6

The galaxy correlation function as a constraint on galaxy formation physics

Semi-analytical models of galaxy formation are generally successful in reproducing the number densities of galaxies as a function of mass. In order to remove possible degeneracies and improve the model, having additional orthogonal constraints like clustering data while exploring parameter space would be useful. However, this is challenging due to the two-point nature of such quantities, which makes using them as a constraint computationally very expensive, as the model would have to be run on the full halo catalogue at every step. Here, we present a fast estimator for the projected galaxy correlation function that produces $\sim 10\%$ accurate results using only a very small subsample of haloes. As a first application, we incorporate it in a recent version of the Munich semi-analytical model and find a set of galaxy formation parameters that simultaneously reproduces the observed $z = 0$ stellar mass function and clustering data from SDSS.

Marcel P. van Daalen, Bruno M. B. Henriques,
Raul E. Angulo and Simon D. M. White
In preparation

6.1 Introduction

Galaxy formation is currently an unsolved problem. Because of this, any model of galaxy formation – be it hydrodynamical, analytical or semi-analytical in nature – has to rely on some set of observations in order to constrain the parameters of the physical processes that cannot be derived from first principles, or be simulated directly.

Hydrodynamical simulations can simulate baryonic processes directly on large scales while relying on sub-grid recipes to model relevant processes below the resolution limit. As such simulations are relatively expensive computationally, the values of the parameters in the sub-grid formulations usually have to be informed by comparing a set of simulations run at lower resolution or in smaller volumes to some observational quantity, though these numerical settings themselves may impact which parameter values are "right" for it. Still, as the available computational resources are ever growing, the number of processes which cannot be simulated directly is slowly decreasing (e.g. Hopkins et al., 2013), and valiant efforts are currently being made to improve the accuracy of direct cosmological simulations (e.g. EAGLE, Schaye et al., in preparation).

Semi-analytical models (hereafter SAMs), on the other hand, necessarily include more physical parameters to calibrate, as baryonic processes are not simulated directly on any scale. However, once the high-resolution collisionless simulations that they are based on have been run a single time, they can be repeated many times with different parameter values at low computational cost. Coupled with a method to efficiently explore parameter space such as Monte Carlo Markov Chains (MCMC, for a review on this and similar methods see Trotta, 2008), this allows one to find the highest-likelihood set of parameters for any given model, based on a set of observational constraints.

Typically, SAMs use observational data sets of one-point functions, such as stellar mass or luminosity functions, as constraints for their model parameters (e.g. Kauffmann, White & Guiderdoni 1993; Baugh, Cole & Frenk 1996; Somerville & Primack 1998; Kauffmann et al. 1999; Cole et al. 2000; Croton et al. 2006; Bower et al. 2006; Monaco, Fontanot & Taffoni 2007; Somerville et al. 2008; Henriques et al. 2009; Guo et al. 2011; Henriques et al. 2013, see Baugh 2006 for a review on the general methodology). The resulting models of galaxy formation can then be tested against other observables (i.e. observables that are independent of those used as constraints) and be used to make predictions for these. A delicate balance must be maintained here: if the model has too many free parameters, prior regions that are too wide, or if there are too few (independent) observational constraints, degeneracies may occur (i.e. separate regions of high likelihood in parameter space), while too little freedom or failing to include some relevant physical process may leave the model unable to match several observables at once.

SAMs generally have trouble matching the small-scale clustering of galaxies while simultaneously matching other observational constraints such as the luminosity function (e.g. Kauffmann et al. 1999; Springel et al. 2005; Li et al. 2007; Guo

et al. 2011; Kang et al. 2012; but see e.g. Kang 2014). In order to determine the cause of this discrepancy, and to test whether the models retain enough freedom to match the observed clustering at all, it would be instructive to use clustering measurements as constraints while exploring parameter space. As galaxy clustering is determined by how galaxies with different properties populate haloes of different mass, it directly constrains galaxy formation, in a way that is complementary to, for example, the luminosity function.

However, this presents a problem: while one-point functions such as the stellar mass function can be quickly estimated with known uncertainty by running the model on only a small sample of representative haloes, allowing large regions of parameter space to be rejected without having to run the model on the full dark matter simulation, the same cannot be done simply for two-point functions such as the correlation function. In principle, any observable that relies on spatial correlations between galaxies can only be calculated by running the model on the full simulation, which is computationally infeasible when thousands of models need to be explored. While running the SAM on a small sub-volume may allow one to measure small-scale correlations to some degree, cosmic variance will be an issue. Additionally, if one aims to compare to observations, where clustering is viewed in projection (unless line-of-sight velocities are used), one still has to account for large-scale correlations, even at small separations.

Here, we present a method to quickly estimate the projected correlation function, $w(r_\mathrm{p})$, to some known uncertainty from a small sample of haloes using a halo model based approach, and apply it to constrain the recent version of the Munich semi-analytical model presented in Guo et al. (2013, , hereafter G13). By measuring the properties of galaxies within individual haloes and making informed assumptions about the distribution of these haloes, we are able to circumvent the aforementioned problems, greatly reducing the CPU time needed to predict their two-point clustering.

This chapter is organised as follows. In Section 6.2, we present our method for estimating $w(r_\mathrm{p})$ and briefly describe the semi-analytical model we apply it to. Next, in Section 6.3, we show the results of using clustering as an additional constraint on parameter space, on top of the often-used $z = 0$ stellar mass function. Finally, in Section 6.4 we present a summary of our work and discuss future improvements and applications.

6.2 Method

6.2.1 Estimating the correlation function

Our approach is slightly different to that of most previous works constructing a correlation function estimator based on the halo model, where the aim is typically to reproduce observations given some halo occupation distribution (HOD). Here, our goal is instead to reproduce the results of the semi-analytical model run on the full dark matter simulation to within some given accuracy, given the galaxy

properties for a small sample of haloes. As we will show, we are able to reproduce the projected correlation function of the full galaxy sample to within about 20%, using the properties of semi-analytical galaxies occupying less than 0.04% of the full halo sample (0.14% of the subhalo sample).

6.2.1.1 The backbone of the model

Our starting point is the linear halo model, introduced independently by Seljak (2000), Ma & Fry (2000) and Peacock & Smith (2000). In what follows, we will adhere to the terminology of Cooray & Sheth (2002). In the analytical halo model the power spectrum, $P(k)$, is written as the sum of two terms:

$$P(k) = P^{\rm 1h}(k) + P^{\rm 2h}(k). \tag{6.1}$$

Here $P^{\rm 1h}(k)$ is the 1-halo term, describing the two-point clustering contribution of points within the same halo, and $P^{\rm 2h}(k)$ is the 2-halo term, describing the contribution of points within separate haloes. For the clustering of matter, these are given by:

$$\begin{aligned} P^{\rm 1h}_{\rm dm}(k) &= \int n(M) \left(\frac{M}{\bar{\rho}}\right)^2 |u(k|M)|^2 {\rm d}M \\ P^{\rm 2h}_{\rm dm}(k) &= \iint n(M_1) \left(\frac{M_1}{\bar{\rho}}\right) u(k|M_1) n(M_2) \left(\frac{M_2}{\bar{\rho}}\right) u(k|M_2) \times \\ &\quad P_{\rm hh}(k|M_1, M_2) {\rm d}M_1 {\rm d}M_2. \end{aligned} \tag{6.2}$$

