As a result, the newly added gas is effectively not actually newly heated up.

4.2.2.2 galform cooling model GFC2

This model was introduced in Benson & Bower (2010). It has several improvements compared to the GFC1 model. In this model, the assumptions about the density profile[2], temperature and metallicity of the hot gas halo are the same as in GFC1, but the influence of halo formation events is mostly removed. The density profile of the hot gas is normalized by requiring

$$4\pi \int_0^{r_{\rm vir}} \rho_{\rm hot} r^2 dr = M_{\rm hot} + M_{\rm cooled} + M_{\rm eject}, \tag{4.2.36}$$

where $M_{\rm eject}$ is the mass of gas ejected by SN feedback and not yet reaccreted, while the definition of $M_{\rm cooled}$ is modified: (a) It is incremented by the mass cooled and accreted onto the central galaxy, and decremented by the mass ejected by SN feedback. (b) The GFC2 model introduces a gradual reduction of $M_{\rm cooled}$ as

$$\dot{M}_{\rm cooled} = -\alpha_{\rm remove} \times M_{\rm cooled}/t_{\rm ff}(r_{\rm vir}), \tag{4.2.37}$$

with $\alpha_{\rm remove} \sim 1$ being a free parameter. (c) When a halo merger occurs, the value of $M_{\rm cooled}$ is propagated to the current halo from its most massive progenitor (rather than being reset to 0 at each halo formation event as in the GFC1 model). Since the density profile normalization for the hot gas is determined by Eq(4.2.36), for a given $M_{\rm hot}$ and $M_{\rm eject}$, the gradual reduction of $M_{\rm cooled}$ due to Eq(4.2.37) lowers the normalization, and so to include the same mass $M_{\rm hot}$ in the density profile, the hot gas must be distributed to smaller radii. This gradual reduction of $M_{\rm cooled}$ thus effectively leads to a contraction of the hot gas halo in response to the removal of hot gas by cooling, which is more physical than the treatment in the GFC1 model. However, here the timescale for this contraction is $t_{\rm ff}(r_{\rm vir})$, while the region where the contraction happens has a radius $\sim r_{\rm cool}$, so there is still a physical mismatch in this timescale. This is improved in the new cooling model introduced in §4.2.1, where the timescale $t_{\rm ff}(r_{\rm cool})$ is adopted instead.

[2]Note that Benson & Bower (2010) actually adopt a different density profile for the hot gas halo, however, here for a fair comparison with other GALFORM cooling models, the β-profile is adopted instead for this model.

In the GFC2 model, as in the new cooling model, $r_{\rm cool}$ calculated using Eq(4.2.5), with $t_{\rm cool,avail}$ being estimated from the energy radiated away. By doing this, the effect of artificial halo formation events on the gas cooling is largely removed. However, instead of directly accumulating this radiated energy as in the new cooling model, the GFC2 model further approximates the integrals involving $\rho_{\rm hot}^2$ in Eqs.(4.2.11) and (4.2.12) as

$$\begin{aligned} 4\pi \int_0^{r_{\rm vir}} \rho_{\rm hot}^2 r^2 dr &\approx \bar{\rho}_{\rm hot} \times 4\pi \int_0^{r_{\rm vir}} \rho_{\rm hot} r^2 dr \\ &= \bar{\rho}_{\rm hot}(M_{\rm hot} + M_{\rm cooled} + M_{\rm eject}), \end{aligned} \tag{4.2.38}$$

where $\bar{\rho}_{\rm hot}$ is the mean density given by the density profile. This approximation is very rough, and while in the new cooling model the integral is limited to the gas that is hot, i.e. between $r_{\rm cool,pre}$ and $r_{\rm vir}$, in the GFC2 model the integration range is extended to $r = 0$, which includes the part of the density profile where the gas has already cooled down. These approximations make the calculation of $t_{\rm cool,avail}$ less accurate and physical than in the new cooling model.

With these approximations, for any time t, the GFC2 model adopts the following equations in place of Eqs.(4.2.11) and (4.2.12) in the new cooling model:

$$L_{\rm cool}(t) = \tilde{\Lambda}(T_{\rm vir}, Z)\bar{\rho}_{\rm hot}(M_{\rm hot} + M_{\rm cooled} + M_{\rm eject}) \tag{4.2.39}$$

$$\begin{aligned} E_{\rm cool}(t) &= \int_{t_{\rm init}}^{t} \tilde{\Lambda}(T_{\rm vir}, Z)\bar{\rho}_{\rm hot} \times \\ &\quad [M_{\rm hot}(\tau) + M_{\rm cooled}(\tau) + M_{\rm eject}(\tau)]d\tau \\ &+ \int_{t_{\rm init}}^{t} \frac{3k_{\rm B}}{2\mu_{\rm m}} T_{\rm vir} \dot{M}_{\rm cooled} d\tau. \end{aligned} \tag{4.2.40}$$

The second term in Eq(4.2.40), which is negative, is equal in absolute value to the total thermal energy of the cooled mass removed according to Eq(4.2.37), and is designed to remove the contribution to $E_{\rm cool}$ from this cooled mass. Given $E_{\rm cool}$ and $L_{\rm cool}$, $t_{\rm cool,avail}$ for a given time step is calculated from Eq(4.2.13), as in the new model. Note that the approximation made in Eq(4.2.38) leads to the derived $t_{\rm cool,avail}$ being closer to the average cooling history of all shells instead of the cooling history of gas near $r_{\rm cool}$, and so leads to less accurate results than in the new cooling model.

The GFC2 model allows for the effect of the free-fall timescale on the gas mass accreted onto the central galaxy in a similar way to the GFC1 model, by introducing

the radius $r_{\rm ff}$ calculated from Eq(4.2.33), but with $t_{\rm ff,avail}$ calculated in a way similar to that of $t_{\rm cool,avail}$. Specifically, a total energy radiated away similar to $E_{\rm cool}$ is accumulated for $t_{\rm ff,avail}$, but this energy has a limit $t_{\rm ff}(r_{\rm vir})\times L_{\rm cool}$, and once it exceeds this limit, it is then reset to this limit value. This limit ensures $t_{\rm ff,avail} \leq t_{\rm ff}(r_{\rm vir})$. Note that the effect of imposing this limit is usually to lead to a $t_{\rm ff,avail}$ different from $t_{\rm cool,avail}$ and $t_{\rm ff}(r_{\rm vir})$. The calculation of $t_{\rm ff,avail}$ is not very physical, because the calculation of $t_{\rm cool,avail}$ in the GFC2 model is based approximately on the total energy released by the cooling radiation, while the accretion of the cooled gas onto the central galaxy is driven by gravity, which does not depend on the energy loss by radiation. In addition, by introducing $r_{\rm ff}$, the GFC2 model inherits the associated problems already identified for GFC1 model.

The GFC2 model also adopts a specific angular momentum distribution for the hot gas to calculate the angular momentum of the gas that cools down and accretes onto the central galaxy. The simpler method to specify this angular momentum distribution is as a function of radius, namely $j_{\rm hot}(r)$. But in principle this requires calculating the subsequent evolution of $j_{\rm hot}(r)$ as the hot gas halo contracts, which is considered in the new cooling model but not in the GFC1 or GFC2 models. A more complex method is to specify $j_{\rm hot}$ as a function of the gas mass enclosed by a given radius, i.e. $j_{\rm hot}(< M)$. This implicitly includes the effect of contraction of the hot gas halo in the case of a static halo, where no new gas joins the hot gas halo, because while the radius of each gas shell changes during contraction , the enclosed mass of it is kept constant and can be used to track each Lagrangian shell. However, when there is new gas being added to the hot gas halo, this method also fails, because the newly joining gas mixes with the hot gas halo after contraction, and in this case, the contraction has to be considered explicitly. Since even the more complex method is not fully self-consistent, for the sake of simplicity, in this work we adopt the simpler method to calculate the angular momentum, without allowing for contraction of the hot gas halo.

This model also adopts the treatments for the gas ejected by feedback and for halo mergers similar to those of the new cooling model, but unlike the latter, here $E_{\rm cool}$ of the hot gas in the infalling halos is also transferred. This again gives the

newly added gas some previous cooling history, so it is not newly heated up.

4.2.2.3 Cooling model in l-galaxies

The cooling model used in L-GALAXIES (see e.g. Croton et al. (2006); Guo et al. (2011); Henriques et al. (2015)) assumes that the hot gas is always distributed from $r = 0$ to $r = r_{\rm vir}$, and its density profile is singular isothermal, namely $\rho_{\rm hot}(r) \propto r^{-2}$, with a single metallicity and a single temperature equaling $T_{\rm vir}$. The total mass inside this profile is $M_{\rm hot}$.

Then a cooling radius $r_{\rm cool}$ is calculated from $t_{\rm cool}(r_{\rm cool}) = t_{\rm cool,avail}$, with $t_{\rm cool,avail} = t_{\rm dyn}(r_{\rm vir}) = r_{\rm vir}/V_{\rm vir}$. If $r_{\rm cool} \leq r_{\rm vir}$, then the mass accreted onto the central galaxy in a time step Δt is [3]

$$\begin{aligned} \Delta M_{\rm acc,gal} &= 4\pi\rho_{\rm hot}(r_{\rm cool}) \times r_{\rm cool}^2 \frac{dr_{\rm cool}}{dt}\Delta t \\ &= \frac{M_{\rm hot}}{r_{\rm vir}} \frac{r_{\rm cool}}{t_{\rm dyn}(r_{\rm vir})}\Delta t, \end{aligned} \tag{4.2.41}$$

with $dr_{\rm cool}/dt$ being estimated as $dr_{\rm cool}/dt = r_{\rm cool}/t_{\rm cool,avail} = r_{\rm cool}/t_{\rm dyn}(r_{\rm vir})$. If instead $r_{\rm cool} > r_{\rm vir}$, then

$$\Delta M_{\rm acc,gal} = \frac{M_{\rm hot}}{t_{\rm dyn}(r_{\rm vir})}\Delta t. \tag{4.2.42}$$

Mass accretion rates onto central galaxies calculated using the above expressions have been shown to be in good agreement with stripped-down SPH hydrodynamical simulations, in which cooling is included but other process such as star formation and feedback are ignored (Yoshida et al., 2002; Monaco et al., 2014). However, the physical reason for adopting $t_{\rm cool,avail} = t_{\rm dyn}(r_{\rm vir})$ is not very clear. Moreover, the physical picture underlying this calculation is not very self-consistent. The density profile assumed for the hot gas halo in this model implies that the cooling time scale $t_{\rm cool}$ increases monotonically with radius, and together with the definition of $r_{\rm cool}$, this means that at any time, all gas within $r_{\rm cool}$ should cool down. All gas

[3] Here we adopted the equation for $\Delta M_{\rm acc,gal}$ from recent versions of the L-GALAXIES model (e.g. Croton et al., 2006; Henriques et al., 2015). In earlier versions (e.g. Springel et al., 2001), an extra factor 0.5 is introduced in front of the second line of Eq(4.2.41). See the footnote in Guo et al. (2011) for more information.

at $r < r_{\rm cool}$ should therefore fall under gravity onto the central galaxy, while the cooling model only allows the gas near $r = r_{\rm cool}$ to be accreted onto the central galaxy, ignoring the contributions from other smaller radii. Gas at $r < r_{\rm cool}$ is left in the halo and continues to be treated as hot gas in the next time step. The above calculation of the mass accretion rate nonetheless provides a good fit to the stripped down SPH simulations, but because of the inconsistency in its physical formulation, this goodness of fit does not imply its validity in the full galaxy formation context.

The angular momentum of the cooled down gas that accretes onto the central galaxy is calculated as

$$\Delta J_{\rm acc,gal} = \Delta M_{\rm acc,gal} \times \bar{j}_{\rm halo}, \tag{4.2.43}$$

where $\bar{j}_{\rm halo} = J_{\rm halo}/M_{\rm halo}$ is the specific angular momentum of the entire dark matter halo, with $J_{\rm halo}$ and $M_{\rm halo}$ being the total angular momentum and mass of the dark matter halo respectively. This corresponds to a specific angular momentum distribution for the hot halo gas very different from the $j_{\rm hot}(r)$ adopted in GALFORM cooling models.

4.2.2.4 Cooling model in morgana

The full details of this cooling model are given in Monaco et al. (2007) and Viola et al. (2008). The hot gas in a dark matter halo is assumed to be in hydrostatic equilibrium, and a cold halo gas component similar to that in the new cooling model is also introduced. As in the new cooling model, in the continuous time limit, the boundary between the hot gas halo and the cold halo gas is the cooling radius $r_{\rm cool}$. The hot gas halo density and temperature profiles are determined by the assumptions of hydrostatic equilibrium and that the hot gas between $r_{\rm cool}$ and $r_{\rm vir}$ follows a polytropic equation of state. This generally gives more complex profiles than those used in GALFORM and L-GALAXIES but typically the derived density profile is close to the cored β-distribution used in GALFORM while the temperature profile is very flat and close to $T_{\rm vir}$. Therefore in this work, when calculating predictions for this cooling model, for simplicity we will adopt the β-distribution as the hot gas density profile and a constant temperature equaling $T_{\rm vir}$ as the temperature profile. Just as

in the new cooling model, the density profile and temperature of the hot gas halo are updated at every time step.

The MORGANA cooling model then calculates the cooling rate $\dot{M}_{\rm cool}$. However, unlike the cooling models described previously, this does not explicitly depend on the cooling history of the hot gas, as expressed in $t_{\rm cool,avail}$, but instead it assumes that at any given time, each shell of hot gas contributes to $\dot{M}_{\rm cool}$ according to its own cooling time scale. [4], Specifically, this is

$$\dot{M}_{\rm cool} = 4\pi \int_{r_{\rm cool}}^{r_{\rm vir}} \frac{\rho_{\rm hot}(r)}{t_{\rm cool}(r)} r^2 dr, \tag{4.2.44}$$

where $\rho_{\rm hot}(r)$ is the hot gas density at radius r, while $t_{\rm cool}(r)$ is the cooling time scale corresponding to gas density $\rho_{\rm hot}(r)$ and temperature $T_{\rm vir}$, and is given by Eq(4.2.6). This equation is supplemented by another equation

$$\dot{r}_{\rm cool} = \frac{\dot{M}_{\rm cool}}{4\pi\rho_{\rm hot}(r_{\rm cool})r_{\rm cool}^2} - c_{\rm s}(r_{\rm cool}), \tag{4.2.45}$$

where $c_{\rm s}(r_{\rm cool})$ is the local sound speed at radius $r_{\rm cool}$. The first term in Eq(4.2.45) describes the increase of $r_{\rm cool}$ due to cooling. The form of this term is derived from the picture that the cooled down gas all comes from the region near $r_{\rm cool}$, and then mass conservation for a spherical shell gives $\dot{M}_{\rm cool}dt = 4\pi\rho_{\rm hot}(r_{\rm cool})r_{\rm cool}^2 dr_{\rm cool}$. The second term describes the contraction of the hot gas halo due to the reduction of pressure support induced by cooling. Because the hot gas halo is in hydrostatic equilibrium in the gravitational potential well of the dark matter halo, $c_{\rm s}(r_{\rm cool})$ is close to the circular velocity at $r_{\rm cool}$, so the contraction time scale is comparable to $t_{\rm ff}(r_{\rm cool})$. Thus the contraction here is similar to that introduced in the new cooling model, but note that in the MORGANA cooling model the contraction does not include the effect from halo growth, which is included explicitly in the new cooling model using Eq(4.2.21). Together, Eqs.(4.2.44) and (4.2.45) enable the calculation of $r_{\rm cool}$ and $\dot{M}_{\rm cool}$ for each time step.

[4] Viola et al. (2008) introduces a modification of this for a static halo, in which the onset of cooling is delayed by a time duration equaling $t_{\rm cool}(r = 0)$. But this modification is not applied in the full MORGANA model, so here we ignore it and use the cooling model described in Monaco et al. (2007).

There are some physical inconsistencies between Eqs.(4.2.44) and (4.2.45). In Eq(4.2.44), it is assumed the cooled down gas comes from the whole region between $r_{\rm cool}$ and $r_{\rm vir}$, but in Eq(4.2.45) the cooled down gas is assumed to only come from a shell around $r = r_{\rm cool}$. Unless $r_{\rm cool}$ is very close to $r_{\rm vir}$, these two assumptions about the spatial origin of the cooled down gas conflict with each other. Furthermore, Eq(4.2.44) implies that there is differential cooling within a single hot gas shell, with a fraction of the gas cooling completely and the remainder not cooling at all. However, since the gas inside one shell all has the same density and temperature, the whole shell should cool down simultaneously, namely all gas in it cools down after a time $t_{\rm cool}$, but no gas cools down before that time.

The mass of gas cooled down in one timestep is then $\Delta M_{\rm cool} = \dot{M}_{\rm cool}\Delta t$. This mass is used to update the mass of the cold halo gas component, $M_{\rm halo,cold}$, and then the mass accreted onto the central galaxy, $\Delta M_{\rm acc,gal}$, is derived assuming gravitational infall of the cold halo gas component, which is calculated in the same as the new cooling model, using Eq(4.2.18).

The MORGANA cooling model does not explicitly follow the flow of angular momentum. Instead, it assumes that the central galaxy always has its specific angular momentum equal to that of its host dark matter halo, $\bar{j}_{\rm halo}$, with $\bar{j}_{\rm halo} = J_{\rm halo}/M_{\rm halo}$, and $J_{\rm halo}$ and $M_{\rm halo}$ the total angular momentum and mass of the dark matter halo respectively. This is assumption is even more approximate than the treatment in L-GALAXIES.

4.2.3 Halo spin and concentration

All of the cooling models described above require the knowledge of the density profile and angular momentum of the dark matter halo. The former is needed for calculating the free-fall time scale from a given radius, while the latter is required for the calculations of the angular momentum of the gas. Assuming the NFW profile for the dark matter halo, the major task of determining the profile is to determine the halo concentration $c_{\rm NFW}$, because other parameters of the profile, such as halo mass and virial radius, are relatively straightforward to derive given the merger tree. The angular momentum of a halo is usually expressed in terms of the halo spin parameter

$\lambda_{\rm halo}$, which is defined as

$$\lambda_{\rm halo} = \frac{J_{\rm halo}\,|E_{\rm halo}|^{1/2}}{GM_{\rm halo}^{5/2}}, \tag{4.2.46}$$

where $J_{\rm halo}$, $E_{\rm halo}$ and $M_{\rm halo}$ are the total angular momentum, energy and mass of a dark matter halo respectively, and G is the gravitational constant. Thus, the major task of determining halo angular momentum is to determine $\lambda_{\rm halo}$ for a given halo.

Different semi-analytical models use different methods to assign these two parameters to each halo in a merger tree.

The main GALFORM models (e.g. Baugh et al., 2005; Bower et al., 2006; Gonzalez-Perez et al., 2014; Lacey et al., 2016) follow the method introduced in Cole et al. (2000), in which a halo inherits the $c_{\rm NFW}$ and $\lambda_{\rm halo}$ of its most massive progenitor until it undergoes a halo formation event. At a halo formation event, a new $c_{\rm NFW}$ is assigned according to the mass and redshift of this halo through $M_{\rm halo}$-$c_{\rm NFW}$ correlation (Navarro et al., 1997), and a new $\lambda_{\rm halo}$ is randomly selected according to a lognormal distribution. The lognormal distribution of $\lambda_{\rm halo}$ is derived from N-body simulations [e.g. Cole & Lacey (1996); Warren et al. (1992); Gardner (2001), but see Bett et al. (2007) for a different fitting form]. This method introduces sudden jumps in $c_{\rm NFW}$ and $\lambda_{\rm halo}$ at halo formation events even if the halo growth is smooth, which is unphysical. Also, the possible evolution of $c_{\rm NFW}$ and $\lambda_{\rm halo}$ between two adjacent halo formation events is ignored.

L-GALAXIES models use halo merger trees from N-body simulations, and adopt $c_{\rm NFW}$ and $\lambda_{\rm halo}$ measured directly from the halos in these simulations. In principle, this provides the most accurate way to assign $c_{\rm NFW}$ and $\lambda_{\rm halo}$ to a given halo, however, it also has some limitations. Firstly, resolving the halo mass accretion history and thus building merger trees only requires marginal resolution, i.e. a halo should be resolved by at least several tens of particles, but robust measurement of $c_{\rm NFW}$ and $\lambda_{\rm halo}$ requires higher resolution, i.e. a halo should be resolved by at least several hundreds of particles (Neto et al., 2007; Bett et al., 2007). Therefore $c_{\rm NFW}$ and $\lambda_{\rm halo}$ values measured for the smaller halos from an N-body simulation are not be so reliable. Secondly, a semi-analytical model employing this method cannot use Monte Carlo merger trees, which limits its applicability, particularly in building large statistical samples.

The MORGANA model also assigns $c_{\rm NFW}$ according to the $M_{\rm halo}$-$c_{\rm NFW}$ correlation, but it does this at each time step instead of at each halo formation event. By doing so, the artificial sudden jump in $c_{\rm NFW}$ at halo formation events is removed. In the MORGANA model each halo inherits the $\lambda_{\rm halo}$ from its most massive progenitor, while for each halo without progenitor, a value of $\lambda_{\rm halo}$ is assigned randomly according to the lognormal distribution. In this way, $\lambda_{\rm halo}$ is constant in each branch of a merger tree, so there is no artificial jump in its value as in GALFORM models, but the evolution of $\lambda_{\rm halo}$ due to halo growth is completely ignored.

Benson & Bower (2010) and Vitvitska et al. (2002) [see also Maller et al. (2002)] propose another way to assign a value of $\lambda_{\rm halo}$ to each halo. In their method, halos with no progenitor are assigned $\lambda_{\rm halo}$ values randomly according to the $\lambda_{\rm halo}$ distribution derived from N-body simulations, but then the evolution of $\lambda_{\rm halo}$ is calculated based on the orbital angular momenta of the accreted halos. With the halo accretion history given by the merger tree and distributions of orbital parameters derived from N-body simulations, the evolution of $\lambda_{\rm halo}$ can be calculated. One potential problem with this this method is that it assumes that smoothly accreted mass makes no contribution to the evolution of $\lambda_{\rm halo}$. This may not be true, and also whether the accretion is smooth or clumpy is resolution dependent, so this approach omits the effect from unresolved accreted halos, which may affect the long term evolution of $\lambda_{\rm halo}$.

In the present chapter, we follow Cole et al. (2000) to set $c_{\rm NFW}$ and $\lambda_{\rm halo}$ for the GFC1 model, to remain consistent with its original assumptions. For other models, we adopt the method used in the MORGANA model for setting $c_{\rm NFW}$ (i.e. setting it according to the adopted $c_{\rm NFW}$-$M_{\rm halo}$ relation at each timestep), while for the assignment of $\lambda_{\rm halo}$, we introduce a new and simple method. Specifically, a lognormal distribution is adopted to randomly generate spins for halos at the tips of merger trees. The subsequent evolution of $\lambda_{\rm halo}$ is then modelled by a Markov random walk, in which the spins of a halo and its progenitor become approximately uncorrelated when this halo reaches twice its progenitor's mass. In each time step, a conditional probability distribution for the new spin can be constructed for each halo given the mass increase and progenitor $\lambda_{\rm halo}$, and then a value of $\lambda_{\rm halo}$ is assigned

randomly according to this conditional distribution. This method allows large spin changes when the halo mass increases by a large factor, i.e. in major mergers, and small but usually nonzero changes for small mass increases, which are typical in smooth halo growth. More details of this random walk method are provided in Appendix C, together with some comparisons of the predictions of this method with results from N-body simulations.

We have checked that all the results presented in this chapter are not sensitive to the methods used for assigning $c_{\rm NFW}$ and $\lambda_{\rm halo}$.

4.3 Results

This section presents predictions from the new cooling model, and compares them with the corresponding results from the earlier cooling models described in the previous section. We start, in §4.3.1, by considering the cooling histories for the simplest case of a static halos, and then, in §4.3.2, consider the more realistic case of evolving halos with full merger histories. Finally, in §4.3.3, we show the effects of using the new cooling model within a full galaxy formation model.

4.3.1 Static halo

For the static halo case, we consider dark matter halos having fixed mass $M_{\rm halo}$, and also a fixed density profile, corresponding to a halo that forms at redshift z. We present 4 cases that show the range of behaviour: $M_{\rm halo} = 10^{11}\,{\rm M}_{\odot}$ (low mass and fast cooling halo), $M_{\rm halo} = 10^{12}\,{\rm M}_{\odot}$ (Milky Way like halo), $M_{\rm halo} = 10^{13}\,{\rm M}_{\odot}$ (group halo) and $M_{\rm halo} = 10^{14}\,{\rm M}_{\odot}$ (cluster halos). For $M_{\rm halo} = 10^{11}\,{\rm M}_{\odot}$ we choose $z = 3$, while for the other cases we choose $z = 0$. The core radius of the β-distribution for hot gas is set to be $r_{\rm core} = 0.07 r_{\rm vir}$. To isolate the effects of the different cooling models, star formation and feedback processes are turned off.

Fig. 4.2 shows the total mass, angular momentum and the specific angular momentum of the gas that has cooled down and accreted onto the central galaxy, as predicted by the different cooling models. Results are plotted against the time t since radiative cooling is turned on in the halo. For the fast cooling halo ($M_{\rm halo} = 10^{11}\,{\rm M}_{\odot}$),

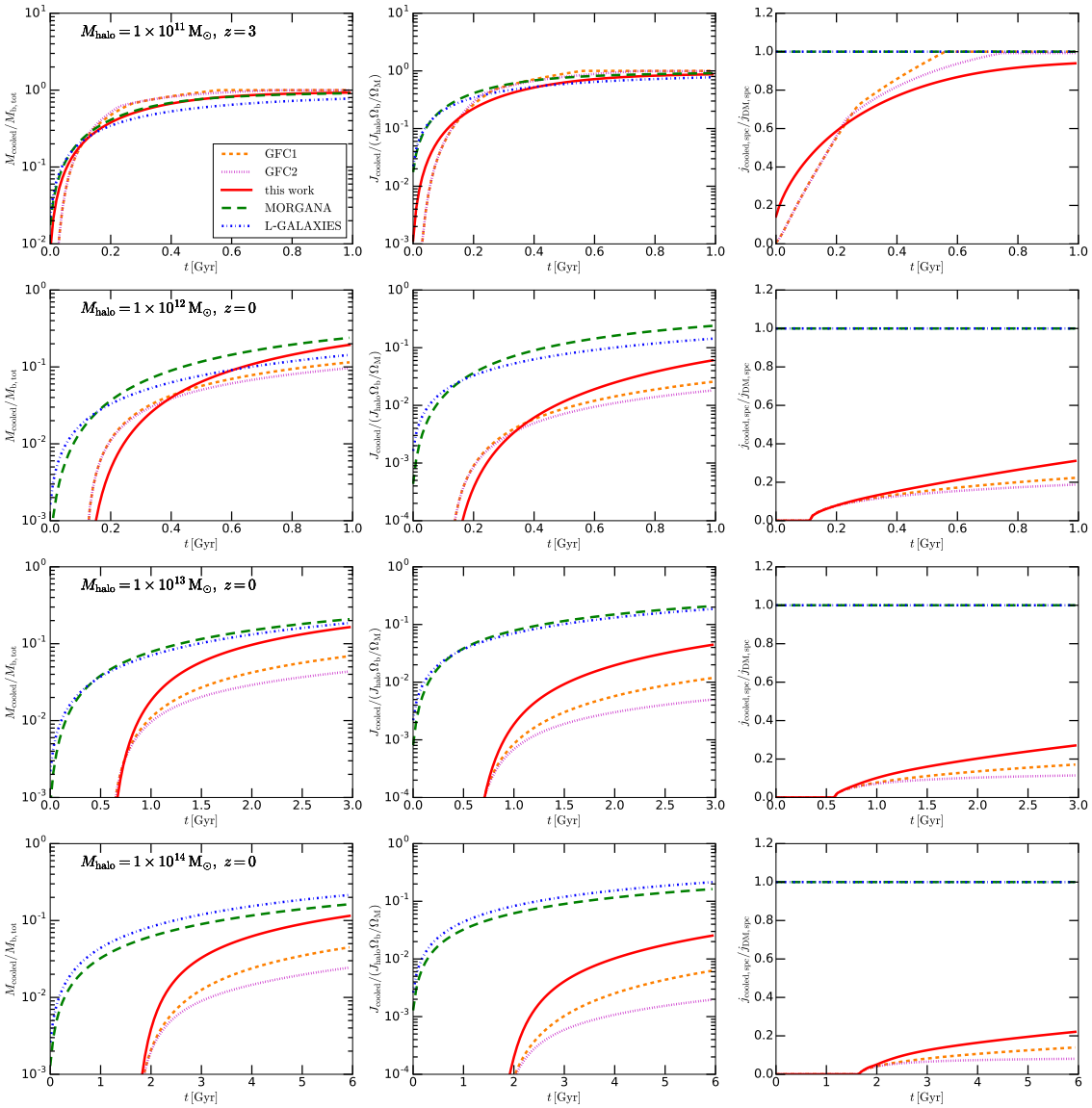

Figure 4.2: Cooling histories for static halos, measured from when radiative cooling is turned on. The different line styles show the predictions for different cooling models, as labelled in the line key. Each row corresponds to a different halo mass and assembly redshift, as labelled. Left column: the ratio of cooled mass to total baryon mass. Middle column: the ratio of angular momentum of cooled gas to total baryon angular momentum, which is defined as $J_{\rm halo}\Omega_{\rm b}/\Omega_{\rm M}$, where $J_{\rm halo}$ is the angular momentum of the halo. Right column: the ratio of the specific angular momentum of the cooled gas to the specific angular momentum of the halo.

all cooling models predict very similar results for all three quantities. This is because in the fast cooling regime, the accretion of the cooled down gas is mainly limited by the timescale for free-fall rather than that for radiative cooling, and all of the cooling models model the free-fall accretion in similar ways. For the more massive halos ($M_{\rm halo} = 10^{12} - 10^{14}\,{\rm M}_{\odot}$), while the results for the L-GALAXIES and MORGANA cooling models remain very similar, the results for the GFC1, GFC2 and new cooling models diverge from those models and from each other.

For halos of all masses, gas starts to accrete onto the central galaxy from $t = 0$ in the L-GALAXIES and MORGANA cooling models, while for the GFC1, GFC2 and new cooling models there is a time delay that varies with halo mass. This time delay is equal to the central radiative cooling timescale, $t_{\rm cool}(r = 0)$. It is a consequence of the assumption that the hot gas density decreases monotonically with radius, so that $t_{\rm cool} \propto \rho_{\rm hot}(r)^{-1}$ increases with radius. In the GALFORM cooling models, the hot gas density at $r = 0$ is finite, and gas cools shell by shell, so no gas can cool and accrete before the gas at the centre cools. In contrast, in the L-GALAXIES cooling model, the hot gas density at $r = 0$ is infinite, while in MORGANA, the gas does not cool shell by shell, so there is no time delay.

