

# Multiwavelength consensus of large-scale linear bias

Hengxing Pan,<sup>1,2,3,4★</sup> Danail Obreschkow<sup>1,5</sup>, Cullan Howlett<sup>1,5</sup>,  
 Claudia del P. Lagos<sup>1,5,6</sup>, Pascal J. Elahi<sup>1,5</sup>, Carlton Baugh<sup>7</sup>  
 and Violeta Gonzalez-Perez<sup>1,8</sup>

<sup>1</sup>International Centre for Radio Astronomy Research, University of Western Australia, 35 Stirling Hwy, Crawley, WA 6009, Australia

<sup>2</sup>Purple Mountain Observatory, 2 West Beijing Road, Nanjing 210008, China

<sup>3</sup>University of the Chinese Academy of Sciences, 19A, Yuquan Road, Beijing 100049, China

<sup>4</sup>Astrophysics, University of Oxford, Denys wilkinson Building, Keble Road, Oxford OX1 3RH, UK

<sup>5</sup>ARC Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D)

<sup>6</sup>Cosmic Dawn Center (DAWN), Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark

<sup>7</sup>Institute for Computational Cosmology, Department of Physics, Durham University, Durham DH1 3LE, UK

<sup>8</sup>Astrophysics Research Institute, Liverpool John Moores University, 146 Brownlow Hill, Liverpool L3 5RF, UK

Accepted 2020 January 5. Received 2019 December 12; in original form 2019 July 6

## ABSTRACT

We model the large-scale linear galaxy bias  $b_g(x, z)$  as a function of redshift  $z$  and observed absolute magnitude threshold  $x$  for broad-band continuum emission from the far-infrared to ultraviolet, as well as for prominent emission lines, such as the  $H\alpha$ ,  $H\beta$ , Ly  $\alpha$ , and  $[O\ II]$  lines. The modelling relies on the semi-analytic galaxy formation model GALFORM, run on the state-of-the-art  $N$ -body simulation SURFS with the Planck 2015 cosmology. We find that both the differential bias at observed absolute magnitude  $x$  and the cumulative bias for magnitudes brighter than  $x$  can be fitted with a five-parameter model:  $b_g(x, z) = a + b(1 + z)^e(1 + \exp[(x - c)d])$ . We also find that the bias for the continuum bands follows a very similar form regardless of wavelength due to the mixing of star-forming and quiescent galaxies in a magnitude-limited survey. Differences in bias only become apparent when an additional colour separation is included, which suggest extensions to this work could look at different colours at fixed magnitude limits. We test our fitting formula against observations, finding reasonable agreement with some measurements within  $1\sigma$  statistical uncertainties, and highlighting areas of improvement. We provide the fitting parameters for various continuum bands, emission lines, and intrinsic galaxy properties, enabling a quick estimation of the linear bias in any typical survey of large-scale structure.

**Key words:** galaxies: large-scale bias – galaxies: surveys – galaxies: formation.

## 1 INTRODUCTION

Most surveys of the cosmic large-scale structure (LSS) rely on galaxies as tracers of the dark matter distribution. In order to extract the cosmological information from such surveys, it is crucial to understand the difference between the spatial statistics of the detectable galaxies and the underlying density field. At the largest scales, this difference takes the form of a scaling factor, known as linear galaxy bias  $b_g$  (Kaiser 1984), between the matter and galaxy power spectrum. Prior knowledge of this bias is essential in designing cosmological surveys.

Moreover, the galaxy bias has an important role in the redshift space distortions (RSDs) and multitracer analyses as a nuisance

variable in the former case (e.g. Clerkin et al. 2015; Li, Schmittfull & Seljak 2018) and an integral part of target selection in the latter (e.g. McDonald & Seljak 2009; Abramo, Secco & Loureiro 2016). Predicting the galaxy bias from simulations can be used to place priors on galaxy bias in RSD measurements, and compared to lensing and LSS measurements to improve our understanding of galaxy physics. Knowing the galaxy bias can also help to break the classic degeneracy between the linear bias and cosmological parameters such as  $\sigma_8$  (the linear rms of the dark matter density perturbations on scales of  $8 h^{-1}\text{Mpc}$ ) and the linear growth rate  $f$  present in two-point statistics (e.g. Ali et al. 2018 and references therein); hence priors on the bias might help better determine these parameters.