Here $M = M_{\rm 200mean}$ is the halo mass definition[1] we will be using throughout, $n(M)$ is the halo mass function, $\bar{\rho}$ is the mean matter density of the Universe, $u(k|M)$ is the normalised Fourier transform of the density profile of a halo of mass M, and $P_{\rm hh}(k|M_1, M_2)$ is the halo-halo power contributed by two haloes of masses M_1 and M_2 on a Fourier scale k. We can rewrite the latter term assuming a linear scale-independent bias relation, $P_{\rm hh}(k|M_1, M_2) = b(M_1)b(M_2)P_{\rm lin}(k)$, where $b(M)$ is the halo bias and $P_{\rm lin}$ the linear theory matter power spectrum. We then obtain:

$$P^{\rm 2h}_{\rm dm}(k) = P_{\rm lin}(k) \left[\int n(M) b(M) \left(\frac{M}{\bar{\rho}}\right) u(k|M) {\rm d}M\right]^2. \tag{6.3}$$

From these expressions, one can easily derive a model for the galaxy power spectrum. For this we assume that the number of galaxies scales with the halo mass M; specifically, $M \propto \langle N_{\rm gal}|M\rangle$ and $M^2 \propto \langle N_{\rm gal}(N_{\rm gal}-1)|M\rangle$, leading to:

$$\begin{aligned} P^{\rm 1h}_{\rm gal}(k) &= \int n(M) \frac{\langle N_{\rm gal}(N_{\rm gal}-1)|M\rangle}{\bar{n}^2_{\rm gal}} |u_{\rm gal}(k|M)|^p {\rm d}M \\ P^{\rm 2h}_{\rm gal}(k) &= P_{\rm lin}(k) \left[\int n(M) b(M) \frac{\langle N_{\rm gal}|M\rangle}{\bar{n}_{\rm gal}} u_{\rm gal}(k|M) {\rm d}M\right]^2. \end{aligned} \tag{6.4}$$

[1] $M_{\rm 200mean}$ is the mass within a spherical region with radius $R_{\rm 200mean}$ and internal density $200 \times \bar{\rho} = 200 \times \Omega_{\rm m} \rho_{\rm crit}$.

Here the mean number density of galaxies is given by $n_{\rm gal} = \int n(M) \langle N_{\rm gal}|M\rangle \, {\rm d}M$. Note that we have followed Cooray & Sheth (2002) in replacing the normalised Fourier transform of the halo density profile, $u(k|M)$, by one describing the distribution of (satellite) galaxies, $u_{\rm gal}(k|M)$, and subsequently in changing the power-law index on this term in the 1-halo term by p. This is often done in the literature in order to be able to differentiate between contributions from central-satellite and satellite-satellite terms, with $p = 1$ for the former and $p = 2$ for the latter, based on the value of $\langle N_{\rm gal}(N_{\rm gal} - 1)\rangle$. $\langle N_{\rm gal}|M\rangle$ – the most common form of the HOD – is often separated into contributions from centrals and satellites as well, with the former ($N_{\rm cen}$) following a roughly lognormal distribution with respect to M, and the latter ($N_{\rm sat}$) being very well approximated by a (linear) power law (e.g. Guzik & Seljak, 2002; Kravtsov et al., 2004; Zehavi et al., 2005; Tinker et al., 2005; Zheng et al., 2005). From this approximate expressions for $\langle N_{\rm gal}(N_{\rm gal} - 1)\rangle$ in terms of $N_{\rm cen}$ and $N_{\rm sat}$ can be derived as well.

However, as our aim is to reproduce the results of the semi-analytical model, for which information on the HOD and the galaxy type is much more readily available than for observations, we can explicitly separate the contributions from central and satellite galaxies to the galaxy power spectrum without approximation. Keeping in mind that a halo will contain at most one central, meaning that $\langle N_{\rm cen}(N_{\rm cen} - 1)|M\rangle = 0$, that $\langle N_{\rm cen}N_{\rm sat}|M\rangle = \langle N_{\rm sat}N_{\rm cen}|M\rangle$, and using that central galaxies reside in the centre of the halo and should therefore not be weighted by the profile, we derive:

$$\begin{aligned} P^{\rm 1h}_{\rm gal}(k) &= 2\int n(M)\frac{\langle N_{\rm cen}N_{\rm sat}|M\rangle}{\bar{n}^2_{\rm gal}}\left[u_{\rm gal}(k|M) - W(kR)\right]{\rm d}M \; + \\ &\quad \int n(M)\frac{\langle N_{\rm sat}(N_{\rm sat}-1)|M\rangle}{\bar{n}^2_{\rm gal}}\left[|u_{\rm gal}(k|M)|^2 - W(kR)^2\right]{\rm d}M \\ P^{\rm 2h}_{\rm gal}(k) &= P_{\rm lin}(k)\left[\int n(M)b(M)\frac{\langle N_{\rm cen}|M\rangle}{\bar{n}_{\rm gal}}{\rm d}M + \right. \\ &\quad \left. \int n(M)b(M)\frac{\langle N_{\rm sat}|M\rangle}{\bar{n}_{\rm gal}}u_{\rm gal}(k|M){\rm d}M\right]^2 . \end{aligned} \tag{6.5}$$

Note that we have followed Valageas & Nishimichi (2011) in adding a counterterm to the halo profiles in the 1-halo term, which ensures the 1-halo term goes to zero for $k \to 0$. Here $W(kR)$ is the Fourier transform of a spherical top-hat of radius $R(M) = [3M/(4\pi\bar{\rho})]^{1/3}$, given by:

$$W(kR) = 3\left(\frac{\sin(kR)}{(kR)^3} - \frac{\cos(kR)}{(kR)^2}\right). \tag{6.6}$$

In our model, we take $P_{\rm lin}(k)$ to be the realised linear input power spectrum from the dark matter initial conditions. We calculate the halo mass function, $n(M)$, directly from the dark matter simulation as well and spline-fit the results. Furthermore, we use the fit for the $M_{\rm 200mean}$ halo bias function provided by Tinker

et al. (2010) for $b(M)$, and compute each of the four HOD terms directly from the SAM run on our halo subsample, spline-fitting these results as well.

6.2.1.2 The galaxy distribution

The normalised Fourier transform of the galaxy distribution, $u_{\rm gal}(k|M)$, is often derived from the dark matter mass profile of the halo. This in turn is usually assumed to be equal to the Navarro, Frenk & White (1997, , NFW) profile, cut off at the virial radius $r_{\rm vir} = R_{\rm 200mean}$, with some concentration-mass relation $c(M)$:

$$\rho_{\rm NFW}(r) = \frac{\rho_0}{(r/r_{\rm s})(1+r/r_{\rm s})^2}, \tag{6.7}$$

where $r_{\rm s} = r_{\rm vir}/c$ is the scale radius. The main advantage of using the one-parameter NFW profile is that this leads to an analytic expression for $u(k|M)$. However, many authors have shown that the Einasto (1965) profile provides a more accurate fit to the mean profile of haloes of a given mass, and to the distribution of dark matter substructure (e.g. Navarro et al., 2004; Merritt et al., 2005, 2006; Gao et al., 2008; Springel et al., 2008; Stadel et al., 2009; Navarro et al., 2010; Reed, Koushiappas & Gao, 2011; Dutton & Macciò, 2014). The two-parameter Einasto density profile is given by:

$$\rho_{\rm Ein}(r) = \rho_0 \exp\left\{-\frac{2}{\alpha}\left[\left(\frac{r}{r_{\rm s}}\right)^{\alpha} - 1\right]\right\}, \tag{6.8}$$

where the shape parameter α allows additional freedom in the slope of the profile. This function does not have an analytic Fourier transform, and an extra numerical integration step is therefore needed when replacing the NFW profile by an Einasto one. The larger degeneracies in fitting a two-parameter model also mean more data points are needed to obtain a reliable fit. Still, when the computational expense is acceptable and enough information on the measured profile is available, the increased accuracy may be worth the cost.