For the Milky Way like halo, the GFC1 and GFC2 models generally predict lower accreted masses than the new cooling model, and this difference grows with halo mass. For the $10^{14}\,{\rm M}_{\odot}$ halo, the difference can be a factor > 4. The origin of this difference can be understood by looking at the cooling in more detail, as is done in Fig. 4.3. For conciseness, we only show the most massive halo, where the abovementioned difference is largest. The results for less massive halos are similar.

The upper left panel of Fig. 4.3 shows the time evolution of the cooling radius $r_{\rm cool}$. The GFC1 model predicts that $r_{\rm cool}$ increases monotonically with time. This is expected for a fixed hot gas halo, in which the hot gas cools down at larger and larger radii with increasing time. For the GFC2 and new cooling models, however, $r_{\rm cool}$ tends to reach a stable value instead of increasing with time. This is caused by the contraction of hot gas halo included in these two models. Although the radiative cooling leads to the increase of $r_{\rm cool}$ just as in the GFC1 model, the contraction moves the hot gas shells to smaller radii, and the competition of these two factors

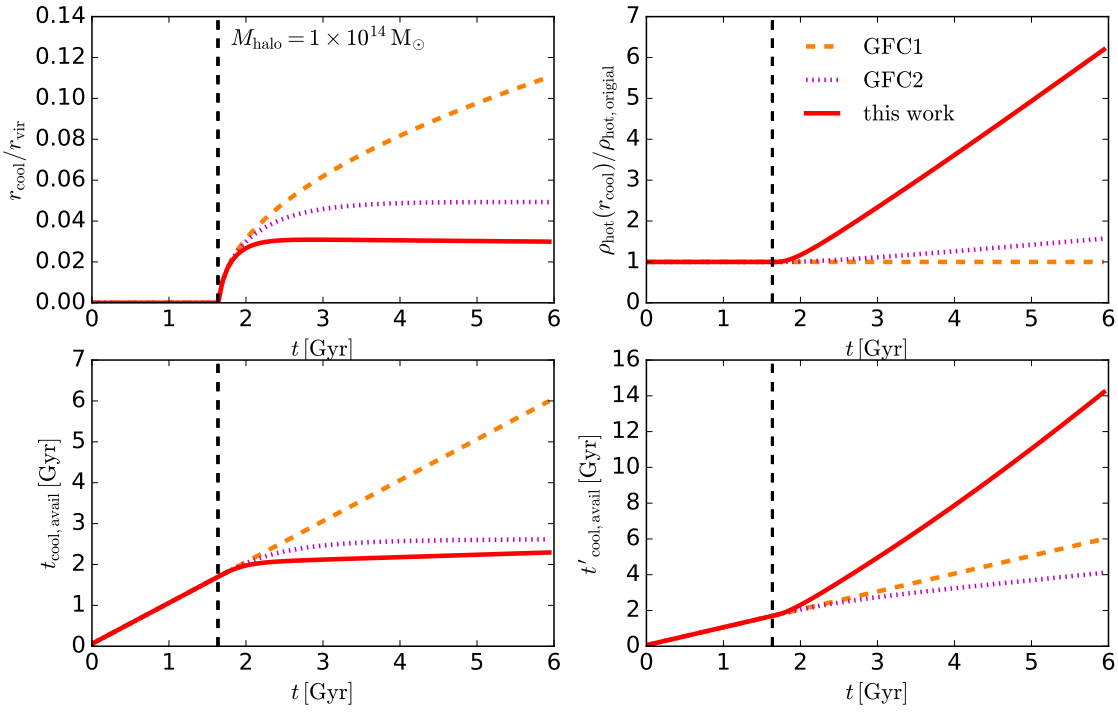

Figure 4.3: More detailed information on the cooling in static halos in the three GALFORM cooling models for $M_{\rm halo} = 10^{14}\,{\rm M}_\odot$. From upper left to lower right, these four panels respectively show the time evolutions of the cooling radius $r_{\rm cool}$, ratio of the density of the shell at $r_{\rm cool}$ to the density of the same Lagrangian at $t = 0$, the time available for cooling $t_{\rm cool,avail}$ and the scaled time available for cooling $t'_{\rm cool,avaik}$, predicted by the three models. Each line style corresponds to a different model, with the model name given in the key in the upper right panel. The vertical dashed line in each panel indicates the moment at which cooling starts.

results in $r_{\rm cool}$ approaching a nearly constant value. The GFC2 model predicts larger values of $r_{\rm cool}$ than the new cooling model, because, as mentioned in §4.2.2.2, these two models adopt different contraction time scales, and the GFC2 model tends to overestimate the contraction timescale, leading to a slower contraction, and resulting in values of $r_{\rm cool}$ intermediate between the GFC1 and new cooling models.

When a hot gas shell moves to smaller radius, it is compressed to a higher density. This effect is shown in the upper right panel of Fig. 4.3. This panel gives the ratio of the density of the gas at $r_{\rm cool}$ to the density, $\rho_{\rm hot,original}$, in the same Lagrangian gas shell at $t = 0$. This ratio is always 1 for the GFC1 model, because it assumes a static hot gas halo, while for the GFC2 and new cooling models it is larger than 1, due to the compression induced by the hot halo contraction.

The lower left panel of Fig. 4.3 shows the $t_{\rm cool,avail}$ predicted by the three models. The prediction of the GFC1 model is just the physical time, while those of the GFC2 and new cooling models tend to level off at a constant value. $t_{\rm cool,avail}$ represents the previous cooling history of the hot gas. The advance of the cooling tends to increase $t_{\rm cool,avail}$ by increasing $E_{\rm cool}$ in Eq(4.2.13), while the hot gas halo contraction in the GFC2 and new cooling models increases the shell density, the cooling rate and thus tends to reduce $t_{\rm cool,avail}$ by increasing $L_{\rm cool}$ in in Eq(4.2.13). The combination of these two effects causes $t_{\rm cool,avail}$ to approach a roughly stable value.

In the GFC2 and the new cooling models $t_{\rm cool,avail}$ is used to calculate the cooled mass for the hot gas halo after contraction. As shown in the upper right panel of Fig. 4.3, the extent of contraction is different in these two models, while the GFC1 model does not have this contraction. Thus, the $t_{\rm cool,avail}$ in these three models are for different hot gas halos. This makes it complicated to analyze the origin of the differences in predicted cool mass based on $t_{\rm cool,avail}$. Therefore, we introduce another quantity, $t'_{\rm cool,avail}$, which is defined as

$$t'_{cool,avail} = t_{\rm cool,avail}\frac{\rho_{\rm hot}(r_{\rm cool})}{\rho_{\rm hot,original}}, \tag{4.3.1}$$

where $\rho_{\rm hot}(r_{\rm cool})$ is the density of the shell that has just cooled down, while $\rho_{\rm hot,original}$ is the density at $t = 0$ of the same Lagrangian shell, and this density ratio is that shown in the upper right panel of Fig. 4.3. Because for the shell just cooled down

one has $t_{\rm cool,avail} = t_{\rm cool}(r_{\rm cool})$, so Eq(4.3.1) implies that

$$\begin{aligned} t'_{\rm cool,avail} &= t_{\rm cool}(r_{\rm cool})\frac{\rho_{\rm hot}(r_{\rm cool})}{\rho_{\rm hot,original}} \\ &= t_{\rm cool,original}, \end{aligned} \tag{4.3.2}$$

where $t_{\rm cool,original}$ is the cooling timescale of this Lagrangian shell at $t = 0$. Then it is clear that $t'_{\rm cool,avail}$ is linked to the cooling timescale at the initial moment, at which the hot gas halo is the same in all three models, and so is easier to compare between models.

The lower right panel of Fig. 4.3 shows the $t'_{\rm cool,avail}$ predicted by the three models. The new cooling model predicts the highest $t'_{\rm cool,avail}$, which means at any given time, the shell at $r_{\rm cool}$ in this model has the largest initial radius among the three models, and because at $t = 0$ the hot gas halo density profile is the same for these three models, the largest radius implies the highest cooled mass. In contrast, the GFC2 model predicts the smallest $t'_{\rm cool,avail}$, so it predicts the lowest cool mass.

The density enhancement ($\rho_{\rm hot}(r_{\rm cool})/\rho_{\rm hot,original} > 1$) seen in the GFC2 and new cooling models implies a higher cooling luminosity than for the case of a fixed hot gas halo as in the GFC1 model. This higher cooling luminosity means more thermal energy is radiated away by a given time, and because the hot gas halos in these three models all have the same temperature, this higher thermal energy loss should mean higher cooled mass. Therefore, it would be expected that for a cooling model with density enhancement, its predicted cooled mass should be higher than that of a model assuming a fixed hot gas halo. Also, a higher cooled mass means the shell cooled down was initially at larger radius, and because the density decreases with increasing radius for the assumed initial density profile, this larger radius implies lower initial density and longer original cooling timescale $t_{\rm cool,original}$. Therefore, for a given $\rho_{\rm hot}(r_{\rm cool})/\rho_{\rm hot,original}$, insofar as this ratio is greater than one, the expected $t'_{\rm cool,avail}$ should be larger than in a model with a fixed hot gas halo, i.e. the GFC1 model.

The new cooling model does predict cooled mass and $t'_{\rm cool,avail}$ larger than those in the GFC1 model, but the GFC2 model predicts these to be lower than in the GFC1 model, which contradicts the physical expectation above. Thus, the GFC2

model appears to be physically inconsistent, and the $t'_{\rm cool,avail}$ in it tends to be too small. Furthermore, because $t'_{\rm cool,avail}$ and $t_{\rm cool,avail}$ are related by the density ratio through Eq(4.3.1), for a given density ratio, the underestimation of $t'_{\rm cool,avail}$ also implies an underestimation of $t_{\rm cool,avail}$.

To understand why $t_{\rm cool,avail}$ is underestimated in the GFC2 model, consider the following. As described in §4.2.2.2, the GFC2 model accumulates the total energy radiated away for the current hot gas halo (Eq 4.2.40) and then divides it by the current halo cooling luminosity to estimate $t_{\rm cool,avail}$. When some gas cools down from the hot gas halo, its contribution to the total energy radiated away should be removed, because this gas is no longer part of the hot gas halo, and this is the motivation for the second term in Eq(4.2.40). This term basically removes the total thermal energy corresponding to the mass removed from the hot gas halo. This would be correct if the GFC2 model exactly accumulated the total energy radiated away by cooling. However, GFC2 model adopts a very rough approximation (Eq 4.2.38), in which the cooling luminosity of a gas shell is approximated as $\delta L_{\rm cool} = 4\pi\tilde{\Lambda}\rho^2_{\rm hot}(r)r^2dr \approx 4\pi\tilde{\Lambda}\bar{\rho}_{\rm hot}\rho_{\rm hot}(r)r^2dr$, with $\tilde{\Lambda}$ being the cooling function and $\bar{\rho}_{\rm hot}$ the mean density of the hot gas. For the β-distribution used for the static halo comparison, $\bar{\rho}_{\rm hot} \sim \rho_{\rm hot}(r = 0.5r_{\rm vir})$, and for the group and cluster halos, cooling happens in the region where $\rho_{\rm hot}(r) > \bar{\rho}_{\rm hot}$. Thus the approximation underestimates the energy radiated away, and so the second term in Eq(4.2.40) removes more energy than necessary, leading to an underestimation of $t_{\rm cool,avail}$. The final cooling depends on the relative strength of this underestimation and the density enhancement. For the static halo considered here, this underestimation of $t_{\rm cool,avail}$ exceeds the effects of the density enhancement and leads to even less gas cooling down than in the GFC1 model, but for other cases, the results could be different.

Overall, the introduction of the contraction of the hot gas halo in the new cooling model results in more efficient cooling than in the more traditional GALFORM cooling model GFC1. Some previous works (De Lucia et al., 2010; Monaco et al., 2014) also noticed that the GFC1 model tends to predict less gas cooling than other cooling models such as MORGANA and L-GALAXIES, and also less than SPH hydrodynamical simulations. These works suggested using more centrally concentrated hot

gas density profiles such as the singular isothermal profile to bring semi-analytical predictions into better agreement with SPH simulations. However, the results here suggest that at least part of the reason for the GFC1 model giving low cooling rates is that it does not include contraction of the hot gas halo as cooling proceeds. Note that the enhancement of hot gas gas density and hence cooling induced by contraction is also mentioned in MORGANA papers (e.g. Viola et al., 2008), but take the average over all hot gas shells to calculate the mass cooling rate (as is done in the MORGANA cooling model) may not be the best way to model this effect.

Fig. 4.2 also shows that for the halos other than the fast cooling halo, the different cooling models predict different angular momenta for the gas in the central galaxies. The L-GALAXIES and MORGANA cooling models predict higher angular momentum and higher specific angular momentum than the GFC1, GFC2 and new cooling models. They predict higher total angular momentum in part because these two models tend to predict more cooled down mass, but more importantly because they (implicitly) assume specific angular momentum distributions of the hot gas, $j_{\rm hot}(r)$, that are very different from the three GALFORM models. The L-GALAXIES cooling model assumes that the gas accreting in the current timestep has specific angular momentum equal to the mean specific angular momentum of the dark matter halo. This corresponds to $j_{\rm hot}(r) =$ constant, i.e. no dependence on the radius from which the gas is cooling. The MORGANA cooling model instead assumes that the mean specific angular momentum of all the gas that has cooled down and accreted onto the central galaxy over its past history is equal to the mean specific angular momentum of the current dark matter halo. In the static halo case, in which the mean specific angular momentum of the halo does not change with time, the assumption in the MORGANA model is equivalent to that in L-GALAXIES cooling model. As shown in the right column of Fig. 4.2, this results in the mean specific angular momentum of the cold gas in central galaxies being equal to the mean halo specific angular momentum at all times for these two models, in the case of a static halo.

In contrast, the GFC1, GFC2 and new cooling models assume that $j_{\rm hot}(r)$ increases with radius, and that the mean specific angular momentum of all the baryons in a halo is equal to the mean specific angular momentum of the halo. Under this

assumption, the hot gas in the central region has lower specific angular momentum than the mean for the halo. For the halos other than the fast cooling halo, typically only part of the hot gas cools down, and because the cooling proceeds from halo center outwards, the hot gas having low specific angular momentum cools first, so the predicted mean specific angular momentum of the cold gas in central galaxies is lower than that of the dark matter halo. The new cooling model predicts higher specific angular momentum for the cooled gas in central galaxies compared to the GFC1 and GFC2 models, because it cools more effectively, and so can cool gas shells that were originally at larger radii, which, according to the assumed $j_{\rm hot}(r)$, have higher specific angular momentum.

4.3.2 Cosmologically evolving halos

Having understood the behaviors of the different cooling models in the simplified case of static halos, the next step is to compare the behaviors of these models in the context of cosmic structure formation. To achieve this, we run the cooling models in cosmologically evolving halos, whose formation histories are described by merger trees. We choose 4 different halo masses at $z = 0$, namely $M_{\rm halo} = 10^{11},\ 10^{12},\ 10^{13}$ and $10^{14}\,{\rm M}_{\odot}$. For each of these halo masses, we generate 100 independent merger trees to sample the range of formation histories, using the Monte Carlo method of Parkinson et al. (2008) that is based on the Extended Press-Schechter approach (e.g. Lacey & Cole, 1993). (We use Monte Carlo rather than N-body merger trees for this comparison because it is then easier to build equal size samples for different $z = 0$ halo masses.) The merger trees are are built with halo mass resolution $M_{\rm res} = 5 \times 10^9\,{\rm M}_{\odot}$. We choose this relatively high $M_{\rm res}$ mainly to avoid too much cooling in small halos, which would leave little gas for the slow cooling regime in high mass halos. Star formation, SN and AGN feedback processes and galaxy mergers are all turned off in order to isolate the effects of the different cooling models. For each merger tree, the mass and angular momentum of the gas cooled and accreted onto the central galaxy in the halos in the major branch of this merger tree are recorded.

Fig. 4.4 shows the medians over 100 realizations for each halo mass of the mass,

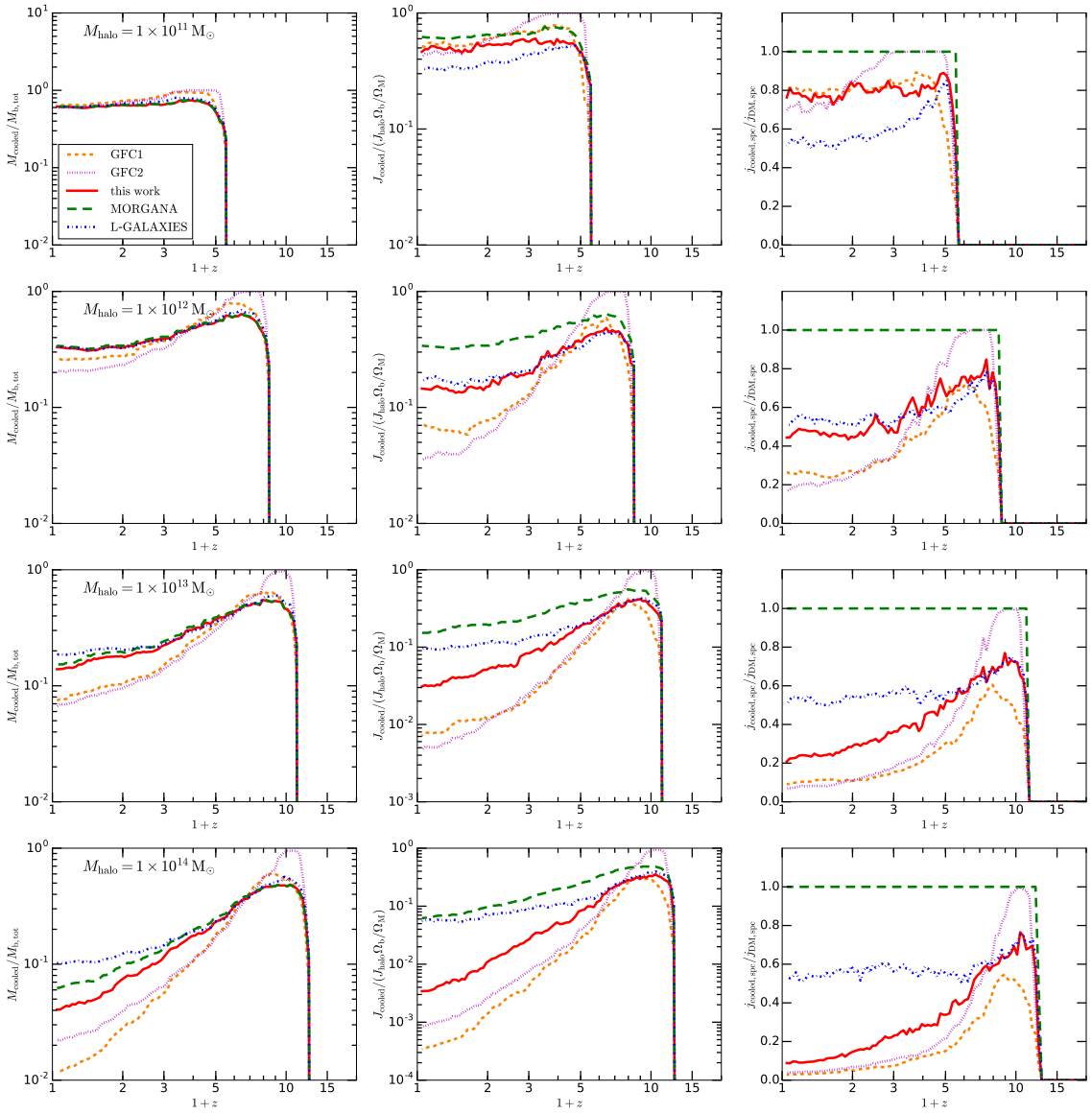

Figure 4.4: The same as Fig. 4.2, but for cosmologically evolving halos. For each $z = 0$ halo mass shown in the figure, 100 merger trees are constructed, and the cooling models run on all branches of each merger tree. Star formation, SN and AGN feedback, and galaxy mergers are turned off. The results are recorded for the main branch of each merger tree, and the medians over each halo sample are plotted.

angular momentum and specific angular momentum of gas accreted onto the central galaxy in the main branch of the merger tree. Many features seen in the static halo case also appear here. For the fast cooling halos ($M_{\rm halo} = 10^{11}\,{\rm M}_{\odot}$ at $z = 0$), the predictions of the different cooling models are similar, again because in the fast cooling regime the accretion of gas onto galaxies is limited by the free-fall timescale and largely insensitive to the details of the cooling calculation. For the slower-cooling halos ($M_{\rm halo} \geq 10^{12}\,{\rm M}_{\odot}$), the new cooling model predicts larger cooling masses than the GFC1 and GFC2 models, because of the contraction of the hot gas halo. For halos less massive than $10^{14}\,{\rm M}_{\odot}$, the predictions of the new cooling model for the mass cooled down are close to those of the L-GALAXIES and MORGANA cooling models, but for $10^{14}\,{\rm M}_{\odot}$ halos, the predictions of the new cooling model at $z = 0$ are about a factor of 2 lower than those of MORGANA, and a factor of 3 than those of L-GALAXIES.

In the static halo case, the cooled down mass predicted by the GFC2 model is always lower than that of the GFC1 model, but here the relation of their predictions is more complex. For some halo masses, the GFC2 model gives higher cooled down masses, but for other halo masses, its predictions are lower. This is because the diverse halo merger histories affect the comparative strengths of the underestimation of $t_{\rm cool,avail}$ and the enhancement of the hot gas density in the GFC2 model, and the competition of these two factors determines the final cooling efficiency of this model, as described in §4.3.1.

The MORGANA cooling model forces the specific angular momentum of the cooled down gas to always equal to the mean specific angular momentum of the halo by construction. Although the L-GALAXIES cooling model makes the same prediction in the static halo case, for dynamically evolving halos, the L-GALAXIES cooling model predicts lower specific angular momenta. This is because L-GALAXIES assumes that the gas currently cooling and accreting onto the central galaxy has specific angular momentum equal to that of the current halo. For cosmologically evolving halos, the halo specific angular momentum typically increases as the halo grows, so the gas cooled at earlier times tends to have lower specific angular momentum, and so the total mean specific angular momentum of all of the gas that has cooled up to that

time is lower than the mean value of the current halo.

The new cooling model tends to give higher specific angular momentum than the GFC1 and GFC2 models, mainly because the new cooling model can cool gas that was originally at larger radii, which according to the assumed $j_{\rm hot}(r)$ has higher specific angular momentum.

For the dynamical halo case, a new phenomenon is that for halos with $M_{\rm halo} \gtrsim 10^{12}\,{\rm M}_\odot$ at $z = 0$, the GFC1, GFC2 and new cooling models predict lower specific angular momentum for the cooled down gas at $z = 0$ than the L-GALAXIES cooling model, while for halos with $M_{\rm halo} \lesssim 10^{12}\,{\rm M}_\odot$ $M_{\rm halo} = 10^{11}\,{\rm M}_\odot$ at $z = 0$, the reverse is true. This can be understood as follows:

In the absence of cooling, all four models would predict that the mean specific angular momentum of the hot gas always equals that of the dark matter halo. Typically the specific angular momentum of the dark matter halo increases as it grows, which means that the specific angular momentum of gas accreting later is higher than that of gas accreting earlier. In the presence of cooling, the gas that accreted earlier is more likely to cool, so the mean specific angular momentum of the remaining is higher than that of the dark matter halo.

For slower-cooling halos (those with $M_{\rm halo} \gtrsim 10^{12}\,{\rm M}_\odot$ at $z = 0$), typically only a small fraction of the hot gas halo cools down, so the mean specific angular momentum of the hot gas cannot be much different from that of the dark matter halo. Moreover, the cooling in this case typically happens at small radii, and because the GFC1, GFC2 and new cooling models all assume $j_{\rm hot}(r)$ increases with r, they predict that the gas that is currently cooling has lower specific angular momentum than the dark matter halo, and so also lower than the predictions of the L-GALAXIES cooling model.

For the faster cooling halos (those with $M_{\rm halo} \lesssim 10^{12}\,{\rm M}_\odot$ at $z = 0$), most of the gas cools, so the specific angular momentum of the remaining hot gas can end up significantly larger than that of the halo. Since the gas ends up cooling from large radii, the specific angular momentum of the gas that cools in a single timestep may be larger than the mean for the dark halo. This effect is more or less captured in the GFC1, GFC2 and new cooling models, but not in the L-GALAXIES cooling model,

which is why for this case L-GALAXIES predicts lower specific angular momentum for the cooled down gas as a whole compared to the GALFORM cooling models.

4.3.3 Full galaxy formation model

In this section we show the effects of implementing the new cooling model in a full galaxy formation model. The GALFORM, L-GALAXIES and MORGANA semi-analytical models have very different modeling of galaxy sizes, star formation, black hole growth and SN and AGN feedback. A full comparison of these models is not the aim of this chapter, so here we restrict our scope to the GALFORM model, and investigate the effects of the new cooling model on a recent version of GALFORM, namely Lacey16 (Lacey et al., 2016). As previously mentioned, the Lacey16 model adopts the GFC1 model for gas cooling in halos.

In our comparison, we focus on two important galaxy properties. The first one of these is the field galaxy luminosity function (field LF) at $z = 0$. This reflects the abundance of galaxies of different masses, and reproducing the observed field LFs is typically a basic requirement for any successful galaxy formation model. The second one is the galaxy size-luminosity relation. This is of special interest because the new cooling model predicts specific angular momenta for galaxies that are significantly different from previous cooling models.

We first compare the original Lacey16 model to variants using the new cooling model, while keeping the other parameters fixed at their original values. In the original Lacey16 model, as in earlier published GALFORM models using the GFC1 cooling model, the halo virial velocity $v_{\rm vir}$ is updated only at halo formation events, while in the new cooling model $v_{\rm vir}$ is normally updated at every timestep. Changing how $v_{\rm vir}$ is updated by itself results in significant changes in some GALFORM predictions. To separate more clearly the effects of the new cooling model from the effects of how $v_{\rm vir}$ is updated, we define several variants which we then compare: Lacey16+cv, which is identical to the original Lacey16 model except that $v_{\rm vir}$ is updated at every timestep; Lacey16+new cool, which is the Lacey16 model with the new cooling model except with $v_{\rm vir}$ updated at formation events; and Lacey16+cv+new cool model, which is the Lacey16 model with the new cooling model and with $v_{\rm vir}$ updated at every

timestep (the default case for the new cooling model). These variants are discussed in §4.3.3.1. As shown below, the Lacey16+cv+new cool model without retuning does not provide a good match to the observed field galaxy luminosity functions at $z = 0$, so we then introduce a retuned model, Lacey16+cv+new cool + retuned, in which some of the other GALFORM parameters are adjusted to provide a better fit to these data. This retuned model is discussed in §4.3.3.2.

4.3.3.1 Original Lacey16 and its variations

We first consider galaxy luminosity functions. Fig. 4.5 shows the present-day $b_{\rm J}$- and K-band field luminosity functions predicted by the different model variants described above, compared with observational data. The original Lacey16 model was calibrated to provide a good fit to the observed LFs. Updating the halo virial velocity $v_{\rm vir}$ at every timestep, as for the variant Lacey16+cv, is seen by itself to produce only small changes in the LFs, reducing them slightly at the faint end. However, replacing the original cooling model (GFC1) with the new cooling model is seen to produce a large increase in the number of bright galaxies, although this effect is smaller in the model Lacey16+cv+new cool where $v_{\rm vir}$ is updated at every timestep (lower panels), compared to the model Lacey16+new cool where it is only updated at formation events(upper panels). In the Lacey16 model, the bright ends of the LFs at $z = 0$ are controlled mainly by AGN feedback. The excesses seen in the bright ends show that the AGN feedback is too weak when the new cooling model is introduced without adjusting any other parameters. There are two reasons for this. Firstly, as shown in §4.3.1 and 4.3.2, by more carefully modeling the contraction of the hot gaseous halo, the new cooling model predicts higher cooling luminosity and more efficient cooling, which requires stronger AGN feedback to balance it. Secondly, the efficiency of the AGN feedback is tightly correlated with the growth of supermassive black holes (SMBH) at the centres of galaxies. As discussed in Lacey et al. (2016), in the Lacey16 model, the starbursts triggered by bar instabilities in galaxy disks are a major contributor to black hole growth. The new cooling model generally predicts higher angular momentum for the cooled down gas, resulting in larger disk sizes, and delaying the onset of disk instabilities (typically by $\sim 5\,$Gyr).

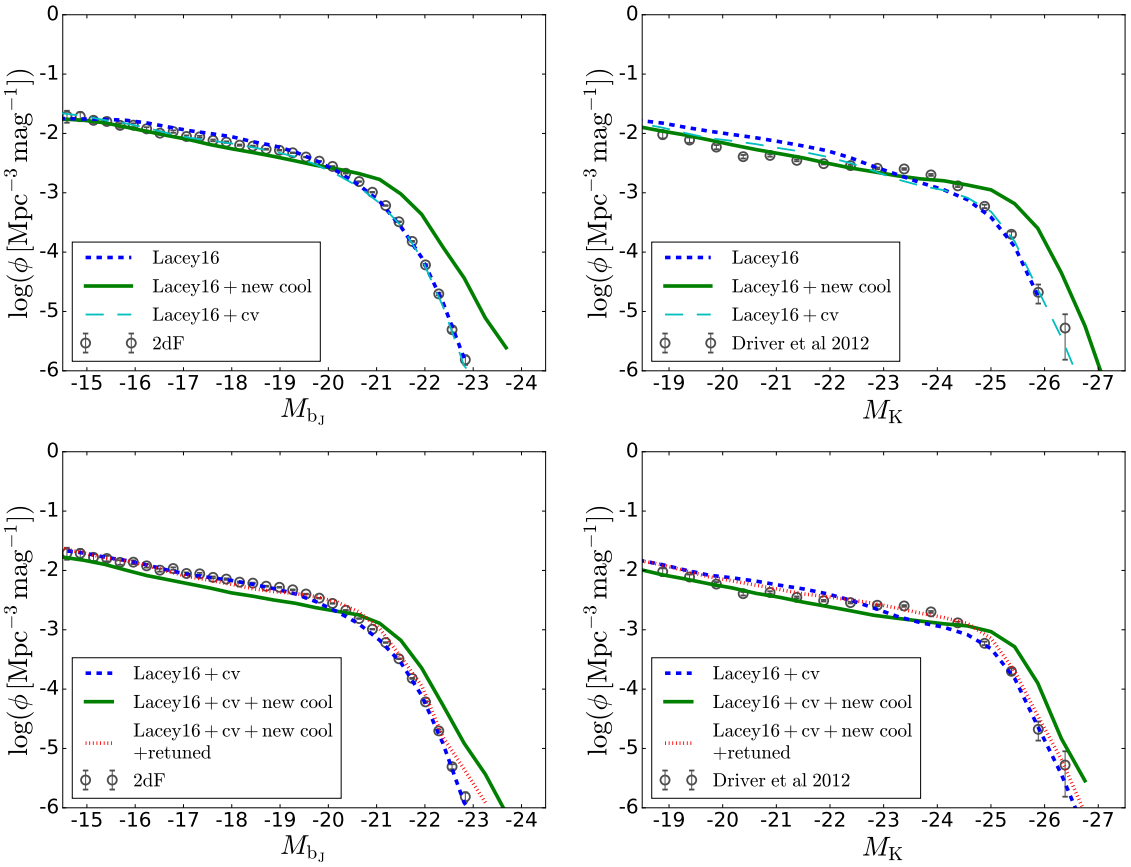

Figure 4.5: The field galaxy luminosity functions at $z = 0$ in the $b_{\rm J}$ and K bands for different variants of the Lacey16 model. Each line shows the prediction for a different model, with the corresponding model name given in the key. The different models are described in the text. The top panels show the original Lacey16 model and the separate effects of changing to the new cooling model or of updating the halo virial velocity at every timestep. In the bottom panels all models have the halo virial velocity updated at every timestep. The bottom panels include the retuned Lacey16 model incorporating the new treatment of gas cooling. The gray open circles with error bars are observational data, from Norberg et al. (2002) for the $b_{\rm J}$-band and from Driver et al. (2012) for the K-band.

This then delays the onset of efficient AGN feedback, leading to ineffective AGN feedback over most of the history of a galaxy.

A further effect of using the new cooling model that is apparent in Fig. 4.5 is to lower the faint ends of the field LFs relative to the corresponding models using the GFC1 cooling model. However, this change is fairly modest, less than a factor of 2. This difference indicates that the new cooling model predicts less gas cooling in the halos forming these faint galaxies, which are typically low mass ($M_{\rm halo} \lesssim 10^{12}\,{\rm M}_{\odot}$) and close to the fast cooling regime. At first sight, this seems to contradict the conclusions in §4.3.1 and 4.3.2, which claim that the cooling in low mass halos predicted by the different cooling models is similar. However, the models used in §4.3.1 and 4.3.2 do not include SN feedback and so there is no reincorporation of the gas ejected out of the halo by SN feedback. In the full model here, this ejected gas plays a central role in the gas cooling, because the faint galaxies have very strong SN feedback, and so a large fraction of their gas is ejected and later reaccreted.

Both the new cooling model and the GFC1 model assume that the ejected gas is gradually reincorporated into the hot gas halo, and when it joins the hot gas halo, it is shock heated to $T_{\rm vir}$, so that it joins as hot gas without any previous cooling history. However, as mentioned in §4.2.2.1, the GFC1 model always calculates $t_{\rm cool,avail}$ as the time since the last halo formation event, which means that ejected gas that is reincorporated between two halo formation events is treated as having been cooling for longer than it has been part of the hot halo. In contrast, the new cooling model estimates the cooling history by accumulating the energy previously radiated away, $E_{\rm cool}$, and the reincorporation of the ejected gas does not change $E_{\rm cool}$. This difference in the treatment of the reincorporated gas causes the new cooling model to predict less cooling in these low mass halos. The strength of this effect depends on the amount of gas ejected, so only the galaxies experiencing strong SN feedback are strongly affected.

We now consider galaxy sizes. Fig. 4.6 shows the r-band half-light radius vs. r-band absolute magnitude relations for both late-type and early-type galaxies at $z = 0$. The original Lacey16 model predicts too large sizes for faint late-type galaxies ($M_{\rm r} \gtrsim -20$) and for faint early-type galaxies $v_{\rm vir}$ ($M_{\rm r} \gtrsim -21$). The Lacey16+cv

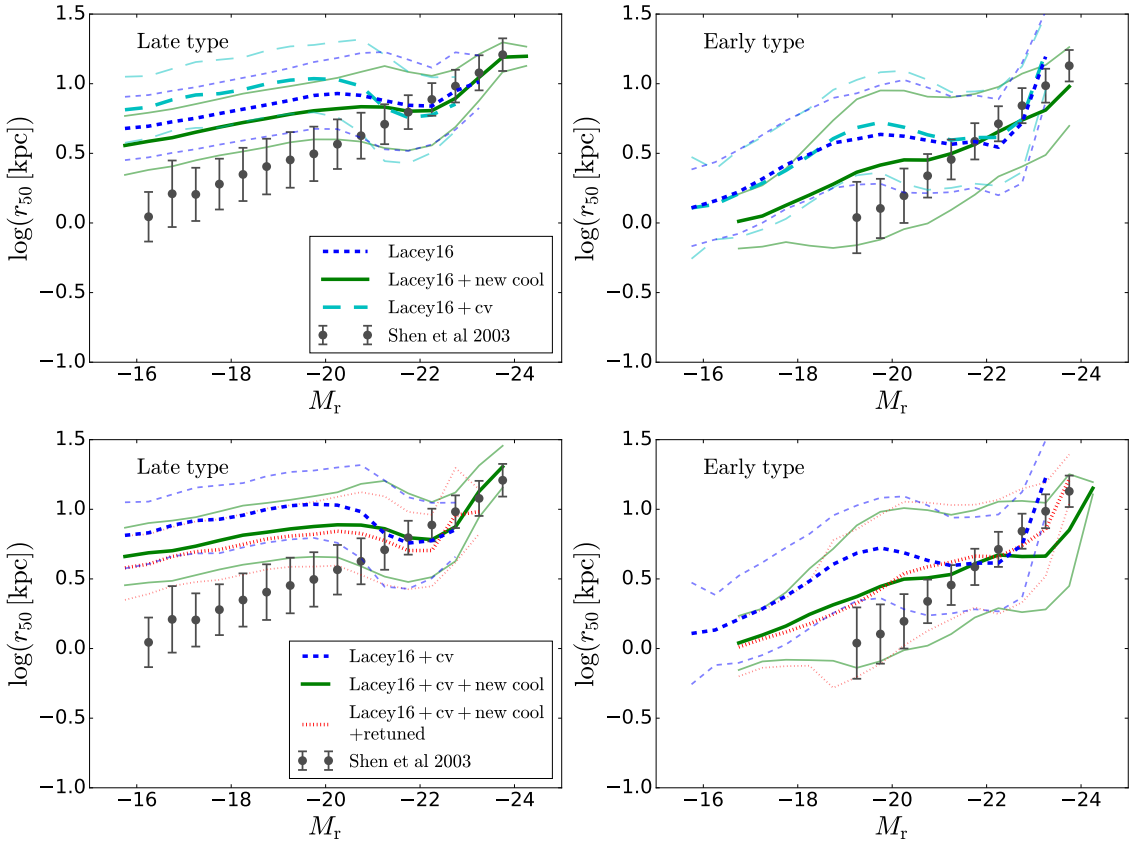

Figure 4.6: Half-light radii of late-type (left column) and early-type (right column) galaxies vs. luminosity at $z = 0$. Both the half-light radius and luminosity are for the r-band. The models plotted and their arrangement between top and bottom rows are the same as in Fig. 4.5. The thick lines show the median relation, while the corresponding thin lines indicate the 10-90% range around this. In the models, galaxies are defined as late- or early-type according to their r-band bulge to total ratio $(B/T)_r$, with $(B/T)_r < 0.5$ for late-type and $(B/T)_r > 0.5$ for early-type. The gray dots with errorbars show medians and 10-90% ranges based on observational data from Shen et al. (2003). Shen et al. (2003) measured the half-light radii by fitting Sersic profiles to galaxy images and defined the late-type and early-type galaxies by Sersic index $n < 2.5$ and $n > 2,5$ respectively. The late-type galaxy sizes have been multiplied by 1.34 to make an average correction to face-on values (see §4.3.2 of Lacey et al. (2016) for more details).

model, in which $v_{\rm vir}$ is updated at every timestep, gives similar results, with the predicted sizes of faint late-type galaxies being even larger. Using the new cooling model, as in Lacey16+cv+new cool, then reduces the sizes of faint late-type galaxies compared to the Lacey16+cv model, due to the reduction of gas cooling when including the reincorporated gas. However, the sizes of late type galaxies in the Lacey16+cv+new cool model are almost the same when compared to the original Lacey16 model. This indicates that some physical effect other than gas cooling in halos may be responsible for the deviation of the model prediction from observations for late-type galaxies.

Using the new cooling model results in a larger improvement in the size-luminosity correlation of the early-type galaxies at $z = 0$. The predicted relation is now in better agreement with observations, much better than both the original Lacey16 and Lacey16+cv models, although the scatter around the median is still much larger than observed. This improvement is mainly due to the reduction in sizes of the faint early-type galaxies. This can again be understood by considering the reduction of cooling in relatively low mass halos when including the reincorporated gas.

4.3.3.2 Retuned Lacey16 model

As already discussed, we retune some of the parameters in the version of the Lacey16 model incorporating the new cooling model, in order to match better the $z = 0$ $b_{\rm J}$ and K-band field LFs at $z = 0$, using the early-type galaxy fraction at different luminosities as a secondary constraint (see §4.2.3 in Lacey et al. (2016)). At the same time, we try to retain the improvement in the size-luminosity correlation of the early-type galaxies at $z = 0$. The retuned parameters are summarized in Table 4.1.

To match the field LF observations, the major problem needing to be solved is the excess at the bright end. As discussed in §4.3.3.1, this is due to the ineffectiveness of AGN feedback, which is a combined effect of enhanced cooling and the less efficient black hole growth induced by the suppression of the disk instabilities. One direct solution would be to increase the number of disk instabilities by raising the stability threshold. However, the faint early-type galaxies are mainly produced by disk instabilities, and raising the stability threshold would let disks with higher

Table 4.1: Retuned parameters and their original values in the Lacey16 model.

parameter	Lacey16	retuned	description
$\alpha_{\rm cool}$	0.8	1.4	threshold of the ratio of the free-fall/cooling time scale
$\gamma_{\rm SN}$	3.2	2.8	slope of the SN feedback power-law scaling
$f_{\rm df}$	1.0	0.7	normalization of the dynamical friction sinking time scale

specific angular momentum, and thus larger sizes, be turned into spheriods. This would increase the median size of the faint early-type galaxies, and thus spoil the improvement achieved by using the new cooling model. Therefore other ways of enhancing the AGN feedback effect should be considered first. The effect of the AGN feedback can also be increased by turning on AGN feedback earlier. This can be done by increasing the parameter $\alpha_{\rm cool}$, which sets the threshold of the ratio of the free-fall timescale over the cooling timescale (both evaluated at $r = r_{\rm cool}$) at which AGN feedback turns on (for more details see Appendix D). Here we increase $\alpha_{\rm cool}$ from 0.8 to 1.4.

We also slightly reduce the SN feedback strength in low-mass galaxies to improve the predictions for the faint ends of the field LFs. In GALFORM, the strength of the SN feedback scales with galaxy circular velocity $V_{\rm c}$ as a power-law, $V_{\rm c}^{-\gamma_{\rm SN}}$. We reduce $\gamma_{\rm SN}$ from 3.2 to 2.8.

We also slightly reduce the galaxy merger timescale to improve the predicted early-type fraction for bright galaxies. The original Lacey16 model and all the variations considered here adopt the fitting formula from Jiang et al. (2008) to calculate the galaxy merger timescale due to dynamical friction. We modify this by introducing an extra factor $f_{\rm df}$ in the formula for the galaxy merger timescale (Eq(14) in Lacey et al. (2016)). The original fit in Jiang et al. (2008) implies $f_{\rm df} = 1$, and this value was effectively assumed in Lacey et al. (2016). Here we reduce $f_{\rm df}$ to

0.7, which is still roughly consistent with the simulation data in Jiang et al. (2008) (see their Fig. 10). The most important effect of this is to increase the number of major mergers.

After this limited retuning of parameters, the predicted field LFs agree with observations again, as shown in the bottom row of Fig. 4.5. The improvements in the predicted galaxy sizes are largely retained.

4.4 summary

We present a new, more physical model for gas cooling and accretion in halos for use in semi-analytical models of galaxy formation. Compared with previous cooling models, this new cooling model adopts a more detailed consideration of the contraction of the hot gas halo induced by cooling, an improved calculation of the cooling history of the hot gas and a more detailed and more self-consistent calculation of the angular momentum gas of the gas that cools and accretes onto the central galaxy.

Compared to the the cooling models previously used in GALFORM, the improved calculation of the cooling history and the detailed modeling of the contraction of the hot gas halo significantly increase the mass that cools in massive halos. Some previous works (e.g. De Lucia et al., 2010; Monaco et al., 2014) argued that the GALFORM cooling model tends to underestimate the gas mass that cools in massive halos, and proposed using a more centrally concentrated hot gas density profile (e.g. singular isothermal profile) to solve this problem. However, in the new cooling model, the predicted cooled mass becomes closer to the predictions of the L-GALAXIES and MORGANA cooling models.

When comparing predictions between different cooling models for the angular momentum of the cooled down gas, even larger differences are seen than for the mass. The new cooling model tends to predict higher specific angular momentum of the cooled down gas than the cooling models previously used in GALFORM. On the other hand, the predictions of the new cooling model for the angular momentum are generally smaller than those from the L-GALAXIES and MORGANA cooling models. This is mainly because different models adopt different distributions for the specific

angular momentum of the hot gas, and different treatments of the effects of cooling on these distributions.

In the full GALFORM model with all other processes such as star formation, supernova (SN) feedback and AGN feedback included, the new cooling model tends to predict less gas cooling in lower mass halos ($M_{\rm halo} \lesssim 10^{12}\,{\rm M}_{\odot}$) than the cooling model previously used in GALFORM, because it models more correctly the effects of the gas that is reincorporated into the hot gas halo after being ejected by SN feedback. This effect improves the predicted size-luminosity relation of both early-type and late-type galaxies relative to observations. However, the improvement in the sizes of late-type galaxy is very small, which indicates that other physical effects may be involved in explaining the discrepancy with observations.

Having understood the behavior of the new cooling model, and having compared the new cooling model with other cooling models, the next step is to compare the predictions of the new cooling model with the results from hydrodynamical simulations. We leave this comparison for future work.

Chapter 5

A comparison between semi-analytical gas cooling models and hydrodynamical simulations

5.1 Introduction

Comparing with more detailed hydrodynamical simulations is a widely used method to assess the accuracy of the highly simplified semi-analytical (SA) gas cooling models. This method has been used in many previous works (e.g. Benson et al., 2001; Yoshida et al., 2002; Helly et al., 2003b; Viola et al., 2008; Lu et al., 2011; Monaco et al., 2014). Here we are going to follow this method to compare the new cooling model constructed in Chapter 4 with hydrodynamical simulations. Along with this model, we also compare other SA models, including the GFC1 (Bower et al., 2006) and GFC2 (Benson & Bower, 2010) cooling model in GALFORM, and the cooling models in L-GALAXIES (e.g. Springel et al., 2001; Croton et al., 2006; Henriques et al., 2015) and MORGANA (Monaco et al., 2007; Viola et al., 2008).

Previous works mainly focused on the mass cooling rate, but the cooling process delivers both mass and angular momentum to central galaxies. Angular momentum is crucial for galaxy formation, because it directly affects galaxy sizes and further indirectly affects star formation and feedback through these sizes. We therefore compare both mass and angular momentum delivered to the central galaxy by cooling

predicted respectively by the SA models and the simulation. Also, the hydrodynamical simulations in previous works mostly used the smoothed particle hydrodynamics (SPH) method. This method is widely used in studying galaxy formation, but some recent works (e.g. Nelson et al., 2013) suggest that there are some artificial effects in SPH that may affect gas cooling rate. Therefore, it would be beneficial if the SA models could also be compared with simulations done with other methods. In this work, we ran the hydrodynamical simulation using the grid-based moving mesh code AREPO. This method has been argued to largely avoid the artificial effects associated with SPH (e.g. Springel, 2010).

The simulation results can be complex to interpret/analyze due to their rich detail, while the outline picture provided by the highly simplified SA cooling models can help to highlight the important physics. Thus this comparison also provides some insights into the cooling physics. In this work, we consider several physical aspects, including cooling in the fast cooling regime, in which cooling is faster than the gravitational infall, cooling from a quasi-hydrostatic hot gas halo and the effects of halo mergers on cooling. All three aspects have been discussed in some previous works. Cooling in the fast cooling regime is closely related to the cold accretion scenario (e.g. Birnboim & Dekel, 2003; Kereš et al., 2005; Nelson et al., 2016), in which the gas is delivered to the central galaxy through cold filaments, which very different from the spherical gas distribution assumed in the SA models. Here we mainly consider how strongly this cold accretion picture can affect the mass cooling rate. The cooling from hot gas halos was previously studied in Viola et al. (2008), but that was mainly for static spherical halos simulated by using the SPH method, while here we directly study this for halos in a cosmological simulation done with a grid-based method. Monaco et al. (2014) studied the effects of halo major mergers on cooling, based on SPH simulations. Here we study these effects in more detail, and based on a grid-based simulation.

This chapter is organized as follows. A brief introduction to the moving mesh method and AREPO code is given in §5.2. Also given in this section are details of the simulation runs, the measurement of accreted mass and angular momentum from simulation results and some further details of the SA cooling models. The results

and associated discussions are given in §5.3, and a brief summary is given in §5.4.

5.2 Method

5.2.1 Moving Mesh Code AREPO for Hydrodynamics

AREPO is a finite volume grid-based hydrodynamical code (Springel, 2010). The grid is generated by a Voronoi tessellation of space, while this tessellation is induced by a set of grid generation particles. These particles are allowed to have arbitrary motions, but usually they are set to largely follow the motion of the fluid itself. Then the fluid fluxes across the boundaries of each cell in the grid are calculated through the exact 1D Riemann solution, and these fluxes are used to update the whole fluid field.

This method can be viewed as a combination of the adaptive mesh refinement (AMR) approach and smooth particle hydrodynamics (SPH). Compared to the more traditional grid-based method AMR, allowing the grid to move with the fluid has several advantages. Firstly, this can largely avoid large fluid velocities relative to the grid. Large fluid velocities lead to the kinetic energy dominating the total energy budget of the flow, leading to a very inaccurate estimation of the internal energy and the thermal state of the fluid, which is crucial for calculating gas cooling. In cosmic structure formation, gas flows with large relative velocities are common, and this means that large velocities relative to the grid are inevitable for a static grid in the AMR method. Secondly, the moving mesh provides a continuous adjustment of the resolution, instead of the discrete jump of resolution in the mesh refinement of AMR. The latter artificially suppresses the structure growth under gravity (e.g. O'Shea et al., 2005; Heitmann et al., 2008).

The SPH method is commonly used in cosmological hydrodynamical simulations. This method is particle-based and quasi-Lagrangian, which also gives it a continuously adaptive resolution. Continuous fluid quantities, such as density, are derived though smoothing over nearby particles. This smoothing introduces relatively large artificial dissipation and diffusion, which can broaden shock fronts, leading to less efficient shock heating, and damp turbulent motions, leading to artificial heating.

As shown in Nelson et al. (2013), these effects bias the cooling calculation. The grid-based flux calculation in AREPO can largely avoid these effects. Some further recipes can be added to SPH to largely remove those artificial effects in some specific situations (e.g. Beck et al., 2016), but whether these recipes are universally workable or have any side effects is unclear.

In summary, the moving mesh code AREPO is an ideal tool for the study of gas cooling in the context of cosmic structure formation.

5.2.2 Simulations

We assume the ΛCDM cosmology with cosmological parameters based on the WMAP-7 data (Komatsu et al., 2011): $\Omega_{\rm m0} = 0.272$, $\Omega_{\Lambda 0} = 0.728$, $\Omega_{\rm b0} = 0.0455$ and $H_0 = 70.4\,{\rm km\,s^{-1}Mpc^{-1}}$, and an initial power spectrum with slope $n_{\rm s} = 0.967$ and normalization $\sigma_8 = 0.810$. The simulations are in a cube with co-moving length 25Mpc and periodic boundary conditions. The initial conditions are generated by the code N-GENIC (Springel et al., 2005), which uses the Zel'dovich approximation to displace particles from a regular cubic grid. we use 376^3 dark matter particles and initially the same number of gas cells. The dark matter particle mass is $9.2\times10^6\,{\rm M_\odot}$. The gravitational softening scale is 0.98 co-moving kpc. There are 128 output times evenly spaced in $\log(1+z)$, from $z = 19$ to $z = 0$. The physical time interval between two adjacent outputs is about one fourth of the halo dynamical timescale, $t_{\rm dyn}$, which is defined as $t_{\rm dyn} = r_{\rm vir}/v_{\rm vir}$, with $r_{\rm vir}$ and $v_{\rm vir}$ respectively the virial radius and velocity of a halo.

We run a simulation with dark matter only to construct merger trees for the semi-analytical (SA) gas cooling models. We also run two simulations with gas. One of these is without cooling or other physical process such as star formation and feedback, and is used to investigate the hot gas density distributions used in the SA models. For the simulation with gas cooling, we adopt the cooling functions of primordial gas based on Katz et al. (1996), and do not include the cooling due to inverse compton scattering on the CMB, which is less important than other cooling mechanisms considered here. There is no UV heating background, but we impose a cooling temperature floor to prevent gas cooling in very small dark matter halos.

Specifically, a gas cell can cool only if its temperature $T_{\rm gas}$ satisfies

$$T_{\rm gas} > T_{\rm cool,lim} = 3.5 \times 10^4 \times [\Omega_{\rm m0}(1+z)^3 + \Omega_{\Lambda 0}]^{1/3}\,{\rm K}, \tag{5.2.1}$$

where z is the redshift of the gas cell, and $T_{\rm cool,lim}$ roughly corresponds to the virial temperature of a halo with $M_{\rm vir} = 2 \times 10^{10}\,{\rm M}_\odot$, which in our simulations is resolved with 2000 particles. According to Monaco et al. (2014), this resolution is high enough for proper cooling calculations. This simulation does not include any feedback or metal enrichment process.

The cooled down gas would accumulate in the halo centre and reach high density. This cold and dense gas has a very short dynamical timescale, leading to large computational cost, but because this gas has already been accreted by the central galaxy, its further fate is irrelevant to the gas cooling model. Thus, we turn this gas into collisionless stellar particles to save computation time. As in Monaco et al. (2014), the gas is turned into stars when its density is higher than $\delta_{\rm sfr,lim}\bar{\rho}_{\rm gas}$ and its temperature is lower than $T_{\rm sfr,lim}$, where $\bar{\rho}_{\rm gas} = \Omega_{\rm b}(z)\rho_{\rm crit}(z)$ is the mean gas density, with $\Omega_{\rm b}(z)$ and $\rho_{\rm crit}(z)$ the baryon fraction and critical density at redshift z respectively, and $\delta_{\rm sfr,lim}$ and $T_{\rm sfr,lim}$ two parameters. We adopt $\delta_{\rm sfr,lim} = 10^4$ and $T_{\rm sfr,lim} = \min[10^5\,{\rm K}, T_{\rm cool,lim}]$. Note that here this star formation is not meant to represent a physical process but is just a numerical technique to enhance simulation speed.

The structures formed are first identified though the friends-of-friends (FOF) approach, and then each FOF group is further split into subgroups using SUBFIND (Springel et al., 2001).

5.2.3 Merger Trees

The merger trees are build using the Dhalo algorithm (Helly et al., 2003a; Jiang et al., 2014). This method is based on the subgroups identified by SUBFIND. It links the subgroups at different snapshots by cross matching their most bound particles. Using this, the merger trees for these subgroups are generated. Then these subgroups are grouped into Dhalos by examining their separations. Namely, if one subgroup lies within twice of the half mass radius of another subgroup, then they are in the

same Dhalo. Thus a structure and the substructures it contains are assembled into a single Dhalo, while the structures enclosed in a single FOF group through artificial low density bridges are separated into different Dhalos. Once a subgroup belongs to a Dhalo, it would never leave this Dhalo. This ensures that a subhalo temporarily leaving its host halo during a merger is treated as being a subhalo since its first infall. Finally, the subgroup merger trees are combined to derive the Dhalo merger trees.

The Dhalo merger trees of the dark matter only simulation are built for calculating SA models, while the merger trees of the hydrodynamical simulation are built to extract the gas cooling histories from this simulation. The merger trees of these two simulations are linked by cross matching the 50 most bound dark matter particles of the base halos at $z = 0$. Two linked merger trees are treated as being of the same halo in different simulations.

Unless otherwise specified, all the halo masses used in this work are the Dhalo mass provided by Dhalo merger trees.

5.2.4 Measuring the Mass and Angular Momentum of the Cooled Down Gas

The cooled down gas in a halo falls towards the minimum of its gravitational potential well, and is accreted by the galaxy there. According to the algorithm of SUBFIND, this potential minimum should be associated to the most massive subgroup in a Dhalo. Thus the cooled down gas should also be found in the region around the potential minimum of this subgroup. We identify this region as the central galaxy in the Dhalo. Further, as mentioned in §5.2.2, the cool gas accreted by galaxies is quickly turned into stars, so finally the cooled down gas is represented by the stars in the central region of the most massive subgroup of a given Dhalo. Here the central region is defined as $r \leq 0.2 r_{200}$. For simplicity, here this r_{200} is calculated based on the Dhalo mass. We checked that it is close to that calculated based on M_{200}. We also checked that our measurements are reasonably stable for different choices of this radius aperture.

The above mentioned selection defines the stars in the central galaxy of a given

Dhalo. However, the gas cooled over the history of this Dhalo along the major branch (formed by the most massive progenitor Dhalos) of its merger tree only forms part of the stars, the other part formed in other galaxies and is delivered to the central galaxy through galaxy mergers. The stars from these two channels can be separated based on two features of galaxy mergers. Firstly, the time from the first infall of a satellite galaxy to its final merger with the central galaxy is typically longer than one halo dynamical timescale. Secondly, the gas cooling after the infall of a satellite halo would not last for long time, so when a satellite has nearly merged with the central galaxy, it should not contain any newly formed stars. The second point is a plausible assumption, and it should be further tested in future works.

Motivated by these two observations, after we pick out the stars in the central galaxy of a given Dhalo at the i-th output time $t_{\rm i}$, we then go back to this halo's main progenitor (defined as the most massive progenitor Dhalo) at the $(i-1)$-th output time $t_{\rm i-1}$, and remove all the selected stars that also exist at $t_{\rm i-1}$. This should only leave the stars formed by the gas cooled down in the given Dhalo between $t_{\rm i-1}$ and $t_{\rm i}$. The reason for this is as follows. The stars in the central galaxy at $t_{\rm i}$ can be divided into three categories, namely the stars in the main progenitor of this central galaxy at $t_{\rm i-1}$, the stars delivered by the merging satellites during $(t_{\rm i-1}, t_{\rm i}]$ and the stars newly formed in the central galaxy between $t_{\rm i-1}$ and $t_{\rm i}$. Because the time interval corresponding to $(t_{\rm i-1}, t_{\rm i}]$ is shorter than halo dynamical timescale, at $t_{\rm i-1}$, these merging satellites should be in the current halo's main progenitor halo, so the above method can cover all the merging satellites, and it removes all stars formed before $t_{\rm i-1}$ in either the main progenitor of the central galaxy or these merging satellites, leaving only new stars formed during $(t_{\rm i-1}, t_{\rm i}]$. Also, because of this relatively short time interval, as assumed above, these merging satellites should have stopped gas cooling by $t_{\rm i-1}$ and thus do not form any star during $(t_{\rm i-1}, t_{\rm i}]$. So, these selected new stars should all be formed in the central galaxy, and this can only be induced by cooling in the given Dhalo.

With these stars selected, the mass of gas cooled down within $(t_{\rm i-1}, t_{\rm i}]$, $\Delta M_{\rm cool,i}$, is measured as

$$\Delta M_{\rm cool,i} = \sum_{j=1}^{N} m_{\rm star,j}, \tag{5.2.2}$$

where the index j labels the selected stellar particles, N is their total number, and $m_{\rm star,j}$ is the mass of the j-th stellar particle. Then the gas cooling rate at $t_{\rm i}$ is estimated as

$$\dot{M}_{\rm cool}(t_{\rm i}) = \frac{\Delta M_{\rm cool,i}}{t_{\rm i} - t_{\rm i-1}}. \tag{5.2.3}$$

The cumulative cooled down mass, $M_{\rm cool}(< t_{\rm i})$, is calculated as

$$M_{\rm cool}(< t_{\rm i}) = \sum_{j=i_{\rm start}}^{i} \Delta M_{\rm cool,j}, \tag{5.2.4}$$

where the summation is along the major branch of a merger tree, (namely it only includes cooling in the main progenitors), and $i_{\rm start}$ is the index of the earliest output time this branch can reach.

The measurement of the angular momentum of the cool gas is more complex, because although galaxy mergers should not affect the mass of this gas, they can change its dynamical state, thus changing its angular momentum. To largely remove this effect, we estimate the angular momentum of the gas cooled down within $(t_{\rm i-1}, t_{\rm i}]$ as the angular momentum change of the central galaxy system, including the central galaxy and all the satellites merging with it within this time interval.

Specifically, for a given Dhalo at output time $t_{\rm i}$, we pick out all stellar particles associated with the most massive subgroup and within $0.2r_{200}$ from the subgroup centre, and calculate its total angular momentum $\boldsymbol{J}_{\rm tot,i}$. Then we pick out the part of these stars that exist at one previous output time $t_{\rm i-1}$, and calculate the corresponding total angular momentum at that time, $\boldsymbol{J}_{\rm tot,i-1}$. Note that at this earlier time, some of these star particles are in the satellites merging with the central galaxy. According to the above galaxy merger timescale argument, we only need to search the main progenitor of the given Dhalo at $t_{\rm i-1}$ to select the particles for $\boldsymbol{J}_{\rm tot,i-1}$. With these two quantities, the angular momentum of the cool gas is calculated as

$$\Delta\boldsymbol{J}_{\rm cool,i} = \boldsymbol{J}_{\rm tot,i} - \boldsymbol{J}_{\rm tot,i-1}. \tag{5.2.5}$$

The reference points of the angular momentum $\boldsymbol{J}_{\rm tot,i}$ and $\boldsymbol{J}_{\rm tot,i-1}$ are the centre of mass of the corresponding selected stellar particles, and the velocities for calculating these angular momenta are relative to the centre of mass. This choice ensures the $\Delta\boldsymbol{J}_{\rm cool,i}$ is associated to the internal motion of the galaxy.

Here we assume that $\Delta \boldsymbol{J}_{\rm cool,i}$ is dominated by gas cooling, while the effects from dark matters can be ignored. This assumption is plausible, but it needs to be tested in the future work.

Note that the directions of angular momenta of halo gas (hot or cold) and galaxies may also slowly evolve with the growth of the dark matter halo. This evolution is included in the vector $\Delta \boldsymbol{J}_{\rm cool,i}$. On the other hand, all SA models considered here attach the angular momenta of the halo gas and galaxies to halo spin, and treat them as scalers. This treatment implicitly assume that the angular momenta of dark matter halo, halo gas and galaxies are all aligned, while the direction evolution in hydrodynamical simulations could go beyond this simple assumption, leading to direction differences between these angular momenta. To further illustrate this effect, as well as to examine the predictions of angular momenta delivered by gas cooling, we introduce the following two quantities in our comparison.

The first quantity is

$$J_{\rm cool}(< t_{\rm i}) = \left| \sum_{j=i_{\rm start}}^{i} \Delta \boldsymbol{J}_{\rm cool,j} \right|. \tag{5.2.6}$$

The vector summation of $\Delta \boldsymbol{J}_{\rm cool,i}$ gives the total angular momentum of a central galaxy with the effects from galaxy mergers largely removed, and $J_{\rm cool}(< t_{\rm i})$ is its magnitude at $t_{\rm i}$. This quantity largely includes the direction evolution described above.