We find that the Einasto profile provides an excellent fit to the distribution of satellite galaxies in the inner parts of haloes in our simulation. But even the Einasto profile over-predicts the number of galaxies at large radii, $r \gtrsim 0.7 r_{\rm vir}$. Additionally, standard practice is to cut off the profile at the virial radius, while we find that $\sim 10\%$ of the satellite galaxies in our simulation are found at distances $1 < r/r_{\rm vir} < 3$. Note that these galaxies are not necessarily outside the virialised region, as haloes are typically not spherical objects. We therefore seek a profile with the same small-scale behaviour as the Einasto profile, while simultaneously fitting the galaxy distribution out to $\sim 3 r_{\rm vir}$.

We find that the following functional form, which we refer to here as the "gamma" profile, is capable of providing an excellent match to the galaxy distribution over the full range of scales we consider, and at any halo mass:

$$n_{\rm g}(r) = n_0 \left(\frac{r}{b}\right)^{ac-3} \exp\left\{-\left(\frac{r}{b}\right)^{c}\right\}. \tag{6.9}$$

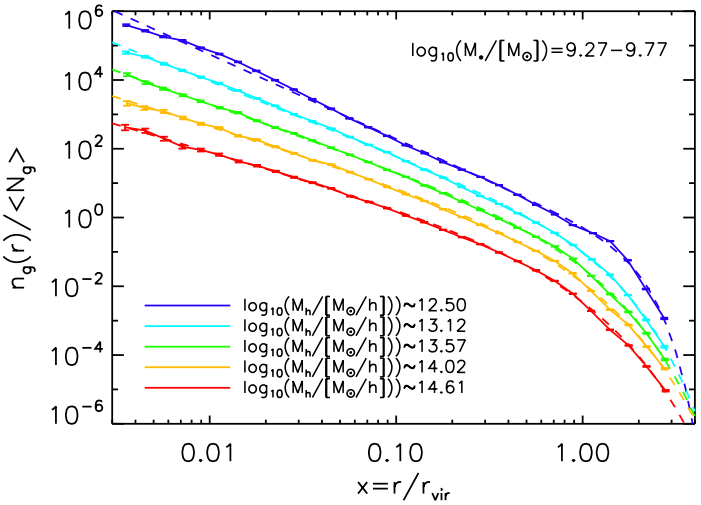

FIGURE 6.1: Galaxy number density profiles for all Guo et al. (2011) galaxies with stellar masses $10.27 < \log_{10}(M_*/\mathrm{M}_\odot) < 10.77$, for five different halo mass bins (shown in different colours). The legend shows the mean logarithmic mass in each of the bins. Solid lines indicate the measured profiles, while dashed lines show the best-fit gamma profiles (see equation 6.12). The halo mass bins are dynamically chosen such that each contains roughly the same number of galaxies, and the fits are performed using 30 radial bins spaced equally in log-space between $\log_{10} x = -2.5$ and $\log_{10} x = 0.5$.

This fitting function has three parameters, a, b and c. Note that the role of b is similar to that of r_s in the Einasto profile. Both the Einasto and gamma profiles are near universal if defined in terms of $x \equiv r/r_\mathrm{vir}$. If we rewrite both profiles in terms of x and integrate them to obtain $N(< r)$, the similarities and differences between the profiles are most easily appreciated. For the Einasto profile:

$$N_\mathrm{Ein}(< r) = N_\mathrm{tot} \frac{\gamma\left[\frac{3}{\alpha}, \frac{2}{\alpha}\left(\frac{x}{r_\mathrm{s}}\right)^\alpha\right]}{\gamma\left[\frac{3}{\alpha}, \frac{2}{\alpha}\left(\frac{x_\mathrm{max}}{r_\mathrm{s}}\right)^\alpha\right]}, \tag{6.10}$$

while for the gamma profile:

$$N_\mathrm{g}(< r) = N_\mathrm{tot} \frac{\gamma\left[a, \left(\frac{x}{b}\right)^c\right]}{\gamma\left[a, \left(\frac{x_\mathrm{max}}{b}\right)^c\right]}. \tag{6.11}$$

Here $\gamma(a, b)$ is the lower incomplete gamma function, and we have assumed the profiles cut off at some x_max. The similarities in the two profiles are clear, and the main difference is that the two parameters of the gamma function are independent for the gamma profile, which effectively allows for a steeper profile at large x and consequently a better match to the galaxy distribution around the virial radius. In practice, we fit a normalised number density profile $n_\mathrm{g}(r)/\langle N_\mathrm{g}\rangle$ to the galaxy distribution before numerically Fourier transforming this to obtain $u_\mathrm{gal}(k|M)$. For

completeness, $n_{\rm g}(r)/\langle N_{\rm g}\rangle$ is given by:

$$\frac{n_{\rm g}(r)}{\langle N_{\rm g}\rangle} = \frac{c}{4\pi b^3 r_{\rm vir}^3 \gamma\left[a, \left(\frac{x_{\rm max}}{b}\right)^c\right]} \left(\frac{x}{b}\right)^{ac-3} \exp\left\{-\left(\frac{x}{b}\right)^c\right\}. \tag{6.12}$$

In our model we set $x_{\rm max} = 3$, as $> 99.9\%$ of satellites in our fiducial model are found inside this radius. Even for small halo samples, the three parameters of the fit are independent enough to ensure degeneracies are not a problem. An example is given in Figure 6.1, where we show the best-fit model for all galaxies with stellar masses $10.27 < \log_{10}(M_*/{\rm M}_\odot) < 10.77$ in the Guo et al. (2011) semi-analytical model, for five different halo mass bins. The solid lines show the measured number density profiles, while the dashed lines show the best-fit gamma profiles. The halo mass bins are dynamically chosen inside the code such that each contains roughly the same number of galaxies. We use 30 radial bins spaced equally in log-space between $\log_{10} x = -2.5$ and $\log_{10} x = 0.5$, and fit an Akima spline through each of the three parameters as a function of halo mass to obtain smooth functions that are stable to outliers.

6.2.1.3 Correction for non-sphericity

As is common, we have assumed a spherical distribution of satellite galaxies around each central. In reality, haloes and consequently their galaxy populations are triaxial. van Daalen, Angulo & White (2012) investigated the effect of assuming a spherical distribution on the two-point correlation function and galaxy power spectrum, and found that the effects can be quite large, with the true power being underestimated by 1% around $k = 0.2\,h\,{\rm Mpc}^{-1}$ to 10% around $k = 25\,h\,{\rm Mpc}^{-1}$, increasing even more towards smaller scales (see the right panel of their Figure 3, or Figure 4.3 in Chapter 4 of this thesis). We have repeated their analysis and found that the functional shape of this underestimation of the power appears to be completely independent of the mass of the galaxies. We therefore fit a function $e(k)$ through these results and use this to correct our halo model power spectra for the combined effects of non-sphericity. The final galaxy power spectrum that comes out of our model for a given set of galaxies is therefore:

$$P_{\rm gal} = [P_{\rm gal}^{\rm 1h}(k) + P_{\rm gal}^{\rm 2h}(k)]/[1 + e(k)], \tag{6.13}$$

with $P_{\rm gal}^{\rm 1h}(k)$ and $P_{\rm gal}^{\rm 2h}(k)$ given by equation (6.5).