The second quantity is

$$\tilde{J}_{\rm cool}(< t_{\rm i}) = \sum_{j=i_{\rm start}}^{i} |\Delta \boldsymbol{J}_{\rm cool,j}|. \tag{5.2.7}$$

When only adding the magnitude of each $\Delta \boldsymbol{J}_{\rm cool,j}$, the direction differences between them are ignored, but note that $\tilde{J}_{\rm cool}(< t_{\rm i})$ still reflects the magnitude of angular momentum delivered by gas cooling.

As assumed above, the angular momentum delivered by gas cooling dominates $\Delta \boldsymbol{J}_{\rm cool,i}$, and if the spins of the halo gas and central galaxy are more or less aligned over time, then one should have $\tilde{J}_{\rm cool}(< t_{\rm i}) \approx J_{\rm cool}(< t_{\rm i})$. Otherwise, $J_{\rm cool}(< t_{\rm i})$ should be obviously lower than $\tilde{J}_{\rm cool}(< t_{\rm i})$.

Corresponding to $J_{\rm cool}(< t_{\rm i})$ and $\tilde{J}_{\rm cool}(< t_{\rm i})$, there are two specific angular momenta, namely

$$j_{\rm cool}(t_{\rm i}) = \frac{J_{\rm cool}(< t_{\rm i})}{M_{\rm cool}(< t_{\rm i})}, \tag{5.2.8}$$

and

$$\tilde{j}_{\rm cool}(t_{\rm i}) = \frac{\tilde{J}_{\rm cool}(< t_{\rm i})}{M_{\rm cool}(< t_{\rm i})}. \tag{5.2.9}$$

5.2.5 Semi-analytical Calculation of Gas Cooling

The SA gas cooling models considered here are from several major SA galaxy formation models, namely GALFORM (e.g. Cole et al., 2000; Baugh et al., 2005; Bower et al., 2006; Lacey et al., 2016), MORGANA (e.g. Monaco et al., 2007; Viola et al., 2008) and L-GALAXIES (e.g. Springel et al., 2001; Croton et al., 2006; Henriques et al., 2015). These are described in Chapter 4. In our comparison, these models use the same cooling functions as in the hydrodynamical simulation. Corresponding to the cooling temperature floor in the simulation, here the gas is only allowed to cool in halos more massive then $2 \times 10^{10}\,{\rm M}_{\odot}$ in all SA models.

In this work, all of the models are run on the same Dhalo merger trees extracted from the dark matter only simulation. This simulation also provides the halo spin, which is required for calculating the angular momentum of the cooled down gas. Because here the cooling occurs only in relatively well resolved halos, the measured halo spin is reliable (Bett et al., 2007).

The cooling calculation also requires the knowledge of the hot gas density and temperature profiles. The L-GALAXIES cooling model assumes a singular isothermal distribution for the hot gas density profile. All of the other models considered here assume a cored profile. The cooling models in GALFORM adopt the so-called β-distribution,

$$\rho_{\rm hot}(r) \propto \frac{1}{r^2 + r_{\rm core}^2}, \tag{5.2.10}$$

where $\rho_{\rm hot}(r)$ is the hot gas density as a function of radius r, and $r_{\rm core}$, the core radius, is a parameter, while the normalization is fixed by the total hot gas mass when $r_{\rm core}$ is known. The MORGANA cooling model adopts a more complex profile, which is derived by requiring hydrostatic equilibrium of the hot gas for a polytropic

pressure-density relation within a potential created by an NFW dark matter halo. For simplicity, here we instead use the β-distribution for the MORGANA cooling model. The cooling models in GALFORM and L-GALAXIES assume that the gas temperature is independent of radius and equals to $T_{\rm vir}$, while the MORGANA cooling model adopts a more complex temperature profile also derived from the hydrostatic requirement. Here, for simplicity, we adopt a constant temperature $T_{\rm vir}$ for the hot gas in all three models. As shown below, this change in the temperature profile should not strongly affect the predictions of the MORGANA cooling model, because the temperature profile suggested by the hydrodynamical simulation is very shallow and close to $T_{\rm vir}$.

To estimate $r_{\rm core}$, we use the simulation without gas cooling introduced in §5.2.2. We extract the spherically averaged density distribution of the most massive subgroup in each of the FOF groups with $M_{200} \geq 10^{12}\,{\rm M}_{\odot}$, and then fit these distributions with the β-distribution, which has $r_{\rm core}$ as the only parameter, while its normalization is fixed by $r_{\rm core}$ and the total gas mass measured from the simulation. This fitting is done for different redshifts, from 0 to 1.5. At $z = 1.5$, there are 17 FOF groups satisfying the selection condition, and this number gradually increases to 35 FOF groups at $z = 0$.

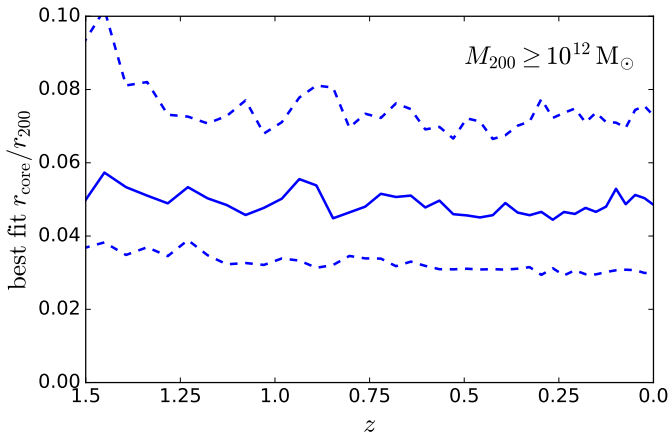

Figure 5.1: Best fit $r_{\rm core}$ of the most massive subgroups in the FOF groups with $M_{200} \geq 10^{12}\,{\rm M}_{\odot}$, for $z = 0 - 1.5$. The blue solid line is the median of the best fit values, while the blue dashed line indicate the $10 - 90\%$ range.

Fig. 5.1 shows the median and $10 - 90\%$ range of the best fit $r_{\rm core}$ for different

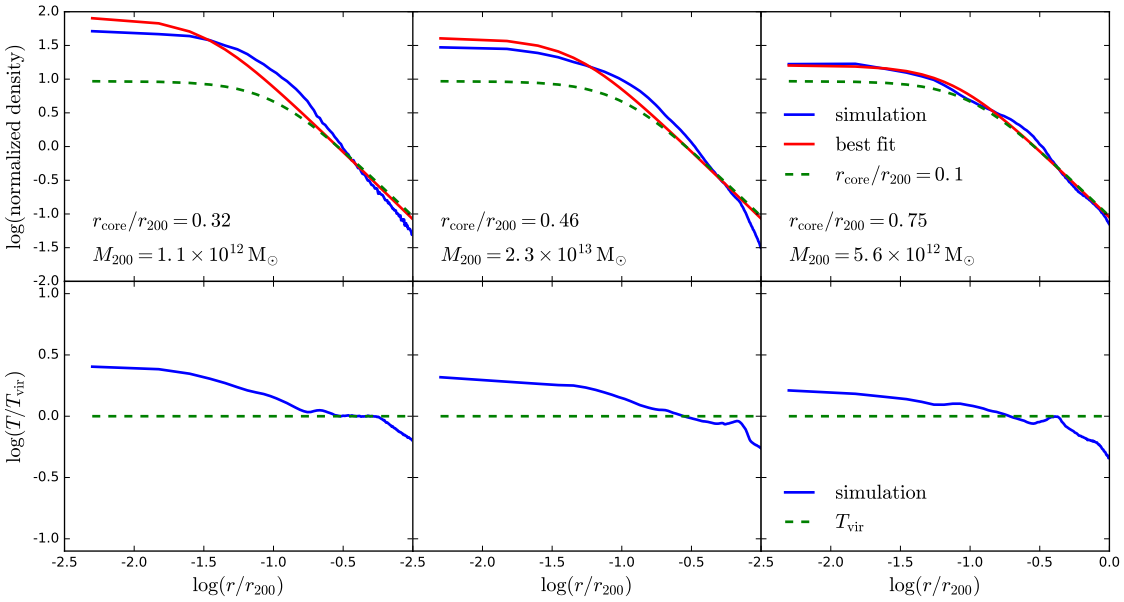

Figure 5.2: The spherical averaged density and temperature profiles of three example halos at $z = 0$. Here the density profile is normalized so that the total mass within r_{200} is 1. The best fit density profile as well as the profile with $r_{\rm core}/r_{200} = 0.1$ are also shown.

redshifts. It shows that $r_{\rm core}/r_{200}$ is very stable over redshift, with the median about 0.05. However, this value should not be directly used in SA cooling models. There are two reasons for this.

Firstly, although the gas temperature profile is very shallow, it is not exactly constant. This can be seen from the three example halos shown in Fig. 5.2. The temperature in the outer region ($r \geq 0.1 r_{200}$) is close to $T_{\rm vir}$, but the central temperature is about a factor of two to three higher than $T_{\rm vir}$. Thus the constant temperature adopted by SA models together with the best fit density profile would lead to an overestimation of cooling in the SA models. In principle, we should adopt a more complex temperature profile in the SA models to solve this problem, but there is a simpler way to reduce this overestimation, which is, by slightly increasing $r_{\rm core}$ to reduce the central density. As can be seen in Fig. 5.2, as far as $r_{\rm core} \ll r_{200}$, this change only significantly affects the density in the central region, where the temperature is higher than $T_{\rm vir}$. Adopting $r_{\rm core}/r_{200} = 0.1$ would lower the central density by a factor of two, which cancels the effect of underestimating central

temperature. We leave the application of non-constant temperature profile in SA models to future work.

Secondly, here the SA models adopt the Dhalo mass as the halo mass. The Dhalo mass is the sum of the subgroup masses in a given halo, and is very close to the total mass of the given nonlinear structure. According to the spherical collapse model, the total mass corresponds to the virial mass $M_{\rm vir}$, with associated radius $r_{\rm vir}$ and mean density $\Delta'_{\rm vir}\rho_{\rm crit}$, where $\Delta'_{\rm vir}$ is the spherical overdensity. On the other hand r_{200} is associated with the mean density $200\rho_{\rm crit}$ and mass M_{200}.

These two radius are related as $r_{200}/r_{\rm vir} = (M_{200}/M_{\rm vir})^{1/3}(\Delta'_{\rm vir}/200)^{1/3}$. We checked that the maximum difference between the Dhalo mass and M_{200} is about 20%, so $r_{200}/r_{\rm vir}$ is dominated by the second term, and thus for the SA models we adopt

$$\frac{r_{\rm core}}{r_{\rm vir}} = \frac{r_{\rm core}}{r_{200}}\frac{r_{200}}{r_{\rm vir}} \approx 0.1\left(\frac{\Delta'_{\rm vir}}{200}\right)^{1/3} \tag{5.2.11}$$

and we evaluate $\Delta'_{\rm vir}$ through the fitting formula (e.g. Eke et al., 1996; Bryan & Norman, 1998):

$$\Delta'_{\rm vir}(z) = 18\pi^2 + 82[\Omega_{\rm m}(z) - 1] - 39[\Omega_{\rm m}(z) - 1]^2, \tag{5.2.12}$$

where $\Omega_{\rm m}(z)$ is the matter density parameter at redshift z. $\Delta'_{\rm vir}$ deviates significantly from 200 only at $z < 1$, so the difference between $r_{\rm vir}$ and r_{200} only appears at very low redshift. At $z = 0$, this difference reaches its maximum, with $r_{\rm vir}$ about 30% larger than r_{200}.

5.3 Results

5.3.1 Cooling Physics

SA models always employ a very simple picture for gas cooling, while hydrodynamical simulations contain more complex details. The comparison of the predictions from these two methods can highlight some important details of the physics of cooling. In this subsection we compare simulation predictions with SA models for several individual halos. The simulation prediction is from the hydrodynamical simulation

introduced in §5.2.2, while the SA model used here is the new gas cooling model introduced in Chapter 4.

Our comparison here mainly focuses on the predicted mass cooling rate $\dot{M}_{\rm cool}(t_{\rm i})$, because it is tightly related to the processing of cooling. Some other quantities are also shown when relevant.

5.3.1.1 Fast Cooling Regime vs. Filamentary Accretion

SA models usually predict that for low mass halos, the gas cools faster than its gravitational infall, and so a part of the halo gas should be cold. This is called the fast cooling regime. This regime ends when a halo reaches a mass around $3\times10^{11}\,{\rm M}_{\odot}$, and afterwards a hot gas halo gradually becomes dominant.

On the simulation side, many previous works have argued for a more complex picture (e.g. Kereš et al., 2005; Dekel & Birnboim, 2006). In this picture, the gas is delivered to dark matter halos through filaments rather than being spherically accreted, and in low mass halos, these cold filaments can reach all the way to the central galaxy, and so never build a spherical cold gas halo. Only at later times, when they become wider and less dense, and the halo has a higher virial temperature, do these filaments join the hot gas halo.

These two pictures are very different for low mass halos ($M_{\rm halo} \lesssim 3\times10^{11}\,{\rm M}_{\odot}$). The importance of their effects on galaxy properties depends on the cooled mass and angular momentum that they predict. Fig. 5.3 shows the cooling histories of a low mass halo predicted by the new SA model and by the hydrodynamical simulation. This halo has mass $2.4\times10^{11}\,{\rm M}_{\odot}$ at $z=0$, and so is close to the end of fast cooling regime only at very late times. The predicted cooling histories include the mass cooling rate, accumulated mass and the specific angular momentum $\tilde{j}_{\rm cool}$ of the cool gas. Here we choose $\tilde{j}_{\rm cool}$ to focus on the gas cooling. The effects related to angular momentum direction are discussed later.

This figure shows that the predictions of the mass cooling rate from the new SA model and from the simulation are generally in good agreement for this halo. The drop in mass cooling rates at $z\sim0$ seen in the simulation results is an artificial effect, and more details about it will be discussed later. The predicted $\tilde{j}_{\rm cool}$ from

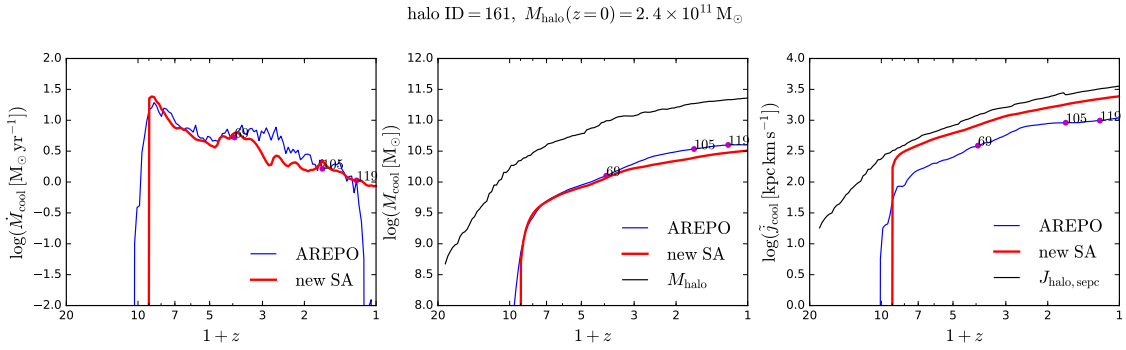

Figure 5.3: *left*: The predicted mass cooling rate. *middle*: The predicted cumulative cooled down mass. The growth of halo mass is also shown for reference. *right*: The predicted specific angular momentum $\tilde{j}_{\rm cool}$ of the cooled down gas. The evolution of the halo specific momentum is also shown as reference. In all three panels, the magenta points label the snapshots where further details are shown, and the associated numbers are the snapshot ids.

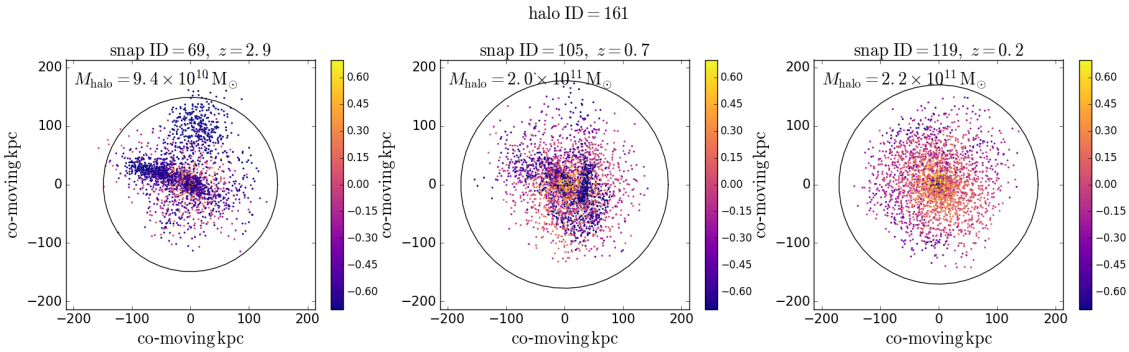

Figure 5.4: The gas distributions of halo 161 in selected snapshots. In each panel, the dots represent the gas cells in the simulation, and their colours indicate gas temperature. The colours are plotted on a scale $\log(T/T_{\rm vir})$, in the range $[\log(1/5), \log(5)]$. The black circle indicates $r_{\rm vir}$, and the black cross shows the halo centre. The halo mass is given in each panel, while the snapshot id and corresponding redshift are given in the title of the corresponding panel.

two methods show larger differences, with the prediction of the SA model higher by about a factor 2, but they are still comparable.

To view further details of the gas cooling, we select three snapshots and plot the gas distributions. These selected snapshots are labeled as magenta dots in Fig. 5.3,

and the corresponding gas distributions are shown in Fig. 5.4.

According to Fig. 5.4, at high redshift, $z \sim 3$, the gas is generally cold ($T < T_{\rm vir}$), and its distribution is clearly filamentary. This confirms the findings in previous works (e.g. Kereš et al., 2005; Dekel & Birnboim, 2006). Only at later times, when $z \sim 0.7$, the gas distribution becomes more spherical, and closer to the picture in the SA model. At $z = 0.7$, there is still an obvious halo cold gas component, as expected in the SA model. The gas distribution becomes more spherical because at low redshift, the filaments become very wide, with radius comparable to $r_{\rm vir}$ of the low mass halo, and so the accretion is close to spherical (e.g. Kereš et al., 2005). Then even later, at $z = 0.2$, the hot gas halo begins to appear, which indicates the transition from cold halo gas to hot gaseous halo. This transition happens at the mass $2.2 \times 10^{11}\,{\rm M}_{\odot}$ for this halo, and is close to the SA model prediction, which is around $3 \times 10^{11}\,{\rm M}_{\odot}$ (e.g. Benson & Bower, 2011).

Although the simulation gives gas distributions very different from the SA model at $z > 0.7$, the predicted mass cooling rates are similar. This is because in both the simulation and the SA model, for this case, the gas accretion onto the central galaxy is controlled by the free-fall timescale. In the simulation, the gas is delivered by the cold filaments, which are difficult to heat up, and is expected to fall freely onto the central galaxy under gravity. On the other hand, in the SA model, although it is assumed that the gas accreted onto the dark matter halo is shock heated to build a hot gas halo, in the fast cooling regime the cooling timescale is very short, so the final accretion rate onto the central galaxy is again limited by the gravitational infall timescale.

Low mass halos can also be the high redshift progenitors of the low redshift massive halos. Compared to the case studied above, in which the halo remains low in mass to $z \sim 0$, these progenitors are formed in very different environments, so the gas accretion can be different. Fig. 5.5 shows the cooling history of a massive halo. Here we should focus on the relatively high redshift range (e.g. $z \gtrsim 2$), in which the halo mass is low. This figure shows that at $z > 2$ the predictions from these two methods are in good agreement, though this agreement of the predicted mass cooling rates is not as good as that for the low mass halos studied above.

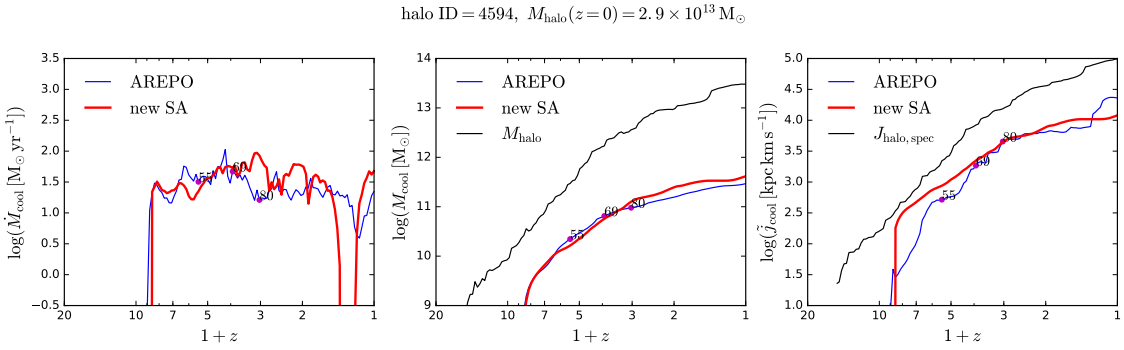

Figure 5.5: The cooling histories predicted by the SA model and simulation for halo with id 4594. The meaning of labels is the same as in Fig. 5.3, and for more information see the caption there.

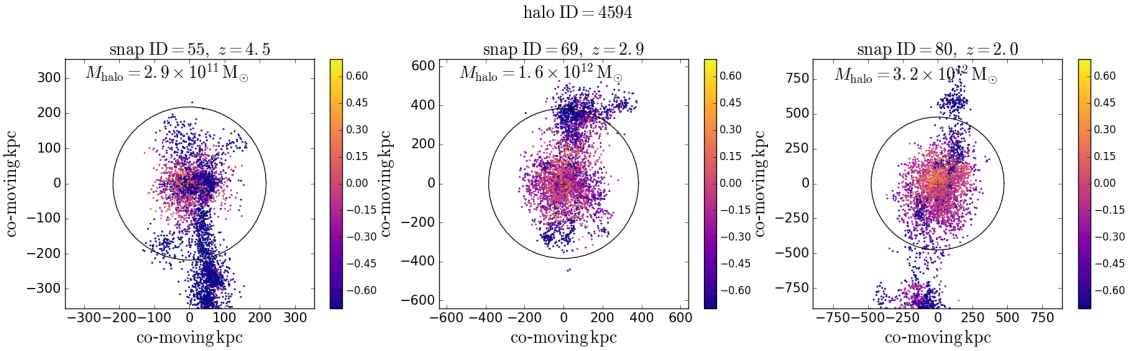

Figure 5.6: The gas distributions of halo 4594 in selected snapshots. The meaning of labels and colour scales is the same as in Fig. 5.4, and for more information see the caption there. This halo has many gas cells, and to keep the plot clear, here only 4000 randomly selected cells are plotted.

At $z > 4$, $M_{\rm halo} \lesssim 3 \times 10^{11}\,{\rm M}_\odot$, and the SA model predicts the halo to be in the fast cooling regime. Between $z = 4$ and $z = 2$, the halo grows from $3 \times 10^{11}\,{\rm M}_\odot$ to about $3 \times 10^{12}\,{\rm M}_\odot$, which is roughly in a transition range from fast cooling to slow cooling. To further investigate the details of cooling at $z > 2$, we select three snapshots and show the corresponding gas distributions in Fig. 5.6. At $z = 4.5$, the gas is obviously cold and filamentary. Then later, this distribution gradually evolves to a more spherical gas halo. At $z = 2.9$, a hot halo has appeared, but its temperature seems to be slightly lower than $T_{\rm vir}$ (purple dots dominate the hot halo), then at $z = 2.0$, the hot gas becomes hotter, with temperature closer to $T_{\rm vir}$.

This transition is different from the simple picture in the SA model, because of the non-spherical filaments, but it seems the SA model still manages to predict the roughly correct cooling history, at least for this specific case.

In summary, filamentary accretion is commonly seen at high redshifts ($z \gtrsim 2$), but in so far as the accretion onto the central galaxy is limited by the free-fall timescale, the simple spherical gas cooling picture in the SA model does not affect much the predictions about cooling. It seems that the SA model also gives roughly the correct cooling histories during the transition from anisotropic filaments to a spherical hot gas halo.

5.3.1.2 Slow cooling Regime

When the halo is massive enough, the SA model predicts the cooling timescale to be longer than the dynamical timescale, and the hot gas in the dark matter halo forms a quasi-hydrostatic hot gaseous halo, and the gas accreted onto the central galaxy cools from this hot halo. This is the so-called slow cooling regime. Because typically the hot gas halo has higher density at smaller radii, the inner part of the halo is easier to cool down. Thus a naive expectation is that the temperature of the gas decreases with radius, the outer part has temperature close to $T_{\rm vir}$ as it experiences less cooling, while the inner part contains gas partially cooling down, with the cold gas (fully cooled down) in the halo centre.

To compare the above picture with hydrodynamical simulations, here we derive the density and temperature maps of the gas in a massive dark matter halo. The one we choose is halo 4594, and its full cooling history is shown in Fig. 5.5. At $z = 0$, this halo has mass about $3 \times 10^{13}\,{\rm M}_{\odot}$, which is massive enough to be in the slow cooling regime predicted by the SA model. Fig. 5.7 shows the projected halo gas temperature and density maps of this halo at $z = 0$. It is clear from this figure that there is a more or less spherical gas halo with temperature close to $T_{\rm vir}$. Thus the expectation from the SA cooling model is confirmed.

However, the temperature maps do not show a decreasing temperature with radius, but instead, the temperature is always close to $T_{\rm vir}$, and is even higher at smaller radii. To further investigate the details in the halo central region, we

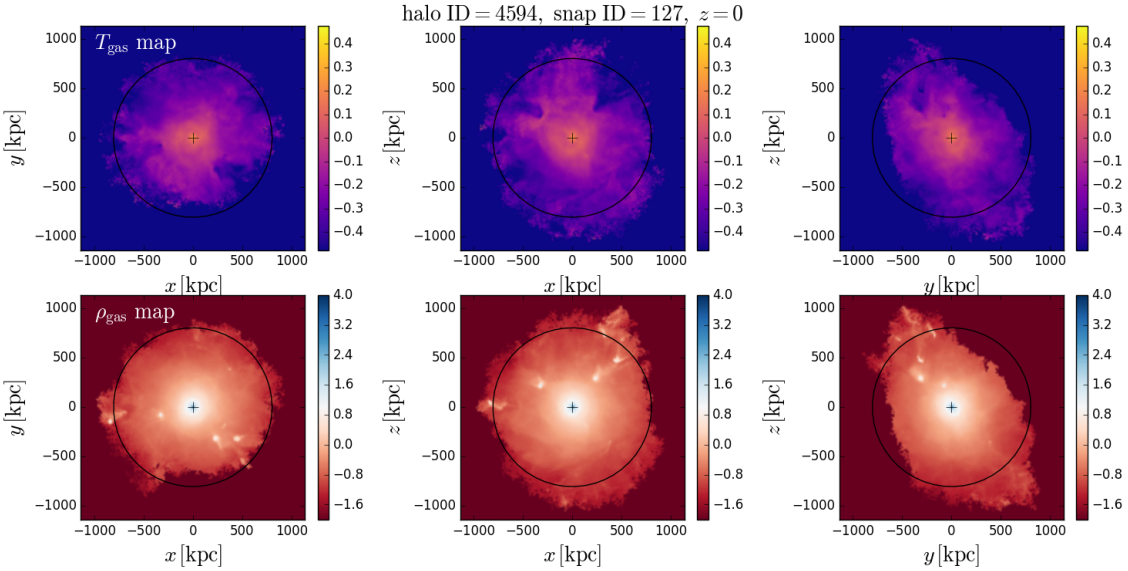

Figure 5.7: The projected halo gas temperature (upper row) and density (lower row) maps for halo 4594 at $z = 0$. Each pixel in the map show the averaged temperature or density of the gas cells along the line of sight. The average is weighted by the gas cell masses. The top row are temperature maps, for three projected directions, while the bottom row are the corresponding density maps. The colours for the temperature maps correspond $\log(T/T_{\rm vir})$, in the range $[\log(1/3), \log(3)]$. The colours for the density maps correspond $\log(\rho/\bar{\rho})$, in the range $[-2, 4]$, and $\bar{\rho} = (\Omega_{\rm b0}/\Omega_{\rm m0})\Delta_{\rm vir}\rho_{\rm crit}$ is the mean baryon density of the halo. These maps show a roughly spherical gaseous halo with temperature about $T_{\rm vir}$. The black circle in each panel shows $r_{\rm vir}$, while the black cross indicates the halo centre.

generated the projected maps of density and temperature for the central region, which are shown in Fig. 5.8. This figure further confirms that there is no gas with temperature significantly lower than $T_{\rm vir}$ in the central region of the gaseous halo. It also shows that in the very central region, the gas becomes very dense while keeping its temperature close to $T_{\rm vir}$, and a disky gas structure forms, with density 10^4 times higher than the mean halo baryon density.

Since there are newly formed stars in this time step, there must be gas cooling, but the temperature maps suggest that the gas keeps a roughly constant temperature during cooling. This means there are heating sources. Because in the current simulation there is no feedback process, the only possible heating energy source is

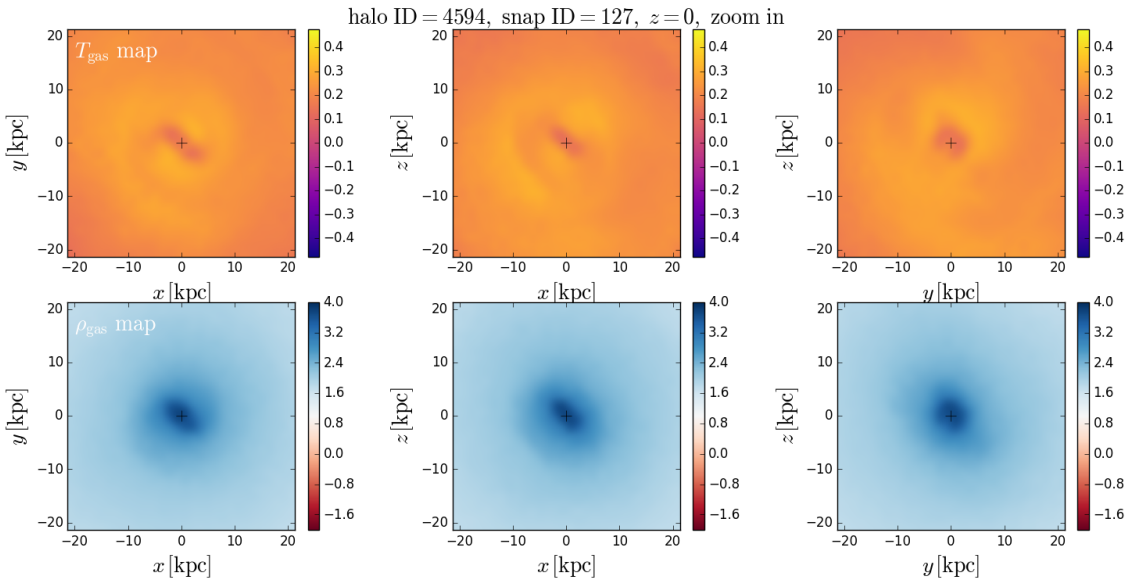

Figure 5.8: The projected halo gas temperature (upper row) and density (lower row) maps for the central region halo 4594 at $z = 0$. The meanings of maps and symbols are the same as Fig. 5.7, and for more information see its caption.

the gravitational potential energy. When the gas in the halo centre finally cools down, it does not provide pressure support to the hot gas halo, and this causes the latter to contract towards the halo centre. During this contraction, gravity does positive work on every shell by compressing the gas, and this balances the energy losses due to cooling. This process continues until the gas reaches very central region, where the radius is small enough that the gas's angular momentum can stop further infall. In this stage, the gas has reached very high density (as indicated by Fig. 5.8), and radiates its thermal energy on a very short timescale, and so becomes cold gas. This gas cooling picture is previously mentioned in Viola et al. (2008).

According to this picture, when a gas shell moves from the outer region to the halo centre, the cooling effectively radiates away the contraction work done by gravity. Because the temperature maps show the gas has roughly a constant temperature around $T_{\rm vir}$, this contraction can be roughly treated as isothermal. Then the total

contraction work for a shell with original radius r is

$$\begin{aligned} W(r) &= -\int_{V(r)}^{V(0)} PdV = m_{\rm gas}\int_{\rho(r)}^{\rho(0)} \frac{P}{\rho^2}d\rho \\ &= \frac{k_{\rm B}T_{\rm vir}m_{\rm gas}}{\mu_{\rm m}}\int_{\rho(r)}^{\rho(0)} \frac{1}{\rho}d\rho \\ &= \frac{2}{3}U\ln\frac{\rho(0)}{\rho(r)}, \end{aligned} \tag{5.3.1}$$

where V is the volume of this gas shell, $m_{\rm gas}$ its mass and ρ its density, $k_{\rm B}$ is the Boltzmann constant, $\mu_{\rm m}$ the mean molecular mass, and $U = (3k_{\rm B}T_{\rm vir}m_{\rm gas})/(2\mu_{\rm m})$ is the thermal energy of the shell. With the shell density given by the β-distribution with $r_{\rm core} \sim 0.05r_{\rm vir}$, the above equation gives $W(r) \sim 3U$, and because $W(r)$ only depends on r through $\ln\rho(r)$, it is not very sensitive to the radius.

Thus, instead of the simple picture assumed in most of the SA models, in which the gas gives away its thermal energy U and cools down, the gravitational contraction requires the gas to radiate away about $3U$ to cool down. The SA model tends to overestimate the cooling rate in this regime.

Note that the gravitational contraction does not play an important role in the fast cooling regime, because there the cooling is faster than the dynamical timescale and the gas completely cools down before significant contraction can happen, so P drops to very low values and little PdV work done. Instead, the related gravitational energy goes into kinetic energy of the infalling cold gas. So for the fast cooling regime the gas still only needs to give away total energy U to cool down. This contraction would not be important either if the gas is delivered by cold filaments.

5.3.1.3 Effects of Halo Mergers

Almost all SA gas cooling models assume that the gas newly accreted onto a dark matter halo is shock heated to $T_{\rm vir}$ of this halo. If a large amount of gas is accreted in a relatively short time duration, then this gas could cause significant heating of the hot gas halo, and thus interrupt its cooling. Halo major mergers are the most natural candidate of this rapid gas accretion.

Different SA cooling models treat the effect of this newly accreted gas differently. The MORGANA model explicitly quenches cooling for some time duration

(e.g. Monaco et al., 2007), while other models use more implicit modelling. We defer the comparison between SA models to §5.3.2, and here we focus on the comparison between the new SA cooling model and the hydrodynamical simulation.

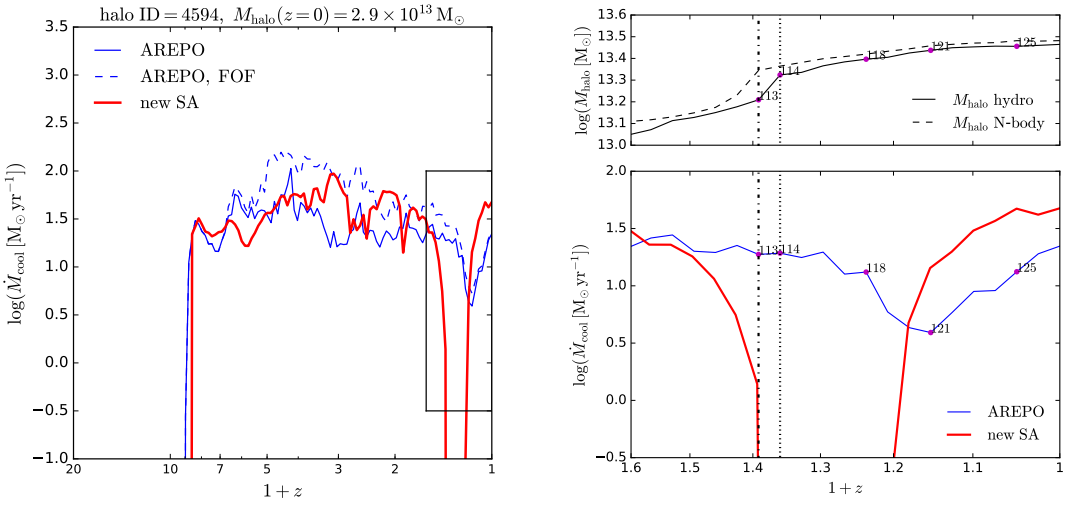

Figure 5.9: *left panel*: the mass cooling rate of halo 4594. The blue lines show the cooling rates measured from the hydrodynamical simulation. Among them, the solid line is the measurement for the central galaxy only, while the dashed line is for the whole FOF group. The red solid line is the result from the new SA cooling model. *right panels*: zoom in plot of the region in the black box in the left panel. The lower right panel shows the cooling rates, with the line meanings the same as in the left panel, and the magenta dots label the selected snapshots for showing further details in Fig. 5.10, with the corresponding numbers the snapshot IDs. The upper right panel shows the growth of halo mass. The vertical dotted line indicates the completion time of the halo merger in the hydrodynamical simulation according to the Dhalo merger tree, while the vertical dashed-dotted line indicates the corresponding time in the dark matter only simulation.

In the new SA model, the effect of this newly accreted gas is modelled straightforwardly. This gas is assumed to be newly heated up and thus has no previous cooling history. The accretion of it increases the total thermal energy of the hot gas halo, but keeps the total energy radiated away unchanged. Due to this effect, the cooling after the accretion could be suppressed, while the extent of this suppression depends on the amount of the gas accreted. When considering that the cooling

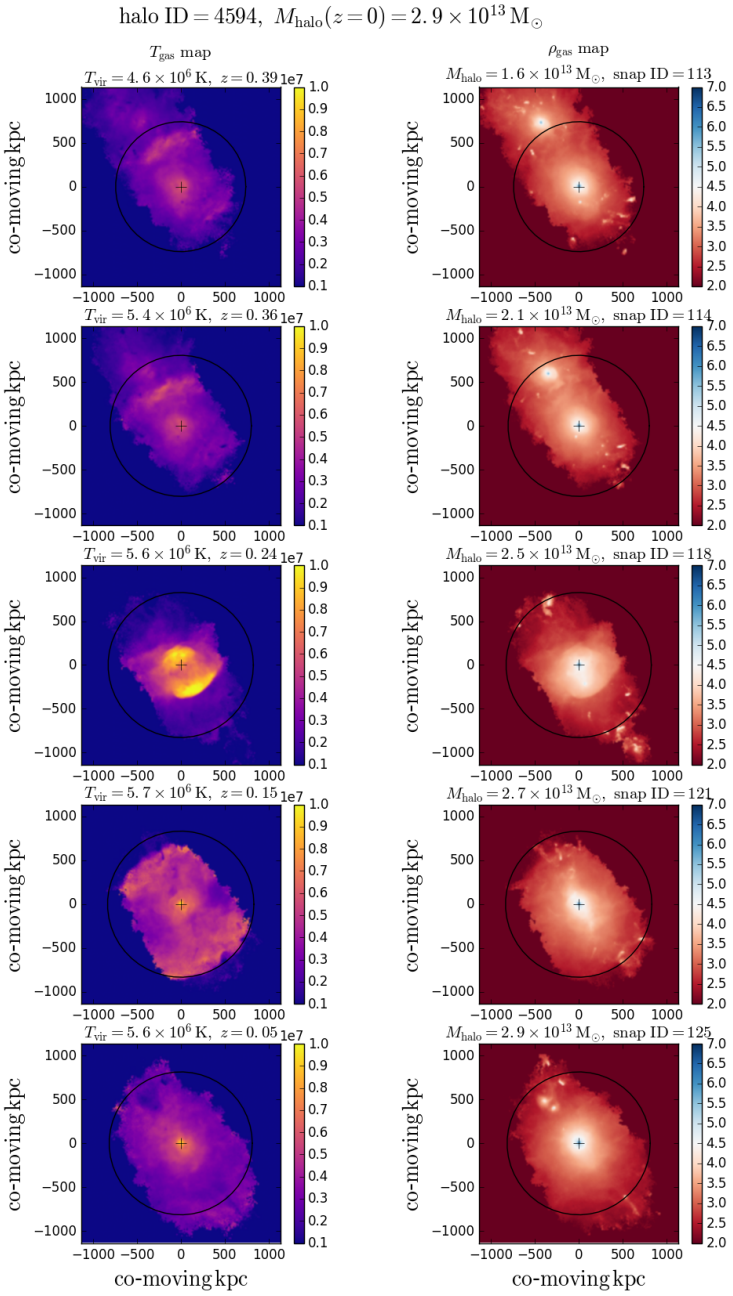

Figure 5.10: The projected temperature (left) and density (right) maps of halo 4594 for the selected snapshots labeled in Fig. 5.9. The left panels are temperature maps, while the right panels are density maps. To clearly view the evolution, here the colour scales for the temperature and density are on absolute scales. The colour scale for temperature is linear in units K, while the scale for density is in log(density), with the density in units $\mathrm{M}_\odot\,\mathrm{kpc}^{-3}$. The snapshot ID, redshift, halo mass and $T_{\rm vir}$ are given in the title of the corresponding row.

actually radiates the compression work done by gravity, as mentioned in the last sub-section, the suppression of cooling is still expected, because this accretion may increase the mean density and temperature of the hot gas halo, which, according to Eq(5.3.1), increases the budget of the total compression work.

It is interesting to see whether this suppression of cooling expected in the new SA cooling model also appears in the hydrodynamical simulation. Fig. 5.9 shows the mass cooling rate as a function of redshift of halo 4594. At $z \sim 0.3$, this halo experiences a major merger with mass ratio about 3 : 1. Accordingly, the SA model prediction gives a sharp drop in the mass cooling rate. From the zoom-in plots on the right it is clear that the sharp drop happens immediately after the merger in the SA model. This is expected, because when using the Dhalo merger tree, the halo merger is treated as an instantaneous event, and in the SA model, the associated gas accretion and heating are also assumed to be instantaneous. In the simulation result, there is also a drop, in which the cooling rate is reduced by a factor about 5. This drop is not as strong as that predicted by the SA model, but more importantly, it appears about 2 halo dynamical timescales (8 snapshots) later than the merger. Although there are small differences between the halo growth histories of the dark matter only (used in constructing merger trees for the SA model) and hydrodynamical simulations, as can be seen from the upper right panel of Fig. 5.9, this time delay is much larger than that, so it must be caused by some other reasons.

To further investigate the details of the drop in the mass cooling rate, we extract the projected gas temperature and density maps for several snapshots covering the halo merger and the drop. These maps are shown in Fig. 5.10, and the selected snapshots are labeled as magenta dots in the right panels of Fig. 5.9. To better view the temperature pattern of the gas, here a linear colour scale is adopted for temperature. For each snapshot, the maps are for the whole FOF group that contains halo 4594. Here the whole FOF group is chosen to better show the merging pair of halos.

In snapshot 113, two merging halos are clearly observable. They are in one FOF group, but the Dhalo algorithm still identifies them as two different Dhalos. In the temperature map, a weak heating due to gas compression can be seen between the

two halos. Then in snapshot 114, these two halos becomes closer and form a single Dhalo, so in the merger tree this snapshot labels the completion of the halo merger, but it seems that the merged structure has not relaxed yet, and the temperature map still only indicates a weak heating between these two merging halos. The merging process continues, and about one halo dynamical timescale later (4 snapshots later), in snapshot 118, a strong shock is generated by this merger, and from this moment, the cooling rate begins to drop, as can be seen from the lower right panel of Fig. 5.9.

About one halo dynamical timescale later, in snapshot 121, the strong shock has expanded and heated up nearly the whole hot gas halo. Accordingly, the mass cooling rate drops to a minimum. From the density map, the gaseous halo appears to be largely relaxed by this time. Then, after about another halo dynamical timescale, in snapshot 125, the hot gas halo becomes cooler, and the mass cooling rate rises back to a level close to that before the merger.

From these maps, two points can be summarized. Firstly, the suppression of cooling is associated with the shock heating induced by the merger, so at least in this case, the halo major merger does suppress cooling. Secondly, the suppression appears few halo dynamical timescales later than the moment of the completion of the Dhalo merger, because the merging halos are identified as one Dhalo before they are fully relaxed, and the time delay from the Dhalo merger to the suppression is the duration needed for the relaxation.

The SA model assumes that the halo is relaxed as soon as the Dhalo merger has completed, thus shifting the drop in cooling rate to earlier times. Also, the SA model predicts a stronger drop than the simulation. However, the general conclusion is that the SA model and hydrodynamical simulation have similar behavior during the halo major merger, and this is a triumph for the simple SA model.

The above example is for a major merger happens at low redshift. Next we consider a halo major merger at high redshift. For this purpose, we selected halo 9181. It experiences a halo merger with mass ratio about $3:1$ at $z \sim 5$. The mass cooling rate as a function of redshift for this halo is shown in Fig. 5.11. The part corresponding to this major merger is labeled as 'A' in the top panel and the zoom-in plots are shown at the lower left.

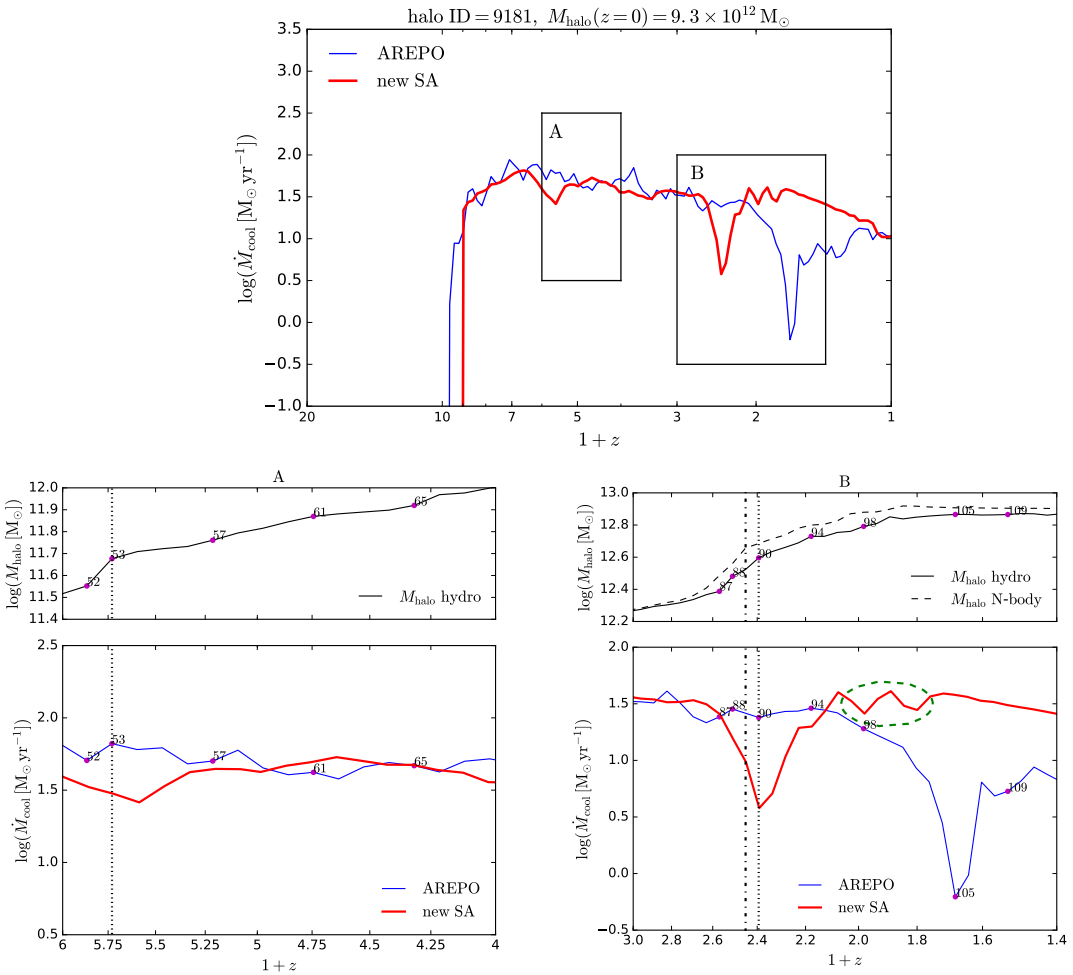

Figure 5.11: *top panel*: the mass cooling rate of halo 9181. The blue solid line is the cooling rate measured from the hydrodynamical simulation, for the central galaxy only, while the red solid line is the prediction of the new SA cooling model. *lower left panels*: zoom-in plot of the region labeled as 'A' in the top panel. The magenta dots label the selected snapshots for showing further details in Fig. 5.12, and the associated numbers are the snapshot IDs. The small upper panel shows the growth in halo mass. In this case the growth in the hydrodynamical and dark matter only simulations are almost the same, so here only the result from the former simulation is plotted. *lower right panels*: zoom in plot of the region labeled as 'B' in the top panel. The meanings of the symbols and lines are the same as for the lower left panels, but the magenta dots are for the snapshots shown in Fig. 5.13. In all lower panels, the vertical lines indicate the completion of the halo merger according to the Dhalo merger trees. The dotted and dashed-dotted lines indicate this in the hydrodynamical and dark matter only simulation respectively.

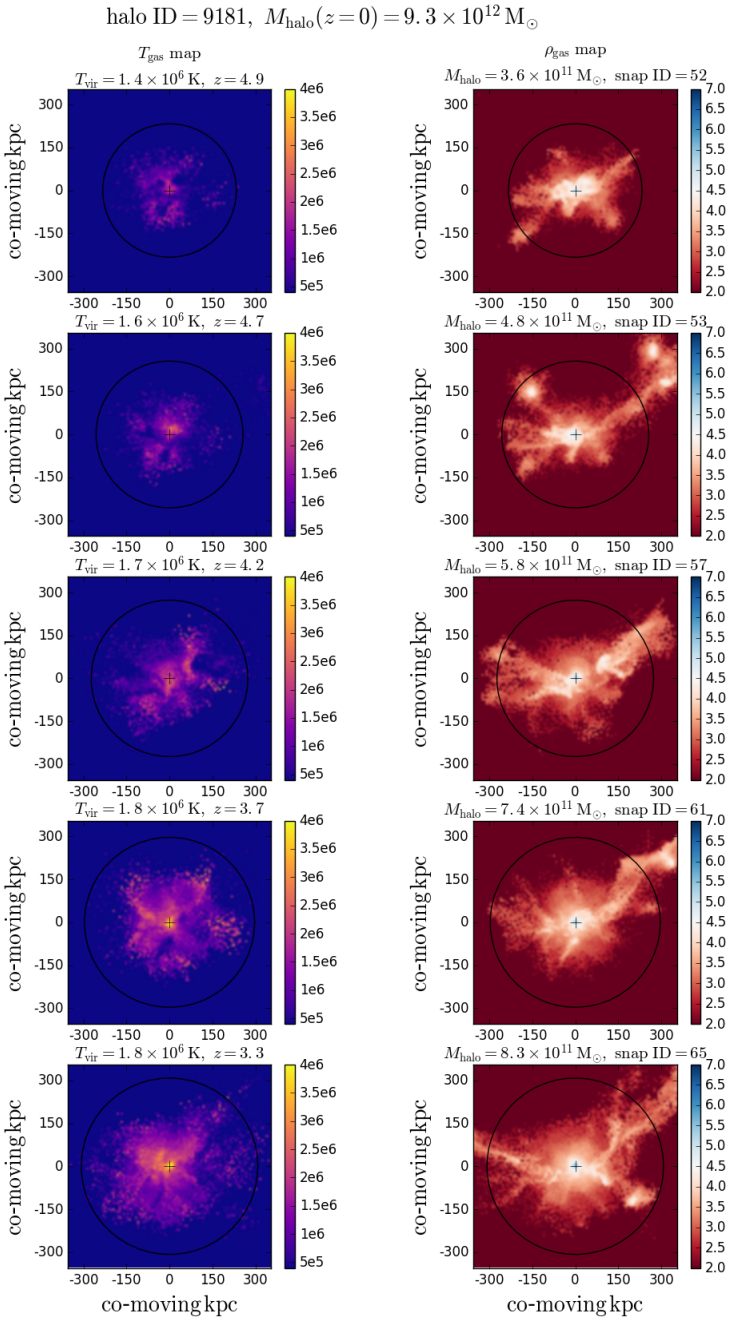

Figure 5.12: The projected temperature (left) and density (right) maps of halo 9181 for the selected snapshots labeled in the lower left panels (region A) of Fig. 5.11. The left panels are temperature maps, while the right panels are density maps. The colour scale for temperature is linear in units K, while the scale for density is in log(density), with the density in units $\mathrm{M}_\odot\,\mathrm{kpc}^{-3}$. The snapshot ID, redshift, halo mass and $T_{\rm vir}$ are given in the title of the corresponding row.

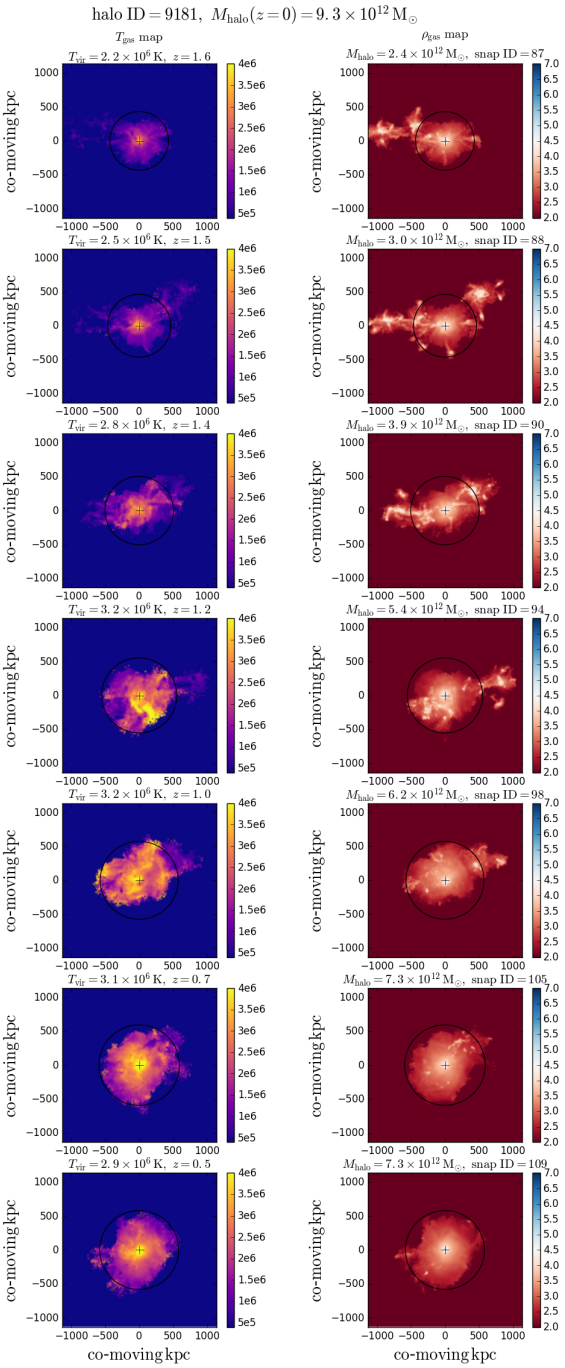

Figure 5.13: The projected temperature (left) and density (right) maps of halo 9181 for the selected snapshots labeled in the lower right panels (region B) of Fig. 5.11. The left panels are temperature maps, while the right panels are density maps. The colour scales have the same meaning as in Fig. 5.12, and see its caption for more information.

It seems that the hydrodynamical simulation does not predict any drop in mass cooling rate related to this merger. The SA model predicts some drop, but the cooling rate is reduced by a factor only about 2, which is much weaker than for the low redshift counterpart discussed above. Thus, neither the simulation nor the SA model predicts strong suppression of cooling for this major merger. To investigate the reason for this, the temperature and density maps of the gas were generated for the snapshots just before the merger, just after the Dhalo merger and one, two and three halo dynamical timescales after the merger. These selected snapshots are shown as magenta dots in the lower left panel of Fig. 5.11, while the maps are shown in Fig. 5.12.

From the density maps in Fig. 5.12, it is clear that at this high redshift, the gas is filamentary rather than in a spherical gas halo. These gas filaments can hardly be seen in the temperature maps. Here the colour scale for the temperature is set to be only sensitive to the gas with $T \gtrsim T_{\rm vir}$, and this means that the filamentary gas has $T < T_{\rm vir}$, namely it is cold. This is consistent with the findings of many other papers (e.g. Nelson et al., 2016; Kereš et al., 2005).

Just before merger, in snapshot 52, the halo gas is dominated by the filamentary cold gas, while the temperature map shows that the hot gas halo is less developed. Then for the snapshots shown after the merger, the density maps continue to indicate the existence of filamentary gas, while a hot gas halo component becomes more and more obvious. There is no strong shock as in Fig. 5.10 for the low redshift merger, and the development of the hot gas halo is more associated with the gradual transition from the cold accretion to the slow cooling regime, as the halo gradually grows from $3\times10^{11}\,{\rm M}_{\odot}$ to $8\times10^{11}\,{\rm M}_{\odot}$, so it is largely irrelevant to the major merger.

The cold filamentary gas is very hard to be heat by shocks. Thus a major merger happening in the cold accretion regime hardly suppresses the cooling. In the SA model, there is no filamentary gas, but for the relatively low halo mass, the assumed hot gas halo is close to the fast cooling regime, in which the cooling timescale is very short, and so significant heating of the gas is also very difficult. Therefore, the SA model does not predict a strong suppression of cooling either.

Halo 9181 shows a deep drop in the mass cooling rate at $z \sim 1$ for the hydrody-

namical simulation, as shown in the region labeled as 'B' in Fig. 5.11. Through the zoom in plots in the lower right corner, it is found that this drop is not caused by a single major merger, but a series of smaller mergers. Two mergers with mass ratios 4 : 1 and 5 : 1 happen successively, and together cause a rapid mass increase. The Dhalo merger completes in snapshot 90, while from the density and temperature maps in Fig. 5.13, again the merged halo has not relaxed yet in this snapshot, and the relaxation happens during the following two halo dynamical timescales. The gaseous halo is heated up during this process, as can be seen from the temperature maps of snapshots 94 and 98.

After snapshot 90, there are still some relatively rapid mass increases, and this together with the heating induced by the two mergers causes a deep drop in cooling rate in the hydrodynamical simulation at snapshot 105, and about one halo dynamical timescale later, in snapshot 109, the cooling rate rises again. In the SA model, the reduction of cooling rate happens immediately after the completion of the Dhalo merger (i.e. in snapshot 89, because the two mergers complete one snapshot earlier in the dark matter only simulation), and because of the lack of a time delay, the effects caused by the further mass increases after snapshot 94 can not superpose onto the effects of these two mergers, so only cause small ripples after the deep drop in the SA prediction (green dashed ellipse in the lower right panel of Fig. 5.11) and the deep drop in the SA result is weaker than that in the simulation.

Overall, we found that rapid gas accretion induced by mergers does suppress gas cooling, but this is only for events happening at low redshifts for halos in the slow cooling regime. Previously Monaco et al. (2014) also investigated the suppression of cooling by major mergers. That work is based on the SPH simulations. Monaco et al. (2014) found no anti-correlation between the mass ratio and the mass cooling rate ratio of halos in two adjacent snapshots. The cooling rates are taken from the same snapshots from which the halo masses are taken, or from snapshots a few halo dynamical timescales later than the snapshots providing the halo masses. This lack of correlation means no systematic suppression of cooling due to major mergers is seen. From our results, this could be partially caused by the mixing of mergers at both high and low redshifts. Monaco et al. (2014) also provided results of two

individual mergers. These two mergers are at $z < 1$, but from Fig.11 of Monaco et al. (2014), there still seems to be no strong drop in cooling rates. We noticed that in Monaco et al. (2014) the mass cooling rate is measured for the entire FOF group, while here we measure this for each central galaxy, but we checked that different measurements do not significantly weaken the drop, as shown by the blue dashed line in the left panel of Fig. 5.9. It is still possible that measuring the total mass cooling rate in the FOF group can erase the drop in cooling rate for some cases. The difference between our results and Monaco et al. (2014) could also be caused by the differences between SPH and moving mesh methods for hydrodynamical simulations.

5.3.1.4 Artificial Effects

In the left panel of Fig. 5.3, the mass cooling rate measured from the simulation drops sharply at $z \sim 0$. This kind of phenomenon is mainly observed in halos with $M_{\rm halo}(z = 0) < 10^{12}\,{\rm M}_{\odot}$. We checked that this is because for these relatively low mass halos, at $z \sim 0$, about 80% of the total baryons in these halos have already cooled down and been turned into stars by our star formation recipe, so that the gas in the central galaxy becomes so diffuse that its density falls below the threshold for star formation. By only counting stars, we then missed this part of the cold gas. Also note that at $z \sim 0$, the remaining gas is typically accreted onto the central galaxy at late times, compared to the gas accreted at early times, this gas tends to have higher angular momentum (because on average the angular momentum of the dark matter halo increases with halo growth), and this is another factor reducing the gas density in the central galaxy. Because this effect only happens in some low mass halos at $z \sim 0$, omitting it would not strongly change our results for the cumulative cool mass and angular momentum, or for the evolution of mass cooling rates over a large redshift range.