6.2.1.4 Converting to the projected correlation function

To obtain the projected correlation function from the galaxy power spectrum, we numerically perform two standard transformations. First, to obtain the 3D correlation function:

$$\xi(r) = \frac{1}{2\pi^2}\int_0^\infty k^2 P(k) \frac{\sin kr}{kr}\,{\rm d}k, \tag{6.14}$$

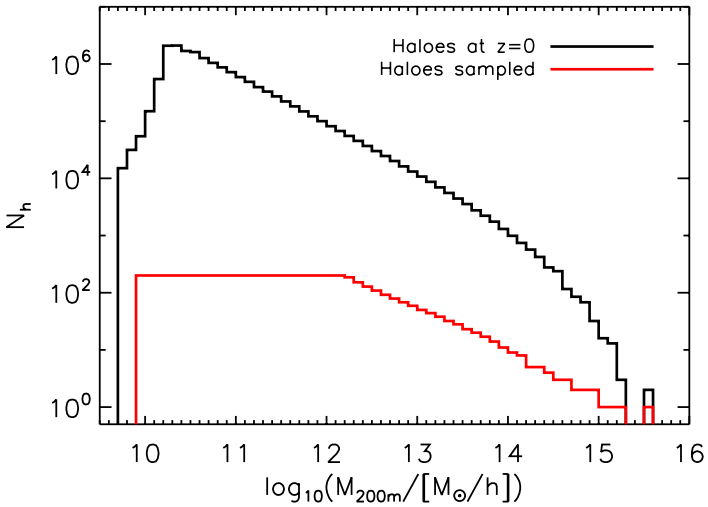

FIGURE 6.2: The FoF halo mass function, showing number of haloes available in the Millennium Simulation at $z = 0$ (black) and the number randomly selected as a function of $M_{200\mathrm{mean}}$ in each subsample (red). The subsamples each comprise of less than 0.04% of the total halo sample, or 0.14% of the total subhalo sample. The selection function was built iteratively by demanding that $\sim 90\%$ of the random samples it generated lead to projected correlation functions that were within 30% of the full sample prediction. Low-mass haloes were favoured over high-mass haloes in order to suppress the size of the trees used in the SAM. Even so, the fraction of FoF groups needed to match the correlation function within some uncertainty at any stellar mass is higher for more massive haloes.

and, finally, to obtain the projected galaxy correlation function:

$$w(r_\mathrm{p}) = 2\int_0^\infty \xi\left(\sqrt{r_\mathrm{p}^2+\pi^2}\right)\mathrm{d}\pi = 2\int_{r_\mathrm{p}}^\infty \frac{r\xi(r)}{\sqrt{r^2-r_\mathrm{p}^2}}\mathrm{d}r. \tag{6.15}$$

Here r_p and π are the projected and line-of-sight separation, respectively. It is in this last step that we also convert the units from Mpc/h to Mpc, in order to directly compare our model $w(r_\mathrm{p})$ to that of observations.

6.2.1.5 Selection function

The selection function we use to create the halo sample the SAM is repeatedly run on while exploring parameter space was built through use of the following algorithm.

At each step, the algorithm adds some number of Friends-of-Friends (FoF) groups to each halo mass bin in turn, and generates a number of random samples for each of the resulting selection functions. The correlation functions predicted using these samples are then compared to determine which mass bin would contribute to the largest reduction in the variance with respect to the full model run

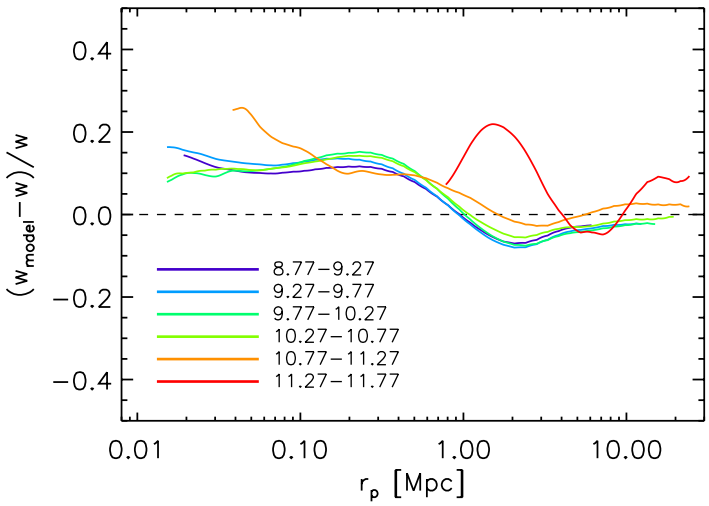

FIGURE 6.3: The fractional difference between our model prediction of the projected galaxy correlation function and a direct calculation, for galaxies in the Guo et al. (2011) semi-analytical model. Here we use the full galaxy sample as an input to our model. Results are shown for six different stellar mass bins, indicated by lines of different colours, over the range where SDSS/DR7 data is available for each. The overall agreement is within 20%, with the model tending to over-predict the clustering on sub-Mpc scales. This can be traced to an overestimation of the power in the 1-halo term by a similar amount around $k = 1\,h\,\mathrm{Mpc}^{-1}$. For our application, our model performs well enough, and we leave improvements to future work.

for all six stellar mass bins. If at any step adding more haloes does not reduce the variance for any halo mass, FoF groups are added to a random bin. This continues until at least 90% of the random samples the current selection function generates lead to projected correlation functions that are within 30% of the full sample prediction.

In order to suppress the size of the merger trees used in the SAM, low-mass haloes were favoured over high-mass haloes by weighting the number of FoF groups added to each mass bin by the inverse of the average number of subhaloes hosted by FoF groups of that mass. Nonetheless, the fraction of haloes selected at high mass is still higher than at low mass, since more massive haloes potentially contribute more galaxies to the sample, increasing the accuracy of the estimates made in the clustering model described above. Additionally, the most massive galaxies probed here, $M_* > 10^{11.27}\,\mathrm{M}_\odot$, preferentially occupy the most massive haloes.

After building several selection functions in this way, we found that on average they were well approximated by the combination of a constant value and a power law (rounded to integer values). This is the near-optimal selection function shown in Figure 6.2 (red line), which takes the constant value $N_\mathrm{h} = 200$ below $M_{200\mathrm{mean}} = 10^{12.2}\,h^{-1}\,\mathrm{M}_\odot$. The subsamples generated by this selection function each comprise less than 0.04% of the total FoF halo sample, or 0.14% of the total subhalo sample.

FIGURE 6.4: The fractional difference between the predictions of our model for 100 halo subsamples and the model prediction for the full sample. Each subsample consists of about 0.14% of the total subhalo sample (see text for details). The colours indicate the same stellar mass bins as in Figure 6.3. *Left:* The predictions of each of the 100 separate realisations, showing the scatter around the full sample result. *Right:* Same as the left panel, but now showing only the median, 16^{th} and 84^{th} percentiles. Even using a small random sample, our model can quickly estimate the projected correlation function to $\sim 10\%$ precision.

6.2.1.6 Performance of the model

We compare our model prediction of $w(r_\mathrm{p})$, using the full halo sample, to that calculated directly for the galaxies in the Guo et al. (2011) model in Figure 6.3. Here we show the relative difference between the two for six different bins in stellar mass, indicated as ranges in $\log_{10}(M_*/\mathrm{M}_\odot)$. We only show the results over the range where we constrain $w(r_\mathrm{p})$ using observations. The model performs well, and any deviations from the true correlation function are typically within 20%. The magnitude of the mismatch tends to increase with stellar mass. The large-scale disagreement is caused by the model slightly under-predicting the power in the transition region between the 1-halo and 2-halo terms, while the small-scale offset is mostly due to the 1-halo term in the power spectrum being slightly overestimated around $k = 1\,h\,\mathrm{Mpc}^{-1}$. However, overall the agreement is good, especially considering our relatively simple treatment of e.g. the halo bias (linear and scale-independent), and we leave further improvements – such as using a halo-halo power spectrum measured from the dark matter only simulation instead of a biased linear power spectrum – to future work.