In the left panels of Fig. 5.3 and 5.5, it is also observed that the increase of cooling rates at high redshift in the simulation is more gradual and appears earlier than in the SA model. This is an artificial effect caused by our temperature threshold for cooling. According to Eq(5.2.1), a gas cell is allowed to cool only if its temperature is high enough. The temperature threshold roughly corresponds to the $T_{\rm vir}$ of a halo

with mass $2 \times 10^{10}\,\mathrm{M}_{\odot}$. In a halo with $M_{\mathrm{halo}} \ll 2 \times 10^{10}\,\mathrm{M}_{\odot}$, an artificial hot gas halo forms due to lack of cooling. As shown by Fig. 5.1, in the simulation, the hot gas in the central region of a halo tends to have higher temperature. Thus, in the simulation, when a halo is still below $2 \times 10^{10}\,\mathrm{M}_{\odot}$, the cooling has already begun in its central region, and later when the halo is more massive, the cooling gradually extends over the whole hot gas halo. Thus the mass cooling rate gradually rises in the simulation. In the SA model, because it is assumed that the hot gas halo has a temperature equaling T_{vir} and independent of radius, so the hot gas halo can only start cooling when the halo mass reaches $2 \times 10^{10}\,\mathrm{M}_{\odot}$, and once the cooling is allowed, the whole hot gas halo starts cooling immediately. Thus the cooling rate rises sharply, but slightly later than in the simulation.

We also note that in some rare cases the SA model starts cooling earlier than the simulation. This is because the Dhalo merger tree used in the SA model is extracted from the dark matter only simulation, and it could be slightly different from the halo growth history in the hydrodynamical simulation. Usually this small difference does not strongly affect cooling, but if it appears near the halo mass $2 \times 10^{10}\,\mathrm{M}_{\odot}$, then it can happen that in the SA model the halo reaches $2 \times 10^{10}\,\mathrm{M}_{\odot}$ first and starts cooling, while the corresponding halo in the hydrodynamical simulation reaches this mass and thus starts cooling slightly later.

5.3.2 Model Comparison

In this section we compare predictions from different SA cooling models with our hydrodynamical simulation. We first investigate some details of the SA models through a case study, because these details cannot be clearly shown in a statistical way. In this case study, we continue to focus on the mass cooling rate. Then we a the statistical comparison between the predictions of the SA models and the simulation. These predictions include the mass cooling rate, cumulative cool mass and two quantities associated with the specific angular momentum, j_{cool} and $\tilde{j}_{\mathrm{cool}}$. The definitions of the latter two quantities are given in §5.2.4 by Eq(5.2.8) and Eq(5.2.9) respectively.

5.3.2.1 Case Study

Fig. 5.14 shows the mass cooling rates predicted by different SA models for halo 4594. The top panel shows the result of the GFC1 cooling model, which is widely used in different versions of GALFORM. This model generates many sharp drops in the cooling rate that do not appear in the simulation results. The majority of these drops are seen to be associated with the artificial halo formation events introduced in the GFC1 model, as many drops appear just after the vertical dotted lines, which indicate the redshifts of this kind of events. The halo formation events cause drops in the cooling rates because the time available for cooling, $t_{\rm cool,avail}$, is reset to be zero at each halo formation event, which means the whole hot gas halo forgets its previous cooling history and is effectively newly heated up. At high redshifts, e.g. $z \geq 6$, the halo formation events only cause small drops, because the cooling timescale is very short for high redshift, low mass halos, while at lower redshifts, the cooling timescale becomes increasingly longer and just after a halo formation event the gas has to wait for a longer and longer time to cool down. During this wait, the cooling rate drops to zero, and correspondingly, wider and wider drops appear.

However, there are some drops not associated with halo formation events. These drops are caused by the way $t_{\rm cool,avail}$ is estimated. The reason for this is as follows. At any given moment, the GFC1 model calculates a cooling radius $r_{\rm cool}$ and assumes that all gas within $r_{\rm cool}$ has cooled down by this time. $r_{\rm cool}$ itself is determined from $t_{\rm cool}(r_{\rm cool}) = t_{\rm cool,avail}$, where $t_{\rm cool}(r)$ is the cooling timescale at radius r for the current hot gas halo, while $t_{\rm cool,avail}$ is determined by the previous cooling history of the halo. The GFC1 model assumes $t_{\rm cool,avail}$ is always the physical time since the last halo formation event, which would be correct if the cooling starts from the last halo formation event, and the hot gas halo remained fixed during cooling. However, this cooling model allows the hot gas to evolve due to the growth of the dark matter halo, and considering that from high to low redshift, the halo tend to have gradually lower mean density (because the mean density of the dark matter halo is always $\Delta'_{\rm vir}\rho_{\rm crit}$, and both quantities decrease with redshifts), $t_{\rm cool}$ gradually becomes longer as a result of the reduction of the density, but the estimation of $t_{\rm cool,avail}$ does not include a corresponding adjustment, so $t_{\rm cool,avail}$ is effectively underestimated, leading to an

underestimated $r_{\rm cool}$. Once this $r_{\rm cool}$ is smaller than the cooling radius of the previous time step, then this model determines that there is no cooling for the current time step, and there is a sharp drop in the cooling rate.

Also note that in this cooling model, the halo virial velocity $v_{\rm vir}$ is only updated at each halo formation event. When the halo grows its mass between two halo formation events, its $v_{\rm vir}$ is still kept unchanged, typically this is smaller than the true virial velocity of the halo after growth, so that $r_{\rm vir}$ calculated based on this velocity is overestimated, while the mean halo density is underestimated. This artificial underestimation in density further worsens the underestimation of $t_{\rm cool,avail}$, and causes these drops to appear more frequently.

The GFC2 model aims to remove the dependence of cooling on the artificial halo formation events, to make the predicted cooling history more continuous. The middle panel of Fig. 5.14 shows the result of the GFC2 cooling model. Surprisingly, although the formulation of this model is intended to make the cooling continuous, the actual predicted cooling history still shows many sharp drops. These drops are mainly caused by the very rough calculation of the total energy lost by cooling and by the method of calculating the time available for free-fall.

Just as in the new cooling model, GFC2 model accumulates the total energy radiated away as a record of the cooling history of the hot gas halo. Then at any given time, this energy is divided by the cooling luminosity of the current hot halo to derive $t_{\rm cool,avail}$ corresponding to the current halo. This method includes the effects of hot gas halo evolution on $t_{\rm cool,avail}$, so in principle it should avoid the problems identified for the GFC1 model above. However, because a very rough approximation is employed to calculate the cooling luminosity and total energy radiated away, the effects of halo evolution in the calculations of these two quantities do not necessarily match those in the calculation of $t_{\rm cool}$, and sometimes this model still underestimates $r_{\rm cool}$. When $r_{\rm cool}$ of the current time step becomes smaller than that of the previous time step, again, the model predicts a drop to zero mass cooling rate.

Now we consider how the calculation of $t_{\rm ff,avail}$ causes drops in cooling rates. In the GFC1 and GFC2 models, although it is assumed that the gas within $r_{\rm cool}$ has cooled down, this gas has not necessarily been accreted by the central galaxy,

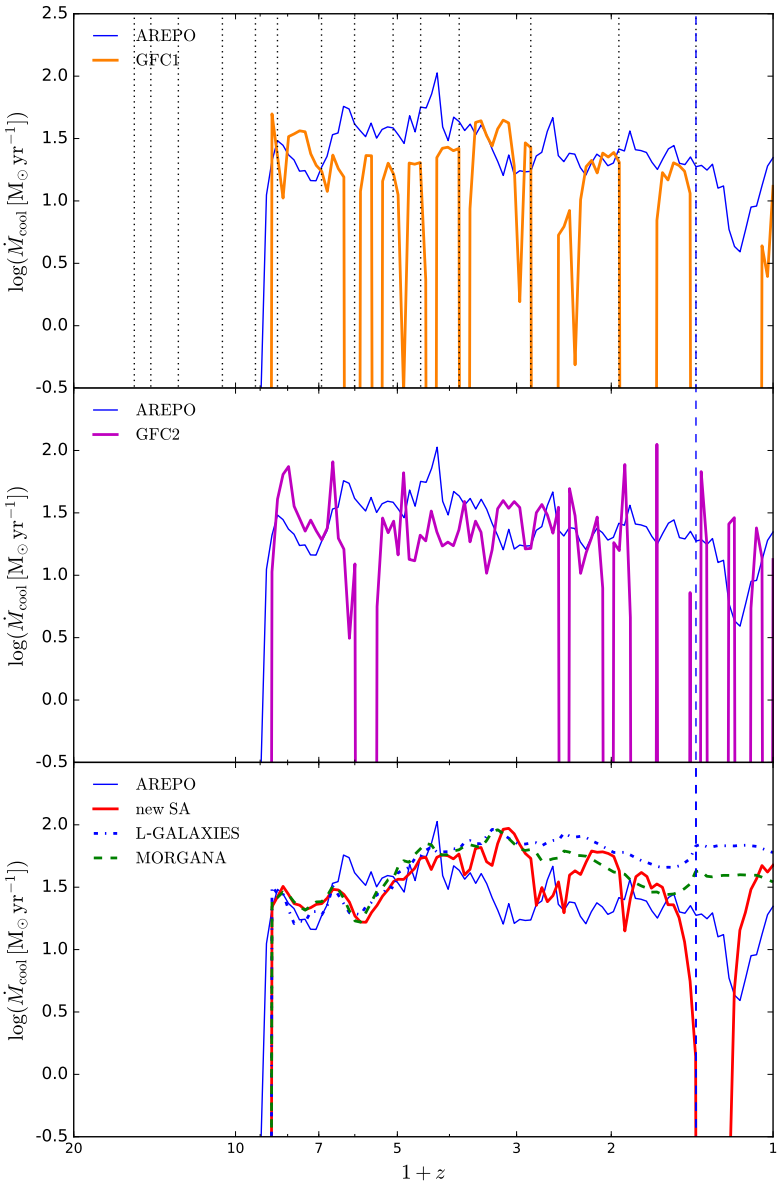

Figure 5.14: The mass cooling rates predicted by different SA models for halo 4594. The blue vertical dashed line in each panel indicates the redshift of a halo major merger, which causes the drop in cooling rate in both the hydrodynamical simulation and the new SA model. Here this redshift is derived from the Dhalo merger tree used in the SA models, i.e. constructed based on the dark matter only simulation. In the top panel, the vertical dotted lines indicate the artificial halo formation events calculated for the GFC1 cooling model. These are not used in the other models, so these lines are omitted in the other panels.

because it may not have had enough time for infall. The gravitational infall is considered in these two models though a free-fall radius $r_{\rm ff}$, with this radius defined through $t_{\rm ff}(r_{\rm ff}) = t_{\rm ff,avail}$, where $t_{\rm ff}(r)$ is the gravitational free-fall timescale for radius r, and $t_{\rm ff,avail}$ is the time available for free-fall. Only the gas within both $r_{\rm cool}$ and $r_{\rm ff}$ is accreted by the central galaxy. The GFC2 model uses the same method to calculate $t_{\rm ff,avail}$ as for the calculation of $t_{\rm cool,avail}$, namely the total energy radiated away divided by the current cooling luminosity. The timescale $t_{\rm ff,avail}$ is not allowed to exceed $t_{\rm ff}(r_{\rm vir})$, and once $t_{\rm ff,avail}$ becomes larger than this limit, the total energy used to derive it is reset to $t_{\rm ff}(r_{\rm vir}) \times L_{\rm cool}$, with $L_{\rm cool}$ the cooling luminosity of the current hot gas halo.

As mentioned in chapter 4, the accumulation of the total energy lost by radiative cooling in the GFC2 model is biased to lower values. More specifically, the mass of the cool gas is gradually removed from the hot gas halo to allow this halo to contract towards the halo center (this is reasonable for a physical model, since the cool gas would not provide pressure support). Accordingly the contribution of this removed gas to the total energy radiated away should also be removed, and this cooling model subtracts the total thermal energy of the removed gas from the total energy radiated away. This subtraction would be correct if this cooling model accumulated the actual energy radiated away, which, for each gas shell of the halo, is $\tilde{\Lambda}\rho^2(r)dV\Delta t$, where $\tilde{\Lambda}$ is the cooling function, $\rho(r)$ is the density of the shell with r its radius and dV its volume, while Δt is the time step length. However, the GFC2 model uses the rough approximation $\tilde{\Lambda}\rho(r)\bar{\rho}dV\Delta t$, with $\bar{\rho}$ the mean density of the hot gas halo, so if the cooling happens in the inner region of the hot gas halo (typical in the slow cooling regime), where $\rho(r) > \bar{\rho}$, then this approximation underestimates the energy lost by cooling, and the above subtraction removes more energy than necessary. This would lead to an underestimation of $t_{\rm cool,avail}$, and because $t_{\rm ff,avail}$ is calculated in a similar way, it is also underestimated. Furthermore, consider that at early times, the cooling is so fast that the derived $t_{\rm ff,avail}$ can easily lead to $r_{\rm ff} > r_{\rm vir}$, so the total energy used to calculate $t_{\rm ff,avail}$ is frequently reset to its limit value described above, while the energy used for $t_{\rm cool,avail}$ gradually accumulates to larger values. Thus, $t_{\rm ff,avail}$ is more sensitive to the biased subtraction. At late times, the underestimation of

$t_{\rm ff,avail}$ can lead to the reduction of $r_{\rm ff}$, and sometimes $r_{\rm ff} < r_{\rm cool}$ even for halos in the slow cooling regime. If $r_{\rm ff}$ at the current time step is smaller than that at the previous step, then no cool gas is accreted by the central galaxy, and there is a drop in the cooling rate.

Note that although the GFC1 and GFC2 cooling models result in many artificial drops in mass cooling rates, the effects on the cumulative cool mass are not very strong, because typically each drop only lasts for a short time.

As shown in the bottom panel of Fig. 5.14, the L-GALAXIES cooling model gives a very smooth evolution of cooling rate, which is better than the results from the previous two cooling models. However, this model predicts that during the low redshift halo major merger (indicated by the blue vertical dashed line), there is no suppression of gas cooling, but instead, the cooling rate increases by about a factor of two. This suppression is expected in the new cooling model and also observed in the simulation, as discussed in detail in §5.3.1.3.

The behaviors of the L-GALAXIES cooling model during major merger can be understood as follows: This cooling model assumes $t_{\rm cool,avail} = t_{\rm dyn}$. Note that $t_{\rm dyn} = r_{\rm vir}/v_{\rm vir}$ is independent of halo mass, but evolves with redshift. Consider that halo major mergers typically only happen over a short time duration, and for mergers at low redshift, the redshift change over the merger can be ignored. Thus, $t_{\rm cool,avail}$ in this model is almost unaffected by major mergers, so the gas accreted through a major merger also has the previous cooling history, namely it is not newly heated up by the merger. This lack of heating leads to the absence of cooling suppression.

Then consider that $r_{\rm vir} \propto M_{\rm halo}^{1/3}$, $T_{\rm vir} \propto M_{\rm halo}^{2/3}$ (the proportionality factors are constants for a given redshift) and for massive halos, the cool mass is still a small part of the total baryon mass, so roughly $M_{\rm hot} \propto M_{\rm halo}$. The L-GALAXIES model assumes the hot gas density profile to be singular isothermal, i.e. $\rho(r) = M_{\rm hot}/(4\pi r_{\rm vir}) r^{-2}$, and according to the above scaling relations, $\rho(r) \propto M_{\rm halo}^{2/3}$ at a given r. A halo major merger increases $M_{\rm halo}$ by up to a factor two, and thus $T_{\rm vir}$ is increased by a factor smaller than two, and for massive halos with $T_{\rm vir} > 10^6$ K, the cooling function $\tilde{\Lambda}$ only slightly increases with this temperature change. So, the cooling timescale $t_{\rm cool}(r) \propto T_{\rm vir}/(\tilde{\Lambda}\rho) \propto 1/\tilde{\Lambda}$ is largely unchanged during a major merger, and thus the cooling

radius $r_{\rm cool}$, which is derived through $t_{\rm cool}(r_{\rm cool}) = t_{\rm dyn}$, is also largely unchanged. For massive halos, typically $r_{\rm cool} < r_{\rm vir}$, and for this case, the L-GALAXIES cooling model calculates the mass cooling rate as $\dot{M}_{\rm cool} = [M_{\rm hot} r_{\rm cool}]/[r_{\rm vir} t_{\rm dyn}]$. Now it is obvious that the increase of $M_{\rm hot}$ during major merger dominates the change of cooling rate, and enhances it by a factor about two.

The MORGANA cooling model also predicts a very smooth cooling history. The model used here does not include the additional suppression of cooling during major mergers that was imposed in Monaco et al. (2007), and in this case, this model always assumes that each shell of the hot gas halo contributes to the current cooling rate, with the contribution being $dm\Delta t/t_{\rm cool}(r)$, where dm is the mass of a shell, r its radius and $t_{\rm cool}(r)$ its cooling timescale, while Δt is the time step length. This calculation also gives a smooth evolution of cooling rate during a major merger. Although the MORGANA model assumes a hot gas profile different from the L-GALAXIES model, the above analysis for the L-GALAXIES model still largely applies. Then we can see that $t_{\rm cool}(r)$ is largely unchanged, while dm is increased due to the merger, so the cooling rate is enhanced during a merger. Of course the gas cooling during a major merger can be suppressed by incorporating the additional recipe in the MORGANA model, but unlike in the new cooling model, here this requires additional parameters to identify major mergers and determine the suppression duration. Also, as discussed in §5.3.1.3 (for halo 9181), the high redshift major mergers do not suppress cooling, while a sequence of low redshift smaller mergers can also suppress cooling, and these cannot be captured by a recipe that simply suppresses cooling during a major merger.

5.3.2.2 Multiple Halos/Statistical Study

We divided halos into several samples according to their mass at $z = 0$. Specifically, these samples are halos in the mass ranges $10^{11}\,{\rm M}_\odot \leq M_{\rm halo} < 3 \times 10^{11}\,{\rm M}_\odot$, $3 \times 10^{11}\,{\rm M}_\odot \leq M_{\rm halo} < 10^{12}\,{\rm M}_\odot$ and $10^{12}\,{\rm M}_\odot \leq M_{\rm halo} < 10^{13}\,{\rm M}_\odot$. The first range corresponds to halos in the fast cooling regime, while the third range correspond to halos going into the slow cooling regime at low redshift, and the second range is a transition region. There are 90, 55 and 35 halos in the three mass ranges respectively.

Because of the small box size [$(25\mathrm{Mpc})^3$] for the current simulation, there are only three halos more massive than $10^{13}\,\mathrm{M}_\odot$ at $z = 0$. Thus no statistical results can be derived for these massive halos. For these, we directly show the predictions for each individual halo.

Fig. 5.15 shows the statistics of the individual halo differences between the SA models and simulation, for the above mentioned three halo mass ranges and for the four quantities related to gas cooling, namely the cumulative cool mass, $M_{\rm cool}$, cooling rate, $\dot{M}_{\rm cool}$, and two quantities associated with the specific angular momentum, $\tilde{j}_{\rm cool}$ and $j_{\rm cool}$. In each panel, the thick lines show the medians of the differences, while the light lines of the same style indicate the $10-90\%$ range.

From this figure it is clear that different SA models predict cumulative cool masses fairly close to the predictions of the simulation. The GFC1 and GFC2 models predict lower $M_{\rm cool}$ than the simulation at low redshifts for halos with $10^{12}\,\mathrm{M}_\odot \leq M_{\rm halo} < 10^{13}\,\mathrm{M}_\odot$. As analyzed in Chapter 4, this is mainly caused by the lack of proper modelling of the hot halo contraction. Without this contraction, the hot gas is always at relatively low density, and so the cooling is low because of the strong dependence of the cooling luminosity on density.

Compared to the simulation, the cooling rates in high redshift low mass halos are slightly underestimated by all of the SA models. In the simulation, the central galaxies in these halos gain cold gas mainly through filamentary accretion (see §5.3.1.1), while in the SA models, these halos are in the fast cooling regime. This underestimation indicates that although the mass accretion rates in both filamentary accretion and the fast cooling regime are mainly limited by the gravitational infall timescale, they are still slightly different from each other. Future direct modelling of the filamentary accretion in SA models may improve this point. At low redshifts, the new cooling model, L-GALAXIES and the MORGANA cooling model tend to give higher cooling rates than the simulation. This is related to the potential overestimation of cooling by the SA models in the slow cooling regime, as discussed in §5.3.1.2. The GFC1 and GFC2 models gives lower cooling rates than the other three, for the reason described above. The 10% envelopes for these two models are much wider than for the other three models. This is because these two models generate

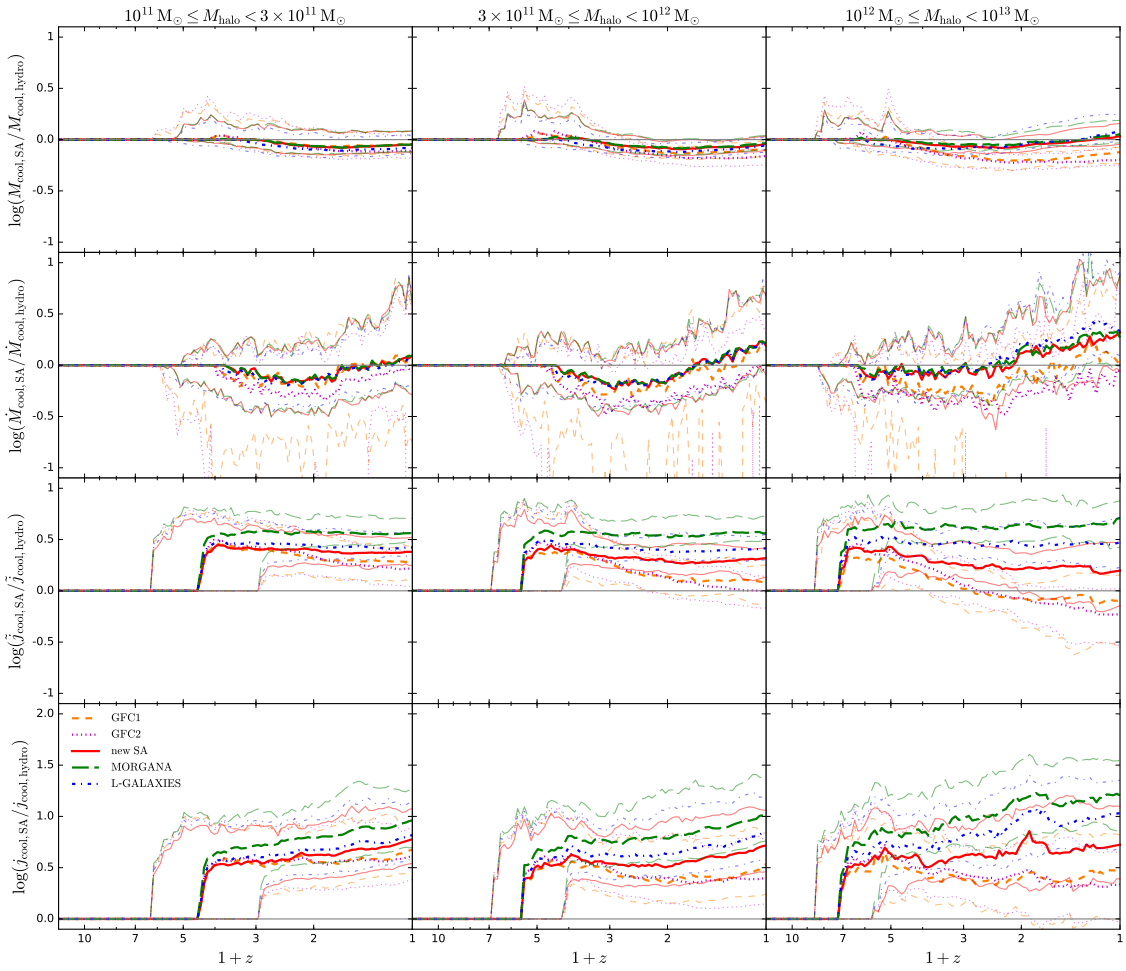

Figure 5.15: Statistical comparison between SA models and the simulation. Each column corresponds to a different halo mass range, with the specific range given at the top of the column. Each row corresponds to a different quantity, and from top to bottom these quantities are respectively the cumulative cool mass, $M_{\rm cool}$, cooling rate, $\dot{M}_{\rm cool}$, and two quantities associated with the specific angular momentum, $\tilde{j}_{\rm cool}$ and $j_{\rm cool}$. Each panel shows the logarithm of the ratio of the SA model prediction over the simulation prediction for the corresponding quantity, and the gray horizontal solid line indicates a ratio of one, i.e. the SA model and simulation giving the same prediction. Each line style corresponds to a different SA model, with the model name given in the key in the lower left corner. The thick lines in each panel indicate the medians of the ratio, while the light lines of the same style indicate the $10-90\%$ range of the ratio.

many artificial drops in the mass cooling rates, which lead to large underestimates compared to the simulation results.

As mentioned in §5.2.4, $\tilde{j}_{\rm cool}$ is mainly sensitive to the magnitude of the angular momentum delivered to the central galaxy by cooling. Compared to the simulation, the SA models give too high $\tilde{j}_{\rm cool}$, especially for low mass halos. The median deviation is about a factor of two to three. The L-GALAXIES and MORGANA models give worse results than the new cooling model, while the GFC1 and GFC2 models tend to give too low $\tilde{j}_{\rm cool}$ for halos with $10^{12}\,{\rm M}_{\odot} \leq M_{\rm halo} < 10^{13}\,{\rm M}_{\odot}$ at low redshifts. Because all of the SA models predict $M_{\rm cool}$ very close to the simulation, the deviations observed here is mainly caused by the calculation of angular momentum. Interestingly, this overestimation gradually decreases with increasing halo mass, which indicates that it may be related to the fast cooling regime. Recently Danovich et al. (2015) found significant angular momentum loss of the cold gas streams infalling towards the central galaxy in an AMR simulation. This loss is believed to be mainly induced by the torque from the inner disk of the central galaxy. This torque could be the reason for the discrepancy between the simulation and SA models observed here, but further, more detailed checking is required to confirm this. I plan to do this in the future.

The SA models always assume that the central galaxy spin and the halo gas spin are aligned. As discussed in §5.2.4, the difference between $\tilde{j}_{\rm cool}$ and $j_{\rm cool}$ is sensitive to this assumption. If $\tilde{j}_{\rm cool}$ and $j_{\rm cool}$ are close, then these two spins are almost aligned, while if $\tilde{j}_{\rm cool}$ is obviously larger than $j_{\rm cool}$, then there are non-negligible direction differences between these two spins. As shown in Fig. 5.15, the specific angular momenta predicted by the SA models differ more from $j_{\rm cool}$ than $\tilde{j}_{\rm cool}$, so in the simulation, there are obvious spin direction differences.

The overestimation of the specific angular momentum in the SA models can affect various aspects of galaxy formation, such as disk sizes, star formation and SN feedback.

Fig. 5.16 shows the cooling predictions from the simulation and SA models for the three halos with $M_{\rm halo} > 10^{13}\,{\rm M}_{\odot}$. The general features of these results are the similar to those shown by the statistical comparison in Fig. 5.15. All SA models give

the total cool masses fairly close to the simulation results, with the GFC1 and GFC2 models tending to slightly underestimate this mass. The L-GALAXIES and MORGANA models predict too high $\tilde{j}_{\rm cool}$, while the GFC1 and GFC2 model predictions are too low at low redshifts. The new cooling model predicts $\tilde{j}_{\rm cool}$ close to the simulation at $z \lesssim 4$, but at higher redshift its predictions are still too high. The specific angular momenta predicted by the SA models are systematically higher than $j_{\rm cool}$ from the simulation, again, indicating the existence of non-negligible direction differences between the spins of the central galaxy and the halo gas.

5.4 Summary

In this chapter we compared the gas cooling models from several major semi-analytical (SA) galaxy formation models with our hydrodynamical simulation. The SA cooling models considered here are the new cooling model introduced in Chapter 4, the GFC1 and GFC2 models for GALFORM, and the cooling models from the L-GALAXIES and MORGANA models. Here the comparison focuses on four quantities related to the gas cooling. These are the total cool mass, $M_{\rm cool}$, mass cooling rate, $\dot{M}_{\rm cool}$, and two quantities related to specific angular momentum, namely $\tilde{j}_{\rm cool}$ and $j_{\rm cool}$. $\tilde{j}_{\rm cool}$ is more sensitive to the magnitude of the angular momentum delivered to central galaxies by cooling in each time step, while $j_{\rm cool}$ includes more effects from the directions of the angular momenta. If the angular momenta of central galaxy and halo gas are aligned, just as assumed in the SA models, then it is expected that $\tilde{j}_{\rm cool}$ and $j_{\rm cool}$ should be similar, while the non-negligible direction difference causes $\tilde{j}_{\rm cool}$ to be obviously larger than $j_{\rm cool}$.

This comparison provides not only an assessment of the accuracy of each cooling model, but also some insights into the physics of gas cooling in cosmological structure formation. Our main conclusions are summarized as follows:

(i) For halos with $M_{\rm halo} \lesssim 3 \times 10^{11}\,{\rm M}_{\odot}$, the SA models predict the cooling to be in the fast cooling regime, in which the cooling is faster than the infall, however, the simulation suggests that for these halos the gas is mainly delivered to the central galaxy through cold filaments, and thus the accretion is

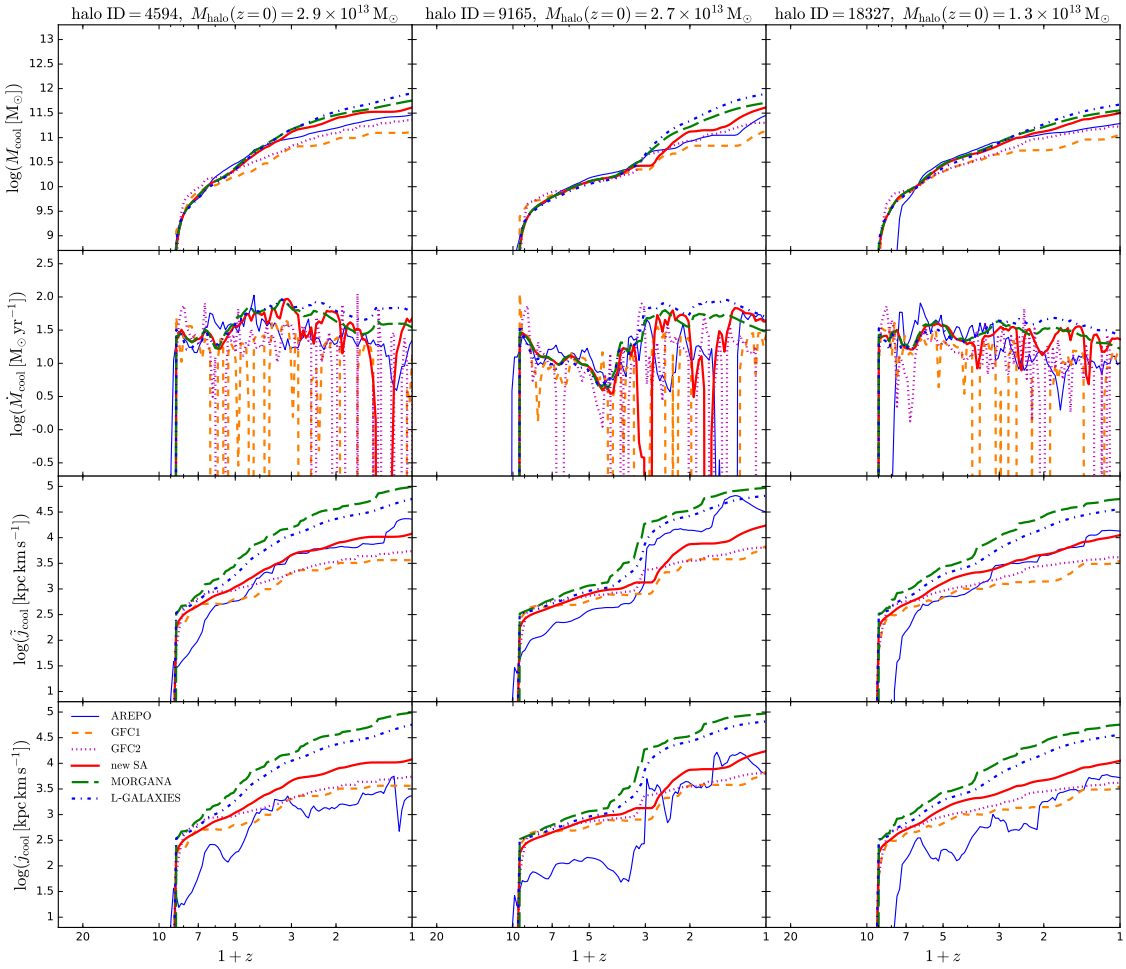

Figure 5.16: Predictions for gas cooling from the SA models and the simulation for the most massive halos. Each column is for a single massive halo, with its halo ID and mass at $z = 0$ given at the top of the column. Each row corresponds to a cooling quantity, and from top to bottom these are respectively the cumulative cool mass, $M_{\rm cool}$, mass cooling rate, $\dot{M}_{\rm cool}$, and two quantities associated with the specific angular momentum, $\tilde{j}_{\rm cool}$ and $J_{\rm cool}$. The thin blue solid lines are for the simulation results, while the other lines correspond to different SA models, with the specific model name given in the key in the lower left corner.

highly anisotropic. This filamentary nature is more obvious at high redshifts. Although these two pictures are very different, the predicted mass accretion rates are close, because in both pictures, these rates are mainly determined by the gravitational infall timescale.

(ii) In low redshift high mass halos, roughly spherical hot gas halos are seen in the simulation, in agreement with the slow cooling regime picture in the SA models. However, the simulation indicates that the gas has a roughly constant temperature around $T_{\rm vir}$ during cooling, until it reaches very high density, after which it cools down rapidly. This constancy of gas temperature is caused by the contraction work done by gravity when the gas gradually falls towards the halo centre. During the whole cooling process, the total work is about three times the initial gas thermal energy. The SA models typically do not consider this work, leading to overestimation of mass cooling rates in the slow cooling regime.

(iii) The simulation suggests that halo major mergers at low redshift can suppress cooling, while those at high redshift do not, because the cold filaments are hardly affected by mergers. At low redshift, a sequence of smaller mergers can also suppress cooling. The new cooling model can better capture these effects of mergers than the other SA cooling models. This is an advantage of the new cooling model. The complex effects of merger on cooling may explain the lack of correlation between the reduction of cooling rate and the merger mass ratio reported in Monaco et al. (2014). Monaco et al. (2014) did not see any obvious suppression of cooling due to low redshift major mergers, and this may be caused by the differences between the SPH and moving mesh hydrodynamical methods.

(iv) Compared to the simulation results, all SA models give total cool masses fairly close to the simulation results, with the GFC1 and GFC2 models tending to slightly underestimate this mass. These two models also generates many artificial drops in cooling rates. The new cooling model, as well as the other two models, predict more continuous evolution of the mass cooling rates. This is an advantage of these three models.

(v) The specific angular momenta predicted by the new cooling model, L-GALAXIES and the MORGANA model are systematically higher than $\tilde{j}_{\rm cool}$ measured from the simulation. The L-GALAXIES and MORGANA models give worse results

than the new cooling model. Because these models predict the total cool masses very close to the simulation values, this difference means they also overestimate the total angular momenta of galaxies. This difference is larger for lower mass halos, which indicates that it may be related to the fast cooling regime. The predictions of the GFC1 and GFC2 models for halos with $10^{11}\,\mathrm{M}_{\odot} \leq M_{\mathrm{halo}} < 3 \times 10^{11}\,\mathrm{M}_{\odot}$ are also higher than the simulation results, as in the other three models, but they tend to underestimate the angular momenta for halos more massive than $10^{12}\,\mathrm{M}_{\odot}$. For the three halos in the simulation more massive than $10^{13}\,\mathrm{M}_{\odot}$, the new cooling model seems to give generally better results than the other SA models. Further statistical comparison is required to fully assess the accuracy of these SA models in this halo mass range.

(vi) The specific angular momenta predicted by the SA models show even larger deviations from j_{cool}. This indicates that in the simulation there are non-negligible direction differences between the spins of the central galaxy and the halo gas.

Chapter 6

Conclusions and future work

In this work we try to improve the modelling of two important processes in galaxy formation, supernova (SN) feedback and gas cooling, in the Durham semi-analytical (SA) model GALFORM. These improvements take their starting point to be the Lacey16 model (Lacey et al., 2016), which is the newest version of GALFORM.

We first improve the SN feedback recipe in a more phenomenological way, namely using the constraints from four observations when combined with a galaxy formation model. These observations include the Milky Way (MW) satellite galaxy luminosity function, the faint end of the field galaxy luminosity function, the redshift at which the universe was half reionized and the stellar metallicity of the MW satellites. The former two are not reproduced when using a too weak SN feedback, while a too strong feedback leads to too late reionization and too low MW satellite metallicities. Together these four observations put tight constraints on SN feedback in galaxy formation models. A simple model in which the mass loading factor for SN feedback depends on galaxy circular velocities, with a normalization that depends on redshift, is able to reproduce the above mentioned observational constraints.

We further apply this improved SN feedback model to investigate some details of the reionization, and find that half of the ionizing photons are emitted by galaxies with rest-frame far-UV absolute magnitudes $M_{\rm AB}(1500\text{Å}) < -17.5$, which implies that already observed galaxy populations contribute about half of the photons responsible for reionization. The $z = 0$ descendants of these galaxies are mainly galaxies with stellar mass $M_* > 10^{10}\,{\rm M}_\odot$ and preferentially inhabit halos with mass

$M_{\rm halo} > 10^{13}\,{\rm M}_{\odot}$.

We then try to improve the modelling of gas cooling in halosinside SA models by revisiting the physical picture and approximation made, though still within the framework of spherically symmetric halos. This leads to a new, more physical model for gas cooling and accretion in halos in semi-analytical models. Compared to previous cooling models, this model incorporates a more physically consistent calculation of the hot gas cooling history, a more detailed modeling of the contraction of the hot gas halo induced by cooling, and a more detailed calculation of the angular momentum of the cooled down gas. This model predicts higher cooled down masses than the cooling models previously used in GALFORM, closer to the predictions of the cooling models in the L-GALAXIES and MORGANA SA models, even though those models are formulated differently. It also predicts cooled down angular momenta higher than in previous GALFORM cooling models, but generally lower than the predictions of L-GALAXIES and MORGANA. When used in the full GALFORM galaxy formation model, this cooling model improves the predictions for early-type galaxy sizes in GALFORM.

Finally we compare this new cooling model with a cosmological hydrodynamical simulation, along with two previous cooling models in GALFORM and the models in L-GALAXIES and MORGANA. The simulation is run using the grid-based moving mesh code AREPO. The star formation and various feedback processes are turned off in these SA models and in the simulation, to focus on the differences in gas cooling modelling. We find that generally all SA cooling models predict the cumulative cool mass close to the simulation, but the mass cooling rates in low redshift massive halos are overestimated, due to not including the compression work done on the gas during the hot gas halo contraction. These SA models tend to overpredict the specific angular momenta of cooled gas for low mass halos, while for low redshift high mass halos, the predicted specific angular momenta in the new cooling model generally agree better with the simulation, while the predictions of the L-GALAXIES and MORGANA models are higher than the simulation, and the predictions of the previous GALFORM cooling models are too low.

We also investigated the cooling in more detail by examining individual halos

in the simulation, and comparing with SA model predictions for the same halos. We find that the gas accretion in high redshift halos is clearly filamentary and very different from the spherical gas distribution assumed in the SA models, but this does not cause huge differences in the mass and angular momentum accretion rates, because in both the SA models and simulation, the accretion onto the central galaxies in these halos is mainly limited by the gravitational infall timescale. We see in the simulation the suppression of gas cooling by a halo major merger at low redshift, but a major merger at high redshift does not strongly affect gas cooling. A rapid mass increase at low redshift caused by a series of small halo mergers is also seen to suppress cooling.

The phenomenological study of SN feedback carried out here could be the first step towards a more complete and physical model for SN feedback. Future developments of this feedback model could also include a better modelling of the interaction between the hot gas halo and the gas ejected by feedback, which could have important effects on gas cooling.

The new cooling model could be further improved. The compression work done by gravity in the slow cooling regime should be included into the model to better predict mass cooling rates. The time delay between Dhalo mergers and the shock heating and suppression of cooling may also be modelled, which could improve the behavior of this cooling model for halo mergers. Also, a direct modelling of the filamentary gas accretion is desired for a even more physical gas cooling model.

A larger cosmological hydrodynamical simulation (with the same resolution but in a $[50\mathrm{Mpc}]^3$ box) has been done. Statistical results for comparison the cooling prediction in halos more massive than $10^{13}\,\mathrm{M}_{\odot}$ can be obtained by analyzing this larger simulation. Further simulations would also help to test the assumptions that we made here for measuring the cool mass and angular momentum from the simulation. More detailed investigations of the simulations may reveal the reason for the overestimation in the SA models of the specific angular momentum of cooled gas in low mass halos.

Bibliography

Abadi M. G., Bower R. G., Navarro J. F., 2000, MNRAS, 314, 759

Alexandroff R. M., Heckman T. M., Borthakur S., Overzier R., Leitherer C., 2015, ApJ, 810, 104

Alpher R. A., Bethe H., Gamow G., 1948, Physical Review, 73, 803

Asplund M., Grevesse N., Sauval A. J., Scott P., 2009, ARA&A, 47, 481

Bahé Y. M. et al., 2016, MNRAS, 456, 1115

Bate M. R., 2012, MNRAS, 419, 3115

Baugh C. M., Lacey C. G., Frenk C. S., Granato G. L., Silva L., Bressan A., Benson A. J., Cole S., 2005, MNRAS, 356, 1191

Beck A. M. et al., 2016, MNRAS, 455, 2110

Benson A. J., 2005, MNRAS, 358, 551

Benson A. J., 2010, Phys. Rep., 495, 33

Benson A. J., 2012, New A, 17, 175

Benson A. J., Bower R., 2010, MNRAS, 405, 1573

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

Benson A. J., Lacey C. G., Baugh C. M., Cole S., Frenk C. S., 2002, MNRAS, 333, 156

Benson A. J., Pearce F. R., Frenk C. S., Baugh C. M., Jenkins A., 2001, MNRAS, 320, 261

Bertschinger E., 1989, ApJ, 340, 666

Bett P., Eke V., Frenk C. S., Jenkins A., Helly J., Navarro J., 2007, MNRAS, 376, 215

Bigiel F. et al., 2011, ApJ, 730, L13

Binney J., 1977, ApJ, 215, 483

Birnboim Y., Dekel A., 2003, MNRAS, 345, 349

Blitz L., Rosolowsky E., 2006, ApJ, 650, 933

Bond J. R., Cole S., Efstathiou G., Kaiser N., 1991, ApJ, 379, 440

Borthakur S., Heckman T. M., Leitherer C., Overzier R. A., 2014, Science, 346, 216

Bournaud F., Jog C. J., Combes F., 2005, A&A, 437, 69

Bouwens R. J. et al., 2011a, Nature, 469, 504

Bouwens R. J. et al., 2011b, ApJ, 737, 90

Bouwens R. J. et al., 2015, ApJ, 803, 34

Bower R. G., Benson A. J., Malbon R., Helly J. C., Frenk C. S., Baugh C. M., Cole S., Lacey C. G., 2006, MNRAS, 370, 645

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

Bowler R. A. A. et al., 2014, MNRAS, 440, 2810

Boylan-Kolchin M., Bullock J. S., Garrison-Kimmel S., 2014, MNRAS, 443, L44

Bryan G. L., Norman M. L., 1998, ApJ, 495, 80

Cattaneo A. et al., 2017, ArXiv:1706.07106

Cattaneo A., Dekel A., Devriendt J., Guiderdoni B., Blaizot J., 2006, MNRAS, 370, 1651

Chabrier G., 2003, PASP, 115, 763

Chandrasekhar S., 1943, ApJ, 97, 255

Christodoulou D. M., Shlosman I., Tohline J. E., 1995, ApJ, 443, 551

Cole S., Lacey C., 1996, MNRAS, 281, 716

Cole S., Lacey C. G., Baugh C. M., Frenk C. S., 2000, MNRAS, 319, 168

Cole S. et al., 2005, MNRAS, 362, 505

Combes F., Debbasch F., Friedli D., Pfenniger D., 1990, A&A, 233, 82

Couchman H. M. P., Rees M. J., 1986, MNRAS, 221, 53

Cox T. J., Jonsson P., Somerville R. S., Primack J. R., Dekel A., 2008, MNRAS, 384, 386

Croton D. J. et al., 2006, MNRAS, 365, 11

Cyburt R. H., Fields B. D., Olive K. A., 2008, J. Cosmology Astropart. Phys., 11, 012

Danovich M., Dekel A., Hahn O., Ceverino D., Primack J., 2015, MNRAS, 449, 2087

Davé R., Katz N., Oppenheimer B. D., Kollmeier J. A., Weinberg D. H., 2013, MNRAS, 434, 2645

Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, ApJ, 292, 371

de Boer T. J. L. et al., 2012, A&A, 544, A73

De Lucia G., Blaizot J., 2007, MNRAS, 375, 2

De Lucia G., Boylan-Kolchin M., Benson A. J., Fontanot F., Monaco P., 2010, MNRAS, 406, 1533

Dekel A., Birnboim Y., 2006, MNRAS, 368, 2

Dekel A., Silk J., 1986, ApJ, 303, 39

Doggett J. B., Branch D., 1985, AJ, 90, 2303

Driver S. P. et al., 2012, MNRAS, 427, 3244

Efstathiou G., 1992, MNRAS, 256, 43P

Efstathiou G., Lake G., Negroponte J., 1982, MNRAS, 199, 1069

Eke V. R., Cole S., Frenk C. S., 1996, MNRAS, 282

Elmegreen B. G., 1993, ApJ, 411, 170

Fanidakis N., Baugh C. M., Benson A. J., Bower R. G., Cole S., Done C., Frenk C. S., 2011, MNRAS, 410, 53

Ferrara A., Bianchi S., Cimatti A., Giovanardi C., 1999, ApJS, 123, 437

Finkelstein S. L. et al., 2014, ArXiv:1410.5439

Finlator K., Oh S. P., Özel F., Davé R., 2012, MNRAS, 427, 2464

Font A. S. et al., 2011, MNRAS, 417, 1260

Fontanot F., Cristiani S., Vanzella E., 2012, MNRAS, 425, 1413

Garcia-Palacios J. L., 2007, eprint arXiv:cond-mat/0701242

Gardner J. P., 2001, ApJ, 557, 616

Giallongo E. et al., 2015, A&A, 578, A83

Gonzalez-Perez V., Lacey C. G., Baugh C. M., Lagos C. D. P., Helly J., Campbell D. J. R., Mitchell P. D., 2014, MNRAS, 439, 264

Granato G. L., Lacey C. G., Silva L., Bressan A., Baugh C. M., Cole S., Frenk C. S., 2000, ApJ, 542, 710

Guo Q. et al., 2011, MNRAS, 413, 101

Haardt F., Madau P., 2012, ApJ, 746, 125

Hamann J., Hannestad S., Lesgourgues J., Rampf C., Wong Y. Y. Y., 2010, J. Cosmology Astropart. Phys., 7, 022

Hatton S., Devriendt J. E. G., Ninin S., Bouchet F. R., Guiderdoni B., Vibert D., 2003, MNRAS, 343, 75

Heitmann K. et al., 2008, Computational Science and Discovery, 1, 015003

Helly J. C., Cole S., Frenk C. S., Baugh C. M., Benson A., Lacey C., 2003a, MNRAS, 338, 903

Helly J. C., Cole S., Frenk C. S., Baugh C. M., Benson A., Lacey C., Pearce F. R., 2003b, MNRAS, 338, 913

Henriques B. M. B., White S. D. M., Thomas P. A., Angulo R., Guo Q., Lemson G., Springel V., Overzier R., 2015, MNRAS, 451, 2663

Hirschmann M., Somerville R. S., Naab T., Burkert A., 2012, MNRAS, 426, 237

Hopkins P. F., Quataert E., Murray N., 2012, MNRAS, 421, 3488

Jiang C. Y., Jing Y. P., Faltenbacher A., Lin W. P., Li C., 2008, ApJ, 675, 1095

Jiang L., Helly J. C., Cole S., Frenk C. S., 2014, MNRAS, 440, 2115

Kaiser N., 1984, ApJ, 284, L9

Katz N., Weinberg D. H., Hernquist L., 1996, ApJS, 105, 19

Kennicutt, Jr. R. C., 1983, ApJ, 272, 54

Kereš D., Katz N., Fardal M., Davé R., Weinberg D. H., 2009, MNRAS, 395, 160

Kereš D., Katz N., Weinberg D. H., Davé R., 2005, MNRAS, 363, 2

Kimm T., Cen R., 2014, ApJ, 788, 121

Klypin A., Kravtsov A. V., Valenzuela O., Prada F., 1999, ApJ, 522, 82

Komatsu E. et al., 2011, ApJS, 192, 18

Koposov S. et al., 2008, ApJ, 686, 279

Kormendy J., Kennicutt, Jr. R. C., 2004, ARA&A, 42, 603

Kregel M., van der Kruit P. C., de Grijs R., 2002, MNRAS, 334, 646

Kroupa P., 2002, Science, 295, 82

Kuhlen M., Faucher-Giguère C.-A., 2012, MNRAS, 423, 862

Lacey C., Cole S., 1993, MNRAS, 262, 627

Lacey C. G. et al., 2016, MNRAS, 462, 3854

Lagos C. d. P., Lacey C. G., Baugh C. M., 2013, MNRAS, 436, 1787

Lagos C. D. P., Lacey C. G., Baugh C. M., Bower R. G., Benson A. J., 2011, MNRAS, 416, 1566

Larson R. B., 1974, MNRAS, 169, 229

Laureijs R. et al., 2011, ArXiv:1110.3193

Leroy A. K., Walter F., Brinks E., Bigiel F., de Blok W. J. G., Madore B., Thornley M. D., 2008, AJ, 136, 2782

Lewis A., 2008, Phys. Rev. D, 78, 023002

Lu Y., Mo H. J., Weinberg M. D., Katz N., 2011, MNRAS, 416, 1949

Madau P., Haardt F., 2015, ApJ, 813, L8

Madau P., Haardt F., Rees M. J., 1999, ApJ, 514, 648

Malbon R. K., Baugh C. M., Frenk C. S., Lacey C. G., 2007, MNRAS, 382, 1394

Maller A. H., Dekel A., Somerville R., 2002, MNRAS, 329, 423

Marasco A., Crain R. A., Schaye J., Bahé Y. M., van der Hulst T., Theuns T., Bower R. G., 2016, MNRAS, 461, 2630

Maraston C., 2005, MNRAS, 362, 799

Massey R. et al., 2007, ApJS, 172, 239

McConnachie A. W., 2012, AJ, 144, 4

McLure R. J. et al., 2013, MNRAS, 432, 2696

Monaco P., Benson A. J., De Lucia G., Fontanot F., Borgani S., Boylan-Kolchin M., 2014, MNRAS, 441, 2058

Monaco P., Fontanot F., Taffoni G., 2007, MNRAS, 375, 1189

Moore B., Ghigna S., Governato F., Lake G., Quinn T., Stadel J., Tozzi P., 1999, ApJ, 524, L19

Murray N., 2011, ApJ, 729, 133

Navarro J. F., Frenk C. S., White S. D. M., 1996, ApJ, 462, 563

Navarro J. F., Frenk C. S., White S. D. M., 1997, ApJ, 490, 493

Nelson D., Genel S., Pillepich A., Vogelsberger M., Springel V., Hernquist L., 2016, MNRAS, 460, 2881

Nelson D., Vogelsberger M., Genel S., Sijacki D., Kereš D., Springel V., Hernquist L., 2013, MNRAS, 429, 3353

Neto A. F. et al., 2007, MNRAS, 381, 1450

Norberg P. et al., 2002, MNRAS, 336, 907

Oesch P. A. et al., 2012, ApJ, 759, 135

Oesch P. A. et al., 2014, ApJ, 786, 108

Okamoto T., Gao L., Theuns T., 2008, MNRAS, 390, 920

O'Shea B. W., Nagamine K., Springel V., Hernquist L., Norman M. L., 2005, ApJS, 160, 1

Paardekooper J.-P., Khochfar S., Dalla Vecchia C., 2015, MNRAS, 451, 2544

Parkinson H., Cole S., Helly J., 2008, MNRAS, 383, 557

Penzias A. A., Wilson R. W., 1965, ApJ, 142, 419

Perlmutter S. et al., 1999, ApJ, 517, 565

Peterson B. M. et al., 1999, ApJ, 510, 659

Planck Collaboration et al., 2014, A&A, 571, A16

Planck Collaboration et al., 2015, ArXiv:1502.01589

Porter L. A., Somerville R. S., Primack J. R., Johansson P. H., 2014, MNRAS, 444, 942

Raičević M., Theuns T., Lacey C., 2011, MNRAS, 410, 775

Rees M. J., Ostriker J. P., 1977, MNRAS, 179, 541

Riess A. G. et al., 1998, AJ, 116, 1009

Rosdahl J., Schaye J., Dubois Y., Kimm T., Teyssier R., 2017, MNRAS, 466, 11

Rozo E. et al., 2010, ApJ, 708, 645

Rubin V. C., Ford, Jr. W. K., Thonnard N., 1980, ApJ, 238, 471

Schaller M. et al., 2015a, MNRAS, 451, 1247

Schaller M. et al., 2015b, MNRAS, 452, 343

Schaye J. et al., 2015, MNRAS, 446, 521

Schenker M. A. et al., 2013, ApJ, 768, 196

Sharma M., Theuns T., Frenk C., Bower R., Crain R., Schaller M., Schaye J., 2016, MNRAS, 458, L94

Shen S., Mo H. J., White S. D. M., Blanton M. R., Kauffmann G., Voges W., Brinkmann J., Csabai I., 2003, MNRAS, 343, 978

Shull J. M., Harness A., Trenti M., Smith B. D., 2012, ApJ, 747, 100

Silk J., 1977, ApJ, 211, 638

Somerville R. S., Hopkins P. F., Cox T. J., Robertson B. E., Hernquist L., 2008, MNRAS, 391, 481

Somerville R. S., Primack J. R., 1999, MNRAS, 310, 1087

Springel V., 2010, MNRAS, 401, 791

Springel V. et al., 2005, Nature, 435, 629

Springel V., White S. D. M., Tormen G., Kauffmann G., 2001, MNRAS, 328, 726

Sutherland R. S., Dopita M. A., 1993, ApJS, 88, 253

Thoul A. A., Weinberg D. H., 1996, ApJ, 465, 608

Toomre A., Toomre J., 1973, Scientific American, 229, 38

Vanzella E. et al., 2010, ApJ, 725, 1011

Vargas L. C., Geha M., Kirby E. N., Simon J. D., 2013, ApJ, 767, 134

Vikhlinin A. et al., 2009, ApJ, 692, 1060

Viola M., Monaco P., Borgani S., Murante G., Tornatore L., 2008, MNRAS, 383, 777

Vitvitska M., Klypin A. A., Kravtsov A. V., Wechsler R. H., Primack J. R., Bullock J. S., 2002, ApJ, 581, 799

Vogelsberger M. et al., 2014, Nature, 509, 177

Wagoner R. V., 1973, ApJ, 179, 343

Warren M. S., Quinn P. J., Salmon J. K., Zurek W. H., 1992, ApJ, 399, 405

White S. D. M., 1996, in Gravitational dynamics, Lahav O., Terlevich E., Terlevich R. J., eds., p. 121

White S. D. M., Frenk C. S., 1991, ApJ, 379, 52

White S. D. M., Rees M. J., 1978, MNRAS, 183, 341

Yoshida N., Stoehr F., Springel V., White S. D. M., 2002, MNRAS, 335, 762

Appendix A

Approximate recursive equation for $E_{\rm cool}$

Here we consider the change of $E_{\rm cool}$ in a time step $(t, t+\Delta t]$, and derive an approximate equation that relates $E_{\rm cool}(t)$ and $E_{\rm cool}(t+\Delta t)$. This equation can then be used to calculate $E_{\rm cool}$ at any given time recursively from the initial time $t_{\rm init}$.

Within this time step, the hot gas halo is treated as fixed, with its inner and outer boundaries respectively at $r_{\rm cool,pre}(t)$ and $r_{\rm vir}(t)$. By $t+\Delta t$, the gas between $r_{\rm cool,pre}(t)$ and $r_{\rm cool}(t+\Delta t)$ has cooled down.

From Eq(4.2.12), one has

$$E_{\rm cool}(t+\Delta t) = 4\pi \int_{t_{\rm init}}^{t+\Delta t} \int_{r'_{\rm p}(\tau)}^{r_{\rm vir}(\tau)} \tilde{\Lambda}\rho^2_{\rm hot}(r,\tau) r^2 dr d\tau, \tag{A.0.1}$$

where $\tilde{\Lambda}$ is the cooling function, $\rho_{\rm hot}(r,\tau)$ is the density of the hot gas at radius r and time τ, and $r'_{\rm p}(\tau)$ is the radius at τ of a shell that has radius $r_{\rm cool}$ at $t+\Delta t$. Note that here we use $r_{\rm cool}$ instead of $r_{\rm cool,pre}(t+\Delta t)$, because the hot halo is fixed here, and in the new cooling model, after the cooling calculation, the halo contraction would change $r_{\rm cool}(t+\Delta t)$ to $r_{\rm cool,pre}(t+\Delta t)$.

Then Eq(A.0.1) can be further expanded as

$$\begin{aligned} E_{\rm cool}(t+\Delta t) &= 4\pi \int_{t_{\rm init}}^{t+\Delta t} \int_{r_{\rm p}(\tau)}^{r_{\rm vir}(\tau)} \tilde{\Lambda}\rho^2_{\rm hot}(r,\tau) r^2 dr d\tau \\ &- 4\pi \int_{t_{\rm init}}^{t+\Delta t} \int_{r_{\rm p}(\tau)}^{r'_{\rm p}(\tau)} \tilde{\Lambda}\rho^2_{\rm hot}(r,\tau) r^2 dr d\tau \\ &= I_1 - I_2, \end{aligned} \tag{A.0.2}$$

where I_1 and I_2 represent respectively the two integrals in the above equation, and $r_{\rm p}(\tau)$ is the radius at τ of a shell that has radius $r_{\rm cool,pre}$ at $t+\Delta t$. Note that at $t+\Delta t$ the hot gas halo inner boundary is still at $r_{\rm cool,pre}(t)$ because the halo is assumed to be static over the interval $(t, t+\Delta t]$. Further

$$\begin{aligned} I_1 &= 4\pi \int_{t_{\rm init}}^{t} \int_{r_{\rm p}(\tau)}^{r_{\rm vir}(\tau)} \tilde{\Lambda}\rho^2_{\rm hot}(r,\tau) r^2 dr d\tau \\ &+ 4\pi \int_{t}^{t+\Delta t} \int_{r_{\rm p}(\tau)}^{r_{\rm vir}(\tau)} \tilde{\Lambda}\rho^2_{\rm hot}(r,\tau) r^2 dr d\tau \\ &= E_{\rm cool}(t) + \Delta t \times 4\pi \int_{r_{\rm cool,pre}}^{r_{\rm vir}} \tilde{\Lambda}\rho^2_{\rm hot}(r,t) r^2 dr \\ &= E_{\rm cool}(t) + L_{\rm cool}(t)\Delta t, \end{aligned} \tag{A.0.3}$$

in which we have used Eq(4.2.12) for the first integral in the above equation, while the second integral is simplified by the assumption that the hot gas halo is fixed within $(t, t+\Delta t]$, with the inner and outer boundaries $r_{\rm cool,pre}$ and $r_{\rm vir}$ respectively, and $L_{\rm cool}(t)$ is defined in Eq(4.2.11).

I_2 can be further written as

$$\begin{aligned} I_2 &= 4\pi \int_{t}^{t+\Delta t} \int_{r_{\rm p}(\tau)}^{r'_{\rm p}(\tau)} \tilde{\Lambda}\rho^2_{\rm hot}(r,\tau) r^2 dr d\tau \\ &+ 4\pi \int_{t_{\rm init}}^{t} \int_{r_{\rm p}(\tau)}^{r'_{\rm p}(\tau)} \tilde{\Lambda}\rho^2_{\rm hot}(r,\tau) r^2 dr d\tau \\ &= L'_{\rm cool}\Delta t + I_3, \end{aligned} \tag{A.0.4}$$

where $L'_{\rm cool}(t)$ is defined in Eq(4.2.15), and the first integral in the above equation is simplified again because the hot gas halo is assumed fixed within $(t, t+\Delta t]$, while I_3 corresponds to the second integral above.

The integral I_3 represents the total energy radiated away by the gas within $r_{\rm cool,pre} \le r \le r_{\rm cool}$ from $t_{\rm init}$ to t, and it can be rewritten as the summation of $\delta E_{\rm cool}$ of each gas shell in this range, namely

$$\begin{aligned} I_3 &= \int_{r_{\rm cool,pre}}^{r_{\rm cool}} \frac{\delta E_{\rm cool}}{\delta r} dr \\ &= \int_{r_{\rm cool,pre}}^{r_{\rm cool}} \frac{\delta E_{\rm cool}}{\delta L_{\rm cool}} \frac{\partial L_{\rm cool}}{\partial r} dr \\ &= \int_{r_{\rm cool,pre}}^{r_{\rm cool}} t_{\rm cool,avail}(r,t) \frac{\partial L_{\rm cool}}{\partial r} dr, \end{aligned} \tag{A.0.5}$$

in which we derive the third line from the second line by virtue of the definition of $t_{\rm cool,avail}$ for an individual gas shell given in Eq(4.2.7).