The true power of the model lies in its ability to reproduce the clustering prediction for the full sample from only a small subsample of FoF groups. In Figure 6.4 we compare the predictions for 100 random subsamples selected according to the selection function shown in Figure 6.2 to the model prediction for the full sample. The dotted lines indicate offsets of 30% for reference, and the colours indicate the same stellar mass bins as in Figure 6.3. The scatter is around $7 - 8\%$ for the first four mass bins, increasing to 10% and 16% for the fifth and sixth mass bin respectively. This shows that the model is capable of reproducing the full sample estimate from relatively few haloes.

6.2.2 The SAM and MCMC

As our estimator is able to quickly and accurately recover the projected correlation function from a very small subsample of haloes, this makes it ideally suited for constraining the parameter space of semi-analytical models using the projected correlation function. In this work we present a first application, where we constrain the model of G13, a recent version of the Munich semi-analytical code, using both the galaxy stellar mass function (SMF) and the projected galaxy correlation function. For this we utilise the same data sets as presented in G13. As we will only utilise the Millennium Simulation, and not Millennium II, we only use constraints above $M_* > 10^9\,h^{-1}\,\mathrm{M}_\odot$.

The G13 model includes 17 parameters which together determine the outcome of galaxy formation. These are (see Table 6.1): the star formation efficiency (α_SF); the star formation criterion ($\tilde{M}_\mathrm{crit}$, or equivalently Σ_crit); the star formation efficiency in the burst phase following a merger ($\alpha_\mathrm{SF,burst}$); the slope on the merger mass ratio determining the stellar mass formed in the burst ($\beta_\mathrm{SF,burst}$); the AGN radio mode efficiency (k_AGN); the black hole growth efficiency (f_BH); the typical halo virial velocity of the black hole growth process (V_BH); three parameters

Parameter	Description	Units
$\alpha_{\rm SF}$	Star formation efficiency	–
$\tilde{M}_{\rm crit}$	Star formation threshold	$\mathrm{M}_\odot\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$
$\alpha_{\rm SF,burst}$	Star formation burst mode efficiency	–
$\beta_{\rm SF,burst}$	Star formation burst mode slope	–
$k_{\rm AGN}$	Radio feedback efficiency	$h^{-1}\,\mathrm{M}_\odot\,\mathrm{yr}^{-1}$
$f_{\rm BH}$	Black hole growth efficiency	–
$V_{\rm BH}$	Quasar growth scale	$\mathrm{km\,s^{-1}}$
ϵ	SN mass-loading efficiency	–
$V_{\rm reheat}$	Mass-loading scale	$\mathrm{km\,s^{-1}}$
β_1	Mass-loading slope	–
η	SN ejection efficiency	–
$V_{\rm eject}$	SN ejection scale	$\mathrm{km\,s^{-1}}$
β_2	SN ejection slope	–
γ	Ejecta reincorporation scale factor	–
y	Metal yield fraction	–
$R_{\rm merger}$	Major-merger threshold ratio	–
$\alpha_{\rm friction}$	Dynamical friction scale factor	–

TABLE 6.1: Parameters varied in the MCMC. The best-fit values (as well as the G13 values for the WMAP1 cosmology and the prior ranges) are shown in Figure 6.8. For more information we refer to G13.

governing the reheating and injection of cold disk gas into the hot halo phase by supernovae, namely the gas reheating efficiency (ϵ), the reheating cut-off velocity ($V_{\rm reheat}$) and the slope of the reheating dependence on $V_{\rm vir}$ (β_1); three parameters governing the ejection of hot halo gas to an external reservoir, namely the gas ejection efficiency (η), the ejection cut-off velocity ($V_{\rm eject}$) and the slope of the ejection dependence on $V_{\rm vir}$ (β_2); a parameter controlling the gas return time from the external reservoir to the hot halo (γ); the yield fraction of metals returned to the gas phase by stars (y); the mass ratio separating major and minor merger events ($R_{\rm merger}$); and finally a parameter controlling the dynamical friction time scale of orphan galaxies, i.e. the time it takes for satellite galaxies of which the dark matter subhalo is disrupted (or at least no longer detected) to merge with the central galaxy ($\alpha_{\rm friction}$).

While in the original G13 paper some of these parameters were held fixed, here we allow all 17 to vary. We start our Monte Carlo Markov Chains (MCMCs) at the position in parameter space used by Guo et al. (2011), which was arrived at by using a combination of SMFs, as well as rest-frame B-band and K-band luminosity functions between $z = 0$ and $z = 3$, as observational constraints. We then use the same techniques as described in G13 to find a new set of best-fit parameters, with the projected correlation function as an additional constraint.

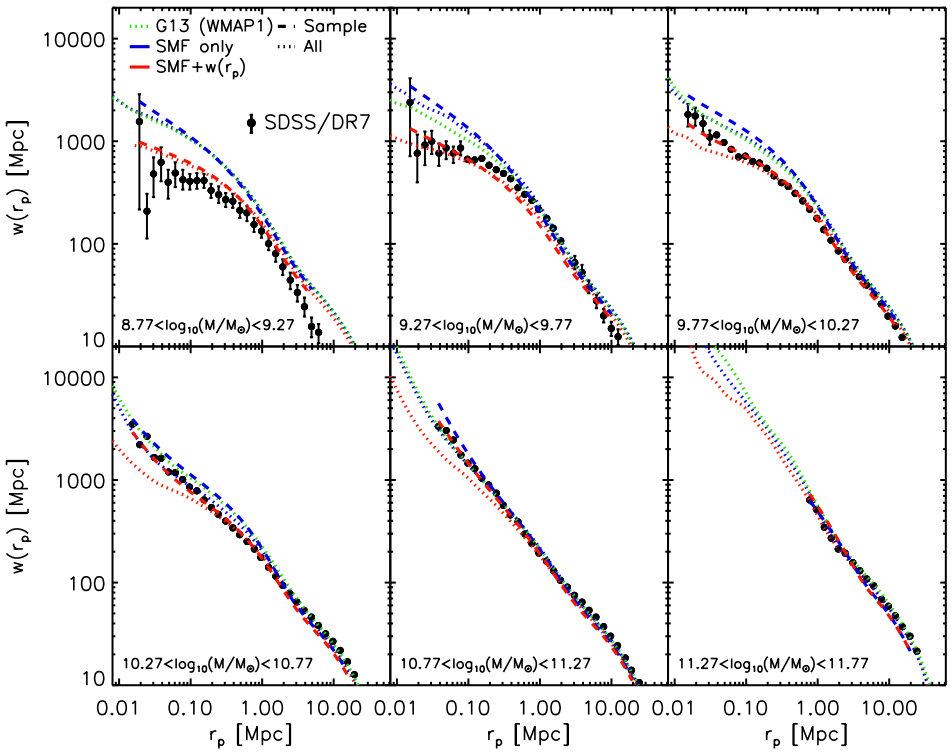

FIGURE 6.5: The projected galaxy correlation function in six bins of stellar mass. The points with error bars show the SDSS data in each bin, while the lines show the model results. The green dotted line shows the results for the original model from G13, in which the parameter values were set manually. The blue lines show the results of only using the stellar mass function as a constraint, while the red lines show the results when the model is simultaneously constrained by the projected correlation function and the stellar mass function. Finally, dashed and dotted lines are used to indicate whether these are the results for the sample haloes or for all haloes, respectively. The clustering on small scales of the full model is systematically underestimated by the sample, which is mostly due to the clustering estimator (see §6.2.1.6). Note that even though the lowest mass bin is not used as a constraint, the match to observations is markedly improved with respect to the other models.