Now consider that in the slow cooling regime, typically $r_{\rm cool}$ is close to $r_{\rm cool,pre}$, thus the radial dependence in $t_{\rm cool,avail}(r,t)$ can be ignored, while in the fast cooling regime, although $r_{\rm cool}$ could be much larger than $r_{\rm cool,pre}$, the cooling is so fast that halo growth and hot gas halo contraction only have weak effects on the cooling, and thus can only introduce weak dependence of $t_{\rm cool,avail}$ on r. In all, for $r_{\rm cool,pre}\leq r\leq r_{\rm cool}$, we can approximate $t_{\rm cool,avail}(r,t)\approx t_{\rm cool,avail}(r_{\rm cool},t)$, so

$$\begin{aligned} I_3 &\approx t_{\rm cool,avail}(r_{\rm cool},t)\int_{r_{\rm cool,pre}}^{r_{\rm cool}}\frac{\partial L_{\rm cool}}{\partial r}dr \\ &= t_{\rm cool,avail}(r_{\rm cool},t)L'_{\rm cool}, \end{aligned} \tag{A.0.6}$$

where $L'_{\rm cool}(t)$ is defined in Eq(4.2.15) and is the total cooling luminosity at t of the gas between $r_{\rm cool,pre}\leq r\leq r_{\rm cool}$. Note that for hot gas halo that is fixed at all times, this approximation becomes exact. Based on this, one has

$$\begin{aligned} I_2 &\approx [\Delta t+t_{\rm cool,avail}(r_{\rm cool},t)]L'_{\rm cool} \\ &= t_{\rm cool,avail}(r_{\rm cool},t+\Delta t)L'_{\rm cool}, \end{aligned} \tag{A.0.7}$$

in which we have used Eq(4.2.8).

Substituting Eqs.(A.0.3) and (A.0.7) into Eq(A.0.2), one reaches the approximate recursive equation for $E_{\rm cool}$, i.e. Eq(4.2.14).

Appendix B

Approximate calculation of change of angular momentum distribution of hot gas halo

B.1 Approximate calculation of $j_{\rm hot}[r(r')]$

In the new cooling model, the hot gas halo evolves with the growth of dark matter halo, and it also contracts in response to gas cooling removing pressure support from the central regions. These effects changes the specific angular momentum distribution of the hot gas halo. We assume that each spherical shell of hot gas conserves its specific angular momentum $j_{\rm hot}$ during this change, but the shell moves from r to r', and thus the angular momentum profile changes from $j_{\rm hot}(r)$ to $j'_{\rm hot}(r') = j_{\rm hot}[r(r')]$. As described in §4.2.1.4, $r(r')$ can be determined through Eq(4.2.22), which is based on mass conservation for each shell. For the assumed form of the hot gas density profile, this equation does not have an explicit analytical solution, leading to no exact analytical expression for $j_{\rm hot}[r(r')]$. While $j_{\rm hot}[r(r')]$ can be derived numerically for each shell at every time step, this is computationally expensive, and so here we present an approximate analytical expression that can be used instead. In Appendix B.2, we test the accuracy of this analytical approximation against a numerical solution of the same equations.

At the end of a time step, the hot gas is distributed between $r_{\rm cool}$ and $r_{\rm vir}$,

following a β-distribution with core radius $r_{\rm core}$. Before the calculation of the next time step, the effects of halo growth and hot gas halo contraction during the current time step should be included. These effects redistribute the hot gas previously in this halo. The inner boundary of the hot gas moves from $r_{\rm cool}$ to $r_{\rm cool,pre}$, while the outer boundary moves from $r_{\rm vir}$ to $r'_{\rm vir}$. According to the assumptions in §4.2.1.2, this adjusted gas still follows a β-distribution, but with a new core radius $r'_{\rm core}$. As mentioned in §4.2.1.4, this adjustment is only for the hot gas previously in this halo, while the newly added hot gas is assumed to mix with the hot gas halo after this adjustment. Therefore the total gas mass, $M_{\rm hot}$, before and after this adjustment is unchanged.

When considering the approximate calculation of $j_{\rm hot}[r(r')]$, it is more convenient to work with the variables $x \equiv r/r_{\rm core}$ and $x' \equiv r'/r'_{\rm core}$ instead of r and r'. Then the angular momentum profile after the adjustment of the hot gas halo can be written as $j_{\rm hot}[x(x')]$. The function $x(x')$ can be derived from Eq(4.2.22), which can be further written as

$$M_{\rm hot}(<x) = M'_{\rm hot}(<x'), \tag{B.1.1}$$

where $M_{\rm hot}(<x)$ is the mass of hot gas within radius x according to the density profile before the adjustment induced by hot gas halo contraction and dark matter halo growth, while $M'_{\rm hot}(<x')$ is the mass of hot gas within x' according to the density profile after this adjustment. As mentioned above, x and x' are respectively the radii of the same Lagrangian shell before and after this adjustment. Note that at the inner boundary the above equation satisfies the condition $M_{\rm hot}(<x_0) = M'_{\rm hot}(<x'_0) = 0$, where $x_0 = r_{\rm cool}/r_{\rm core}$ and $x'_0 = r_{\rm cool,pre}/r'_{\rm core}$, while at the outer boundary it satisfies $M_{\rm hot}(<x_{\rm vir}) = M'_{\rm hot}(<x'_{\rm vir}) = M_{\rm hot}$, where $x_{\rm vir} = r_{\rm vir}/r_{\rm core}$ and $x'_{\rm vir} = r'_{\rm vir}/r'_{\rm core}$.

According to the assumed β-distribution, one has

$$M_{\rm hot}(<x) = \frac{M_{\rm hot}}{Y_{\rm vir} - Y_0}[x - \arctan(x) - Y_0], \tag{B.1.2}$$

where $Y_{\rm vir} = x_{\rm vir} - \arctan(x_{\rm vir})$ and $Y_0 = x_0 - \arctan(x_0)$. Similarly, for the hot gas halo after the adjustment, one has

$$M'_{\rm hot}(<x') = \frac{M_{\rm hot}}{Y'_{\rm vir} - Y'_0}[x' - \arctan(x') - Y'_0], \tag{B.1.3}$$

where $Y'_{\rm vir} = x'_{\rm vir} - \arctan(x'_{\rm vir})$ and $Y'_0 = x'_0 - \arctan(x'_0)$.

Substituting Eq(B.1.2) and (B.1.3) into Eq(B.1.1), one derives an implicit form for the function $x(x')$

$$\begin{aligned} x - \arctan(x) &= \frac{Y_{\rm vir} - Y_0}{Y'_{\rm vir} - Y'_0}[x' - \arctan(x')] \\ &+ \frac{Y'_{\rm vir}Y_0 - Y_{\rm vir}Y'_0}{Y'_{\rm vir} - Y'_0}. \end{aligned} \tag{B.1.4}$$

Eqn.(B.1.4) does not allow an explicit analytical expression for $x(x')$. However, it is still possible to construct simple analytical approximations for $x(x')$ in different ranges of x', and so to derive analytical approximations for $j_{\rm hot}[x(x')]$.

First note that typically $x' \leq x$, because the contraction moves shells from large radii to small radii. When x' is large, then both $x - \arctan(x)$ and $x' - \arctan(x')$ can be well approximated by linear functions. These linear functions then lead to a linear functional form for $j_{\rm hot}[x(x')]$. This linear functional form can be kept during the recursion procedure, which is necessary for deriving the specific angular momentum distribution from its initial value, so for large enough x', $j_{\rm hot}[x(x')]$ can always be expressed as a linear function of x'.

On the other hand, when x' is very close to 0, $x' - \arctan(x') = x'^3/3 - x'^5/5 + O(x'^7)$ according to its Taylor expansion. Note that this typically happens in the slow cooling regime, in which the cooling is limited to the central region of the halo and the induced contraction of the hot gas halo is small in each time step, so typically in this case x is also close to 0, and the Taylor expansion is also a good approximation for $x - \arctan(x)$, i.e. $x - \arctan(x) = x^3/3 - x^5/5 + O(x^7)$. These nonlinear terms in the Taylor expansions cause $j_{\rm hot}$ to gradually deviate from the assumed linear form before the starting of cooling. The nonlinear terms in the Taylor expansions are third and fifth order, this inspires the following expression for $j_{\rm hot}[x(x')]$

$$j_{\rm hot}[x(x')] = c_1 x'^6 + c_2 x'^5 + c_3 x'^3 + c_4 x' + c_5, \tag{B.1.5}$$

where $c_1 - c_5$ are coefficients and we include all terms with orders lower than $O(x'^7)$ that can be generated by the third and fifth order terms, while the linear term is added to include the initial linear form of the angular momentum profile.

When x (and also x') are either not very large or not close to 0, the function $x - \arctan(x)$ has a nonlinear-dependence, but not so strong as in the case when x is close to 0. Thus generally speaking, $j_{\rm hot}[x(x')]$ in this regime can be expressed approximately as a lower order polynomial, and here we choose a second-order polynomial.

In summary, we adopt the following piecewise function as the analytical approximation for $j_{\rm hot}[x(x')]$

$$j_{\rm hot}[x(x')] = \begin{cases} a_1 x' + a_2, & x' \geq 3.5 \\ a_3 x' + a_4, & 2.0 \leq x' < 3.5 \\ a_5 x'^2 + a_6 x' + a_7, & 0.5 \leq x' < 2.0 \\ a_8 x'^6 + a_9 x'^5 + a_{10} x'^3 & \\ + a_{11} x' + a_{12}, & 0.0 \leq x' < 0.5 \end{cases} \tag{B.1.6}$$

where $a_1 - a_{12}$ are coefficients, with the coefficients in Eq(B.1.5) to be renamed as $a_8 - a_{12}$.

The procedure is then as follows. At each time step, several sample points are taken over the whole range of x', and then Eq(B.1.4) is solved numerically for these sample points to find the corresponding x, with the specific angular momentum distribution in the last time step, $j_{\rm hot}[x(x')]$ being known for these sample points. Using these values, Eq(B.1.6) then becomes a set of linear equations for the coefficients $a_1 - a_{12}$, which can be solved easily. Once these coefficients are determined, then the approximate $j_{\rm hot}[r(r')]$ can be calculated for any value of r' for the current time step. Then the contribution from the newly added gas, $j_{\rm new}(r')$, can be added as described in §4.2.1.4. Because it is assumed that $j_{\rm new}(r') \propto r'$, so this further changes the coefficients of the first and zeroth order terms in Eq(B.1.6). After this, the angular momentum profile of this time step is fully determined.

This approximation requires 9 sample points for determining $a_1 - a_{12}$ (two adjacent x' sections share one common sample point), and so Eq(B.1.4) needs to be solved for $x(x')$ only 9 times at each timestep. An alternative to this approximate method would be to numerically evaluate $j_{\rm hot}[r(r')]$ on a radius grid, which would require solving Eq(B.1.4) at each radius grid point, rather than at a handful of sample points. So the approximate method is much computationally faster than the

straightforward radius grid method. Also, the approximate method only requires storing the 12 coefficients, while the radius grid method requires storing the whole radius grid and the numerical $j_{\rm hot}[r(r')]$ on it, and thus would require much more computer memory.

B.2 Comparison with direct calculation

To assess the accuracy of the approximation introduced in the previous section, we compared the angular momentum accretion rates for central galaxies calculated using this approximation with those calculated using a direct (but more more computationally intensive) calculation. This direct calculation evaluates $j_{\rm hot}(r)$ numerically on a radius grid at each time step. The radius grid covers the range between $r_{\rm cool,pre}$ and $r_{\rm vir}$ with 1000 grid points. $j_{\rm hot}(r)$ at a given time step is calculated from $j_{\rm hot}(r)$ at the previous time step by solving Eq(B.1.4) for each grid point, and then using Eq(4.2.25).

The comparison is done for three cases. The first one is for static halos, with no feedback. The second is for dynamically evolving halos, including full halo merger histories, but still without any feedback. The third case is also for dynamically evolving halos, but with strong supernova feedback. Here the supernova feedback is modeled as usual in GALFORM with a mass ejection rate from the galaxy into the ejected gas reservoir $\dot{M}_{\rm eject} = \beta\psi$, where ψ is the star formation rate and the mass-loading factor $\beta = (V_{\rm c}/V_{\rm SN})^{-\gamma_{\rm SN}}$, with $V_{\rm c}$ being the circular velocity of the galaxy and $V_{\rm SN}$ and $\gamma_{\rm SN}$ being parameters. For the calculations here, we use $V_{\rm SN} = 320\,{\rm km\,s^{-1}}$ and $\gamma_{\rm SN} = 3.2$, which are close to the values adopted in recent versions of GALFORM. The calculations are done for four different halo masses at $z = 0$, namely $M_{\rm halo} = 10^{11}$, 10^{12}, 10^{13} and $10^{14}\,{\rm M}_\odot$, which covers both the fast and slow cooling regimes. For the dynamically evolving halos, results are calculated for 100 Monte Carlo merger trees for each halo mass.

For each of these cases, the angular momentum accretion rate onto the central galaxy due to the cooling flow, $\dot{J}_{\rm cool}$, is calculated at each timestep, both for the approximate method in Appendix B.1 ($\dot{J}_{\rm cool,app}$) and for the direct cal-

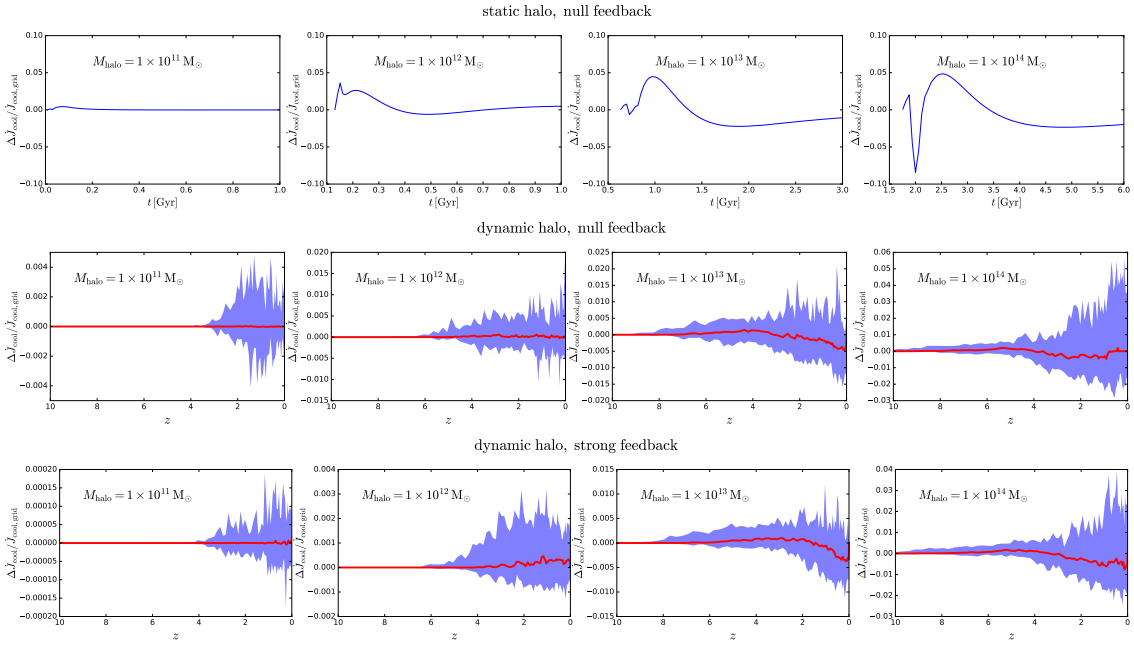

Figure B.1: The relative error in the angular momentum accretion rate calculated using the approximate method for evolving $j_{\rm hot}(r)$ compared with that obtained from the direct calculation. Results are shown for 3 cases (static halo without feedback, dynamically evolving halo without feedback, and dynamically evolving halo with strong supernova feedback) and 4 different halo masses at $z = 0$ (10^{11}, 10^{12}, 10^{13} and $10^{14}\,{\rm M}_\odot$). The dynamic halo cases use full halo merger histories, with 100 Monte Carlo merger trees for each halo mass. For the dynamic halo cases, in each panel the solid line shows the median of the relative error, while the shaded region indicates the $5-95\%$ range. See text for more details.

culation ($\dot{J}_{\rm cool,grid}$). The relative error is then calculated as $\Delta\dot{J}_{\rm cool}/\dot{J}_{\rm cool,grid}$, with $\Delta\dot{J}_{\rm cool} = \dot{J}_{\rm cool,app} - \dot{J}_{\rm cool,grid}$. Fig. B.1 shows this relative error for the three cases and the four different halo masses. From this figure, it can be seen that the relative error is generally less than 10%, so the approximate method works well.

Appendix C

Random walk model for evolution of $\lambda_{\rm halo}$

C.1 Random walk model of halo spin evolution

The evolution of halo spin results from the the angular momentum and mass brought into the halo by accretion and mergers. This angular momentum depends on the tangential component of the infall velocity. A simple model for the halo growth is to assume that these accretion/merger events are random, with random infall velocities. In this case, the evolution of the halo spin accompanying the mass accretion is a kind of random walk (e.g. Vitvitska et al., 2002). For simplicity, we further assume that this random walk for the halo spin is a Markov walk, meaning that each step is statistically independent of previous steps.

In this picture, the spins of the descendant halo and its major progenitors are related by a conditional spin distribution, which gives the probability density for any given descendant spin value given the spin and mass accretion history of the progenitor. We now derive the form of this probability distribution for some plausible assumptions.

C.2 Conditional distribution of descendant halo spin

Mathematically, a random walk is described as a sequence of random variables, $Y(x)$, where x is the sequence index and $Y(x)$ is the random variable at x, with its possible value y and corresponding probability distribution $P(y, x)$. For the random walk considered here, we choose $x = \ln(M_{\rm halo}/M_{\rm i})$, where $M_{\rm halo}$ is the mass of a given halo, and $M_{\rm i}$ is its initial mass. We choose this form because it gives the same Δx whenever the halo mass has increased by a certain factor, and we expect that the change of halo spin is more closely related to the fractional increase in halo mass than to the absolute increase in mass.

N-body simulations of the formation of dark matter halos by hierarchical clustering show that the distribution of $\lambda_{\rm halo}$ is well approximated by a lognormal, with median $\lambda_{\rm med}$ and dispersion σ_λ in $\ln\lambda_{\rm halo}$ that are almost independent of the halo mass and cosmological parameters (e.g. Bett et al., 2007). Motivated by this, we define $Y = [\ln(\lambda_{\rm halo}) - \ln(\lambda_{\rm med})]/\sigma_\lambda$.

For a Markov random walk, $P(y, x)$ is approximately described by the Fokker-Planck equation: [1]

$$\frac{\partial P}{\partial x} = -\frac{\partial}{\partial y}[a_1 P] + \frac{\partial^2}{\partial y^2}[a_2 P], \tag{C.2.1}$$

where a_1 and a_2 are two functions of y and x. Given the results for the spin distribution described above, we want Eq(C.2.1) to have a steady-state asymptotic solution $P(y, x) = 1/\sqrt{2\pi}\exp(-y^2/2)$, which corresponds to a lognormal distribution for $\ln\lambda_{\rm halo}$ with parameters that do not depend on $M_{\rm halo}$. For simplicity, we assume a_2 is a constant. The requirement that $P(y, x) = 1/\sqrt{2\pi}\exp(-y^2/2)$ is a steady-state solution then leads to the relation $a_1 = -a_2 y + c_0\exp(y^2/2)$, with c_0 a constant. However, the term $c_0\exp(y^2/2)$ provides a drag towards $y = +\infty$, which in terms of spin evolution is a trend for $\lambda_{\rm halo}$ to become arbitrarily large, and this is unphysical,

[1] Strictly speaking, Fokker-Planck equation is not valid for an arbitrarily sharp distribution like our initial condition $P(y, 0) = \delta(y - y_0)$, but this distribution would be broadened quickly by diffusion, thus the Fokker-Planck equation is expected to still be valid for the moments not too close to the initial moment.

so we set $c_0 = 0$, leading to $a_1 = -a_2 y$. In terms of the random trajectories $Y(x)$, the first term on the right hand side of Eq(C.2.1) then represents a mean shift back towards $Y = 0$, while the second term represents a diffusion of Y.

With these choices for a_1 and a_2, the Fokker-Planck equation has the following analytical solution (see e.g. Garcia-Palacios (2007) for details) for the initial condition $P(y, 0) = \delta(y - y_0)$:

$$P(y, x|y_0, 0) = \frac{1}{\sqrt{2\pi(1 - e^{-2x/\tau})}} \exp\left[\frac{(y - y_0 e^{-x/\tau})^2}{2(1 - e^{-2x/\tau})}\right], \tag{C.2.2}$$

where $\tau = 1/a_2$ and $P(y, x|y_0, 0)$ is the conditional distribution of y given $y = y_0$ at $x = 0$.

Here τ serves as a relaxation scale for the variable x, with the solution having roughly relaxed to the steady solution for $x = \tau$. We choose $\tau = \ln 2$, so that the correlation between the spin of a halo and its progenitor is nearly disappears when it becomes twice as massive as the progenitor. This value for τ was originally chosen to approximately match the assumption made in earlier GALFORM models that a new spin is assigned randomly at every halo formation event, defined as happening whenever the halo mass has increased by a factor 2. However, we show below that this choice for τ produces results for the spin evolution in quite good agreement with N-body simulations. With the parameter τ fixed, and the definitions of Y and x, it is straightforward to derive the corresponding conditional distribution for $\lambda_{\rm halo}$, with which a halo's spin can be assigned given its progenitor spin and mass growth history.

C.3 Comparison with N-body simulations

We test our simple random walk model for the evolution of $\lambda_{\rm halo}$ by comparing its predictions with results from Vitvitska et al. (2002), measured for halos in cosmological N-body simulations. Fig. 4 in Vitvitska et al. shows the conditional probability distribution of $\lambda_{\rm halo}$ for several ranges of initial spin and halo mass growth. Specifically, they show three ranges for the initial spin, $\lambda_{\rm i}$, namely $\lambda_{\rm i} < 0.025$, $0.025 < \lambda_{\rm i} < 0.055$ and $\lambda_{\rm i} > 0.055$, and three ranges for the mass growth, which

are respectively $M_{\rm f}/M_{\rm i} < 1.1$, $1.1 < M_{\rm f}/M_{\rm i} < 1.25$ and $M_{\rm f}/M_{\rm i} > 1.25$, with $M_{\rm f}$ the halo mass after growth and $M_{\rm i}$ the mass before growth. In order to make a simple comparison between the results of Vitvitska et al. and the predictions from our random walk modeling, we estimate the typical value for each $\lambda_{\rm i}$ and $M_{\rm f}/M_{\rm i}$ range, and then calculate the conditional probability distribution using Eq(C.2.2).

We choose $\lambda_{\rm i} = 0.019, 0.038, 0.08$ as typical values for the three ranges $\lambda_{\rm i} < 0.025$, $0.025 < \lambda_{\rm i} < 0.055$ and $\lambda_{\rm i} > 0.055$ respectively. These are the means over the corresponding ranges according to the lognormal distribution of $\lambda_{\rm halo}$ measured from the same simulation.

For the mass ratio $M_{\rm f}/M_{\rm i}$, we set $M_{\rm f}/M_{\rm i} = 1$ as its lower limit, which means that the halo mass is not allowed to decrease, while $M_{\rm f}/M_{\rm i} = 2$ is set as the upper limit. This is because Vitvitska et al. always measure the change of halo spin between two adjacent N-body snapshots, between which the physical time duration is relatively short, large values of $M_{\rm f}/M_{\rm i}$ should be caused by major mergers instead of smooth accretion, and the number of major mergers for a halo should be at most one in this short time duration. Thus the three ranges of $M_{\rm f}/M_{\rm i}$ in Vitvitska et al. become $1 < M_{\rm f}/M_{\rm i} < 1.1$, $1.1 < M_{\rm f}/M_{\rm i} < 1.25$ and $1.25 < M_{\rm f}/M_{\rm i} < 2$ respectively. We take the geometric mean of the range boundaries as the typical value for the corresponding mass range, and this leads to $M_{\rm f}/M_{\rm i} = 1.049, 1.173, 1.581$ for the three ranges.

Using these estimated typical values, the corresponding conditional distributions can be calculated for the random walk model. Fig. C.1 shows the comparison between the predictions of our simple random walk model and the results measured by Vitvitska et al. from their N-body simulations. The agreement is acceptable for a simple comparison.

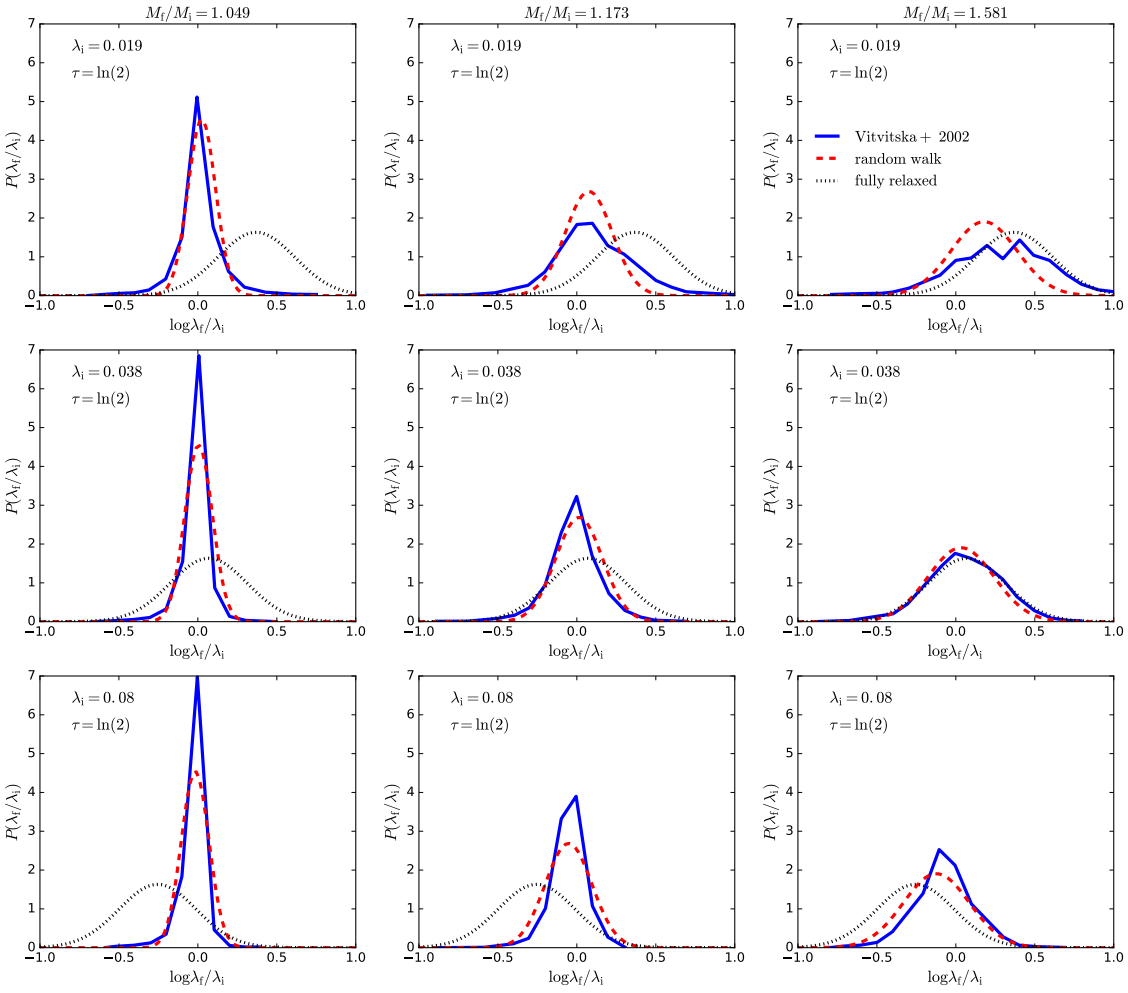

Figure C.1: Comparison of the conditional halo spin distributions predicted by our random walk model with measurements from N-body simulation in Vitvitska et al. (2002). The nine panels correspond to those in Fig. 4 Vitvitska et al.. Each row corresponds to a range of the initial spin λ_i, with our estimated typical λ_i for that range given in the upper left corner of each panel. Each column corresponds to a range of the ratio M_f/M_i, with M_f and M_i being the halo masses at adjacent snapshots, and our estimated typical M_f/M_i being shown at the top of the column. In each panel, the blue solid line is the conditional spin distribution from Vitvitska et al., the red dashed line is the distribution calculated from Eq(C.2.2) based on our random walk model, and the black dotted line shows the fully relaxed distribution expected in the random walk model for reference. The relaxation scale τ is also given in the upper left corner of each panel.

Appendix D

Simple AGN feedback model in GALFORM

The AGN feedback model used in the Lacey16 model was first introduced in Bower et al. (2006). Specifically, it assumes that the AGN feedback is in the radio mode (e.g. Croton et al., 2006), in which a relativistic jet generated by supermassive black hole (SMBH) accretion heats the halo gas and thus suppresses cooling.

In GALFORM there are two conditions for an effective AGN feedback. Firstly, the halo gas should be close to the slow cooling regime, in which the cooling is slower than the gravitational infall and a quasi-hydrostatic hot gaseous halo exists. This is motivated by the idea that only the gas close this regime can maintain its pressure and thus the jet can interact and heat the halo gas effectively. This condition is tested by comparing the cooling time scale, $t_{\rm cool}$, and the free-fall time scale, $t_{\rm ff}$, at the cooling radius $r_{\rm cool}$. Specifically, AGN feedback is assumed to be effective only if

$$t_{\rm cool}(r_{\rm cool})/t_{\rm ff}(r_{\rm cool}) > 1/\alpha_{\rm cool}, \tag{D.0.1}$$

with $\alpha_{\rm cool} \sim 1$ an adjustable parameter. Consider that at earlier times the ratio $t_{\rm cool}(r_{\rm cool})/t_{\rm ff}(r_{\rm cool})$ is typically smaller, then increasing $\alpha_{\rm cool}$ causes AGN feedback to turn on earlier and thus enhances the suppression due to this feedback.

Secondly, the SMBH accretion rate should be significantly lower than the Eddington limit so that jets can be efficiently produced (Fanidakis et al., 2011), and the jet should be energetic enough to balance the cooling radiation. This motivates

the following condition

$$f_{\rm Edd} L_{\rm Edd}(M_{\rm BH}) > L_{\rm cool}, \tag{D.0.2}$$

where $f_{\rm Edd} \ll 1$ is a parameter, $L_{\rm Edd}(M_{\rm BH})$ is the Eddington luminosity of a black hole with mass $M_{\rm BH}$, and $L_{\rm cool}$ is the cooling luminosity of the hot gas halo. In the Lacey16 model, $f_{\rm Edd} = 0.01$.

Once the above two conditions are satisfied, the AGN feedback is assumed to be effective. In the GFC1 model, the increase of $r_{\rm cool}$ due to cooling is then set to zero, and then the associated mass and angular momentum cooling rates become zero.

In the new cooling model, because a different procedure is used to calculate $t_{\rm cool,avail}$, some modifications are needed. Specifically, when AGN feedback turns on, the energy previously radiated away, $E_{\rm cool}$, is set to zero because the halo gas is heated up. This causes $t_{\rm cool,avail}$ to reduce to zero. With this, $r_{\rm cool}$ does not increase and the halo cold gas component stops growing immediately. If this component has nonzero mass, then it can still deliver cold gas to the central galaxy. When a halo is close to the slow cooling regime, the halo cold gas component typically is very small, so the cold gas accretion onto the central galaxy should stop shortly after AGN feedback turns on.