As our model is only accurate to within $\sim 10\%$ on small scales, and additionally since the error bars on the SDSS clustering data were derived from Poisson statistics alone, and so do not include cosmic variance, we artificially increase the error bars on the data points used during the fitting. Each data point of the observed projected correlation function was assumed to have an uncertainty of 20%. As noted before, we do not use the clustering data below $M_* = 10^{9.27}\,\mathrm{M}_\odot$, nor the stellar mass function data below $M_* < 10^9\,\mathrm{M}_\odot$, when constraining the model, as the haloes hosting these galaxies are not well resolved in the original Millennium Simulation which we are using as a basis for the SAM. When fitting to the SMF and clustering data simultaneously, we increase the relative weighting of the fit to

the SMF by a factor of five to compensate for the fact that the clustering data is measured in five separate bins. This helps avoid sacrificing the excellent fit to the SMF in favour of matching the correlation function.

Note that while G13 used a WMAP7 cosmology, here we use the original WMAP1 cosmology to avoid additional complications introduced by scaling to a different cosmology. In future work the results will be explored for more up-to-date cosmologies. Contrary to what is claimed in G13, the change in cosmology has a negligible impact on the resulting correlation functions, which are far more sensitive to the SAM's physical recipes. Besides updating the cosmology, the only change made from the WMAP1 Guo et al. (2011) model to the newer WMAP7 G13 model is that the type 2 (orphan) satellite galaxy positions are now correctly updated in the code, meaning that their orbits now decay as intended and can therefore be disrupted earlier. This change was the main reason for the improved agreement with clustering data with respect to Guo et al. (2011).

6.3 Results

6.3.1 Comparison with observations

The results of our MCMC chains for the projected correlation function are shown in Figure 6.5, for six bins in stellar mass, as indicated in the panels. In each figure, we indicate the original results found by G13, where the galaxy formation parameters were set by hand, as a green dotted line. The new results are shown in blue and red; in blue, we show the correlation functions that follow from only using the stellar mass function as a constraint ("SMF-only"), while in red we show the results of fitting to the clustering data simultaneously ("SMF+$w(r_{\rm p})$").

The dashed lines show the predictions made based on the sample of haloes used in the MCMC, as described in §6.2.1. The dotted lines show the true galaxy correlation function, as calculated directly from the full galaxy catalogue for the same model parameters. The true values are generally below the ones estimated from the sample, as expected from the results of §6.2.1.6, and as a consequence the new results tend to under-predict the amount of clustering on the smallest scales.

Even so, one immediately sees that the SMF+$w(r_{\rm p})$ correlation functions (red lines) generally provide a better fit to the data, bringing the small-scale clustering down considerably in comparison with the original G13 and SMF-only (blue lines) models, which are very close together. This effect is larger for low stellar masses, where the clustering discrepancy between the old model and the data was larger as well. The much improved match to observations indicate that the model retains enough freedom to match the clustering data. Note that the match to the projected galaxy correlation function for galaxies in the first mass bin is greatly improved as well, even though this data is not used to constrain the model. For the highest-mass galaxies, $11.27 < \log_{10}(M_*/\mathrm{M}_\odot) < 11.77$, all models perform equally well, while for galaxies with masses above $10^{10.27}\,\mathrm{M}_\odot$ the SMF-only correlation

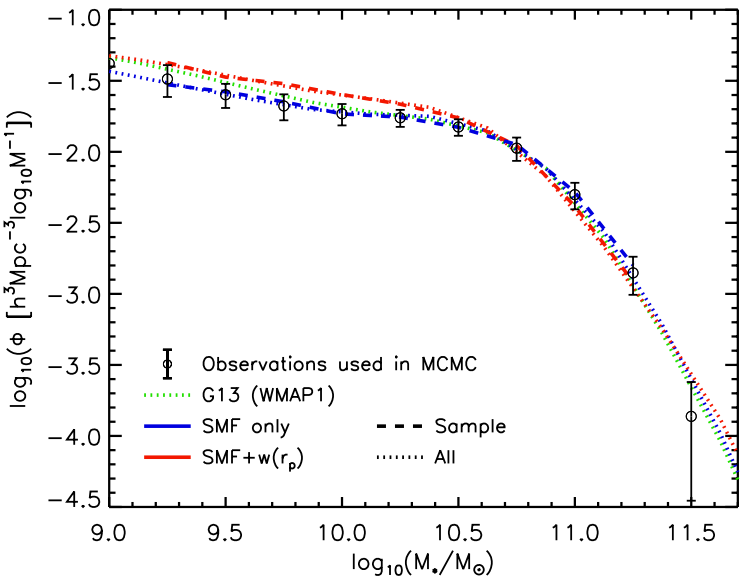

FIGURE 6.6: The stellar mass functions of the models. The green line again refers to the original G13 model, of which the parameters were set manually. The blue lines show the results when the MCMC algorithm is used with only the stellar mass function as a constraint, while the red lines again show the result when clustering constraints are used additionally. When the SMF is the only constraint, the model clearly has enough freedom to reproduce it to high precision. However, the match grows somewhat worse at low mass when the model is additionally constrained by clustering, and is in some places about 2σ away from the combined observational constraints shown in black. Still, the SMF+$w(r_\mathrm{p})$ model performs better in matching both sets of constraints simultaneously.

functions perform better on the smallest scales, due to the clustering estimator overestimating the small-scale clustering.

However, the improved match to the observed clustering data (at least for low-mass galaxies) comes at a price. In Figure 6.6, we show how the models compare to the SMF data used to constrain the models. The black points with error bars are derived by combining several observational data sets (see G13). The original G13 model, in which the parameters were set by hand, is again shown as a green dotted line, which matches the data well. When we use only the SMF as a constraint for the galaxy formation model, shown in blue, we obtain a marginally better fit to the data at low mass.

When the projected galaxy correlation function is used as an additional constraint, shown in red, the agreement with the stellar mass function suffers considerably in favour of the clustering predictions. While the agreement for galaxies with masses $M_* \gtrsim 10^{10.5}\,\mathrm{M}_\odot$ is still comparable to that obtained by G13, the new model over-predicts the number densities of lower-mass galaxies, although the results are still within 2σ of the data. Note that the sample results (dashed lines) agree perfectly with the full catalogue ones (dotted lines) for both SMF-only

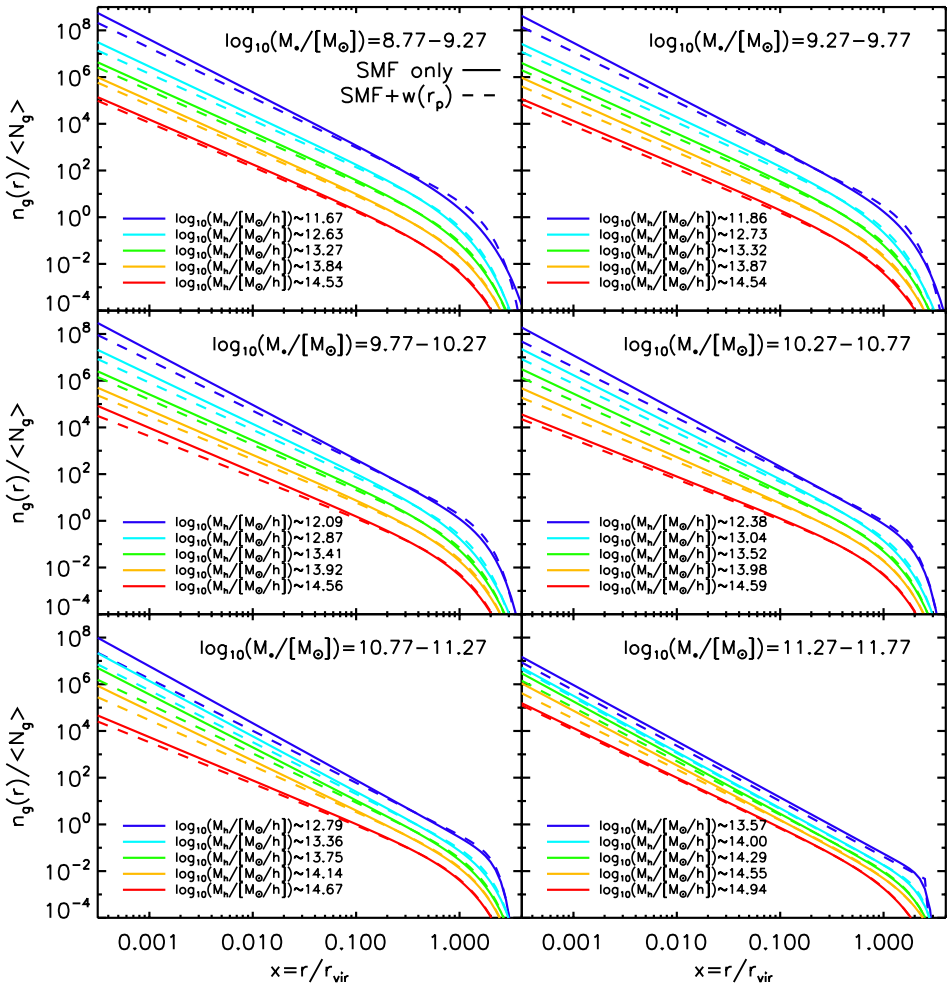

FIGURE 6.7: Comparison of the galaxy distribution profiles for the SMF-only (solid lines) and SMF+$w(r_{\rm p})$ (dashed lines) best-fit parameters. The different panels show the profiles of galaxies in the six correlation function mass bins, as indicated in the top right of each panel. As in Figure 6.1, different colours are used for different halo mass bins, which are set to be the same for both models to allow for an unbiased comparison. Note that the mass bins do change as a function of stellar mass in order to make sure each bin in halo mass is roughly equally populated. For clarity, we show only the fits to the measured profiles (see equation 6.12) here, but stress that each provides an excellent fit over the full range shown here. Note that the dynamic range in scales has been extended relative to Figure 6.1 to better appreciate the differences between the profiles. Mainly because of the reduced dynamical friction time scale in the latter model, the profiles of galaxies in every mass bin are slightly flatter at any halo mass, reducing the correlation function on small scales.

and SMF+$w(r_{\rm p})$, which indicates that the discrepancy observed for the correlation function is indeed due to the inaccuracy of the estimator at small separations.

The slight mismatch for low-mass galaxies could indicate that the SAM is missing some physical ingredient needed in order to reproduce observations, but other viable explanations also exist. For both the clustering and SMF data the uncertainties may be underestimated; for example, the error bars on the correlation function do not take into account cosmic variance, which could have a quite significant effect. If the observed correlation functions are biased low because of this, the clustering in our model may have been brought artificially low, preventing us from matching the SMF simultaneously. Another possible source of errors could be systematic uncertainties in the observations that lead to samples that are not volume limited. Additionally, changing the cosmology to one that is more up-to-date may help. We will explore some of these possibilities in future work. Note, however, that the SMF+$w(r_{\rm p})$ model is in far closer agreement with both the SMF and the clustering data *simultaneously* than both the original G13 and the SMF-only models: while the latter models are in strong disagreement with the clustering data for low-mass galaxies on small scales, the SMF+$w(r_{\rm p})$ model is generally in agreement with both the low-mass clustering data and the SMF within 2σ. This shows the merit of using a clustering estimator while exploring parameter space.

6.3.2 Change in parameters

Even though we vary 17 galaxy formation parameters, by far the largest role in bringing the clustering predictions in agreement with observations is played by only two of these: $\alpha_{\rm friction}$, which controls the time it takes for satellite galaxies to merge with the central once their dark matter subhalo has been disrupted, and γ, which controls the time it takes for ejected gas to re-enter the halo.

The way these parameters influence the clustering and stellar mass function predictions is as follows. When the clustering data is included as an additional constraint, the dynamical friction time scale of orphan galaxies decreases by more than a factor of three with respect to the SMF-only results. This causes galaxies at small separation scales to merge with their centrals much quicker, flattening the galaxy distribution profile within the haloes and greatly decreasing the amount of clustering on small scales, especially for low-mass satellites. This change in the galaxy distribution profiles from the SMF-only to the SMF+$w(r_{\rm p})$ model is shown in Figure 6.7. The halo mass bins are set to be the same for both models to allow for an unbiased comparison. Note that the mass bins do change as a function of stellar mass in order to make sure each bin in halo mass is roughly equally populated. Although we only show the fits to the measured profiles here, we stress that each provides an excellent fit to the data, over the full range in scales shown here. The change in slope of the profiles is relatively small, meaning that the galaxy distributions are still consistent with SDSS data for rich clusters (see Figure 14 of Guo et al., 2011). This is because even though the friction time scale

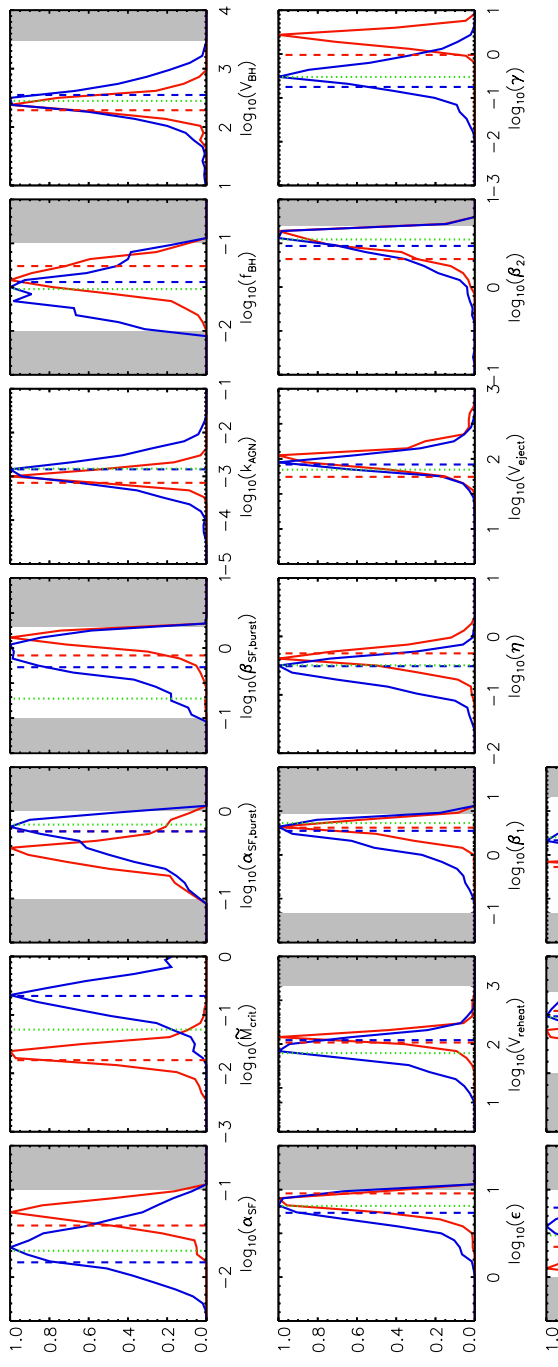

FIGURE 6.8: The preferred parameter values in both models. The best-fit values are shown as dashed vertical lines. The dotted vertical line again shows the result for the original G13 model, and the grey regions indicate values deemed non-physical, and which are therefore made inaccessible to the model. For most parameters the best-fit values are consistent between the model where clustering is not used as a constraint and the model where it is. The largest shifts occur for $\alpha_{\rm SF}$, $\tilde{M}_{\rm crit}$, y, γ and $\alpha_{\rm friction}$. The last two are most important for bringing the clustering data into agreement with data, while the rest mainly serve to preserve the match to the SMF.

decreases by more than a factor of three when using clustering as an additional constraint, the number of type 2 galaxies at $z = 0$ decreases only by a factor 0.87, as the merging time scale for many of these galaxies is still long compared to the Hubble time.

Additionally, however, the decrease in the dynamical friction time scale causes the number density of galaxies above the knee ($M_* > 10^{10.5}\,\mathrm{M}_\odot$) to decrease as well. This counter-intuitive change in the SMF comes about because the cold gas in the merging satellites directly feeds the supermassive black holes in the centres of the central galaxies, increasing feedback from AGN and thereby the suppression of star formation.

The γ parameter, on the other hand, increases by more than a factor of five in SMF+$w(r_\mathrm{p})$ with respect to SMF-only, meaning that the hot gas reincorporation time scale *decreases* by the same factor. This raises the number densities of galaxies at any mass, but most significantly below the knee of the SMF ($M_* < 10^{10.5}\,\mathrm{M}_\odot$). The change in γ is the main source of the higher low-mass number densities from SMF-only to SMF+$w(r_\mathrm{p})$. The upside is that this parameter shift also lowers the clustering of galaxies, especially for galaxies with masses $M_* > 10^{9.77}\,\mathrm{M}_\odot$. While it may seem counter-intuitive to have the number of galaxies at some mass increase while their clustering decreases, keep in mind that it is the (normalised) galaxy distribution within each halo that is driving the clustering prediction, and this distribution flattens when the aforementioned time scales decrease.

The parameter changes in γ and α_friction alone, with respect to the best-fit parameters of the SMF-only data, already produce predictions that are very close to those of the SMF+$w(r_\mathrm{p})$ model. While the decrease in the dynamical friction and reincorporation time scales each bring the clustering into better agreement with data separately, a change in both simultaneously is needed as they affect the SMF in different (adverse) ways.

We show the shift in parameter values in Figure 6.8. We again indicate the results for all three models: the original G13 model (green dotted lines), the SMF-only model (blue lines), and the SMF+$w(r_\mathrm{p})$ model (red lines). Histograms indicate the Bayesian likelihood regions as derived from the full MCMC chains, while the vertical dashed lines indicate the best-fit values. Both the likelihood regions and the best-fit values of the SMF-only and SMF+$w(r_\mathrm{p})$ models are generally consistent. The largest exceptions to these are the star formation efficiency α_SF, the cold gas mass star formation threshold $\tilde{M}_\mathrm{crit}$, the metal yield y, and the previously mentioned reincorporation scale factor γ and dynamical friction scale factor α_friction. The latter two cause the main decrease in the clustering predictions, needed to bring them in agreement with observations. The significant increase in the star formation efficiency and the decrease in the cold gas mass threshold for star formation, on the other hand, mainly affect the SMF, compensating for the decrease in high-mass galaxies due to the more active AGN, caused in turn by the change in α_friction. Finally, the large change in the metal yield is of little consequence, as this parameter is largely unconstrained by both the SMF and correlation functions.

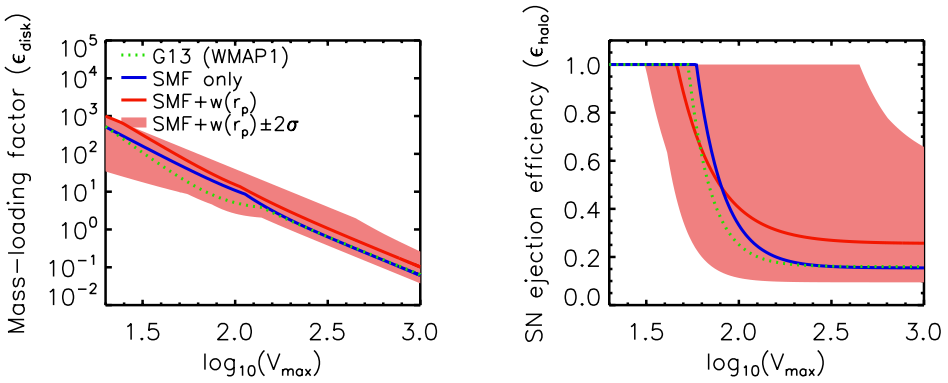

FIGURE 6.9: The effect of the changes in the supernova parameters ϵ, $V_{\rm reheat}$, β_1, η, $V_{\rm eject}$ and β_2. The mass-loading factor (left panel) goes up slightly when using the correlation function as an additional constraint, but the change is not significant with respect to the 2σ regions allowed (also shown). The same goes for the supernova ejection efficiency (right panel).

To show the effect of the changes in the feedback parameters (ϵ, $V_{\rm reheat}$, β_1, η, $V_{\rm eject}$ and β_2), we turn to Figure 6.9. In the left-hand panel, we show the SN mass loading as a function of the maximum virial velocity of the halo, for all three models. We also indicate the 2σ regions allowed by the parameters for the SMF+$w(r_{\rm p})$ model. It is clear that while the supernova mass loading increases when the clustering data is used as an additional constraint, the change is not significant.

The right-hand panel of Figure 6.9 shows how the SN ejection efficiency changes between the different models. Because the parameter η is significantly higher in the SMF+$w(r_{\rm p})$ model with regards to the others, the high-$V_{\rm max}$ horizontal asymptote of this function is increased, meaning SNe are more effective at ejecting material for galaxies occupying massive haloes. However, the large 2σ regions again indicate that the constraints used here are not very sensitive to these changes.

6.4 Summary

We have developed a fast and accurate clustering estimator, capable of predicting the projected galaxy correlation function for a full galaxy catalogue to within $\sim 10\%$ accuracy using only a very small subsample of haloes ($< 0.1\%$ of the total sample). In this work, we have described our estimator and demonstrated its effectiveness for use in constraining parameter space for semi-analytical models of galaxy formation, using the Guo et al. (2013) version of the Munich SAM as a test case.

Our estimator determines the halo occupation distribution of galaxies in the subsample and fits a profile to the galaxy distribution within haloes as a function of halo mass, using these quantities in a halo model based approach to determine the

galaxy clustering of the full sample. By being able to quickly predict the two-point galaxy correlation function for the first time while exploring parameter space, one can use clustering observations to limit the range allowed to the galaxy formation parameters of any SAM, adding constraints complementary to those of one-point functions typically used today, such as the stellar mass or luminosity function. As we have demonstrated, this may lead to different sets of parameters through which the resulting model is able to provide a better match to the observed stellar mass and correlation functions simultaneously.

For the G13 model tested here, the improved match to the correlation function is achieved mainly by significantly decreasing the time it takes for stripped (orphan) satellites galaxies to merge with their centrals, as well as the time it takes for gas ejected into the hot halo by feedback processes to be reincorporated. Both changes cause the galaxy distribution profiles within haloes to flatten, lowering the clustering on small scales. Other parameter shifts mainly serve to keep the changes in the SMF caused by the reduced time scales in check.

While the use of the clustering estimator presented here clearly has merit, some issues remain to be solved. The estimator tends to over-predict clustering on small scales, leading to final results that tend to fall $\sim 10\%$ below the observational constraints. Improving the model, for example by adding higher-order terms to the linear halo bias currently used, or basing the clustering predictions of galaxies directly on the measured clustering of the haloes in N-body simulations may help. Additionally, the agreement with the SMF could be improved at low mass. We will explore these topics in future work.

Acknowledgements

The authors thank Joop Schaye for useful discussions and comments on the manuscript. The Millennium Simulation databases used in this chapter and the web application providing online access to them were constructed as part of the activities of the German Astrophysical Virtual Observatory. This work was supported by the Marie Curie Initial Training Network CosmoComp (PITN-GA-2009-238356) and by Advanced Grant 246797 "GALFORMOD" from the European Research Council.