It is possible to measure the galaxy bias by computing the ratio of the two-point correlation function and three-point correlation function (see Verde et al. 2002; Gaztañaga et al. 2005, and references

\* E-mail: [hengxing.pan@physics.ox.ac.uk](mailto:hengxing.pan@physics.ox.ac.uk)

therein) or by exploiting phase-space correlations (Ali et al. 2018). The bias can further be constrained by combining galaxy redshift surveys with gravitational lensing data (e.g. Simon et al. 2007; Jullo et al. 2012). Existing bias measurements correspond to specific survey selection criteria, such as selecting the galaxies seen in the near-infrared (IR) band (Orsi et al. 2010), the  $H\alpha$  emission line (Amendola et al. 2017), or 21 cm emission from neutral hydrogen (HI) (Castorina & Villaescusa-Navarro 2017). However, these bias measurements are often uncertain and their extrapolation to different surveys, i.e. different wavelengths and sensitivities, is challenged by the complex radiative physics of galaxies. For instance, ongoing and upcoming LSS surveys (BOSS,<sup>1</sup> eBOSS,<sup>2</sup> DESI,<sup>3</sup> LSST,<sup>4</sup> 4MOST,<sup>5</sup> EUCLID<sup>6</sup> for which the references are shown in Section 3) are covering a wide range of wavelengths and many of them also use specific emission lines to probe the cosmic LSS. Predictive bias models would help calibrating the analysis of the expected data from these surveys. In addition, it might also be interesting to explore the galaxy bias dependence on some physical properties, such as star formation rate or stellar age, in order to better understand the physical origin of specific bias values. Providing a consensus of the linear bias for various broad-band wavelengths, emission lines and physical properties, is therefore a pressing goal.

On the theoretical side, there are two classes of approaches to model the clustering and bias of galaxies (Baugh 2013). The first class consists of populating simulated or analytically evolved dark matter haloes with galaxies drawn from observed luminosity functions, for instance using the so-called halo occupation distribution (HOD) (e.g. Berlind & Weinberg 2002; Zheng et al. 2005; Wechsler & Tinker 2018) method or subhalo abundance matching (SHAM, e.g. Vale & Ostriker 2004; Shankar et al. 2006). The second class uses hydrodynamics simulations (e.g. Vogelsberger et al. 2014; Schaye et al. 2015) and semi-analytic models (SAM, e.g. Henriques et al. 2015; Somerville, Popping & Trager 2015; Croton et al. 2016; Lacey et al. 2016; Xie et al. 2017; Lagos et al. 2018) to directly model the baryonic physics involved in galaxy formation. A limitation of the first approach compared to the second is that it is descriptive rather than predictive. Gas dynamics simulations are currently limited to relatively small volumes with large shot noise on linear bias (Governato et al. 2007). To address clustering on scales of hundreds of Mpc, semi-analytics models are therefore a sensible choice. The results can in return serve as useful reference points for HOD/SHAM analyses (e.g. Chaves-Montero et al. 2016; Guo et al. 2016), as well as for comparisons with hydro simulations such as EAGLE (Crain et al. 2017) and IllustrisTNG (Springel et al. 2018), which would aid in building a better understanding of the broad features of galaxy evolution.

In this paper, we use the semi-analytic model GALFORM (Cole et al. 2000), specifically an updated variant of Gonzalez-Perez et al. (2018) to model the linear galaxy bias in the standard  $\Lambda$ CDM cosmology with Planck (2015) parameters (Ade et al. 2016), for various IR/optical/UV bands, emission lines, and intrinsic galaxy properties.

This paper is organized as follows. In Section 2, we describe our dark matter  $N$ -body simulation with a semi-analytic galaxy

formation model, the techniques used for measuring the large-scale biases, and a heuristic five-parameter model for fitting. In Section 3, we show the bias dependence on galaxy properties, emission lines and continuum bands, following up with the large-scale galaxy bias as a function of wavelength. In Section 4, we discuss the comparison of our model with existing surveys and forecasts, as well as the limitations of this work. In Section 5, we conclude with a short synopsis of the paper.

## 2 SIMULATIONS AND METHODS

### 2.1 Simulations and galaxy formation model

The SURFS suite consists of  $N$ -body simulations of a periodic volume of side lengths from 40 to 900  $h^{-1}$ Mpc assuming a Planck (2015) cosmology, where  $h$  is the dimensionless Hubble parameter. The simulation used in this work, L210N1536, has a side length of 210  $h^{-1}$ Mpc and number of dark matter particles of 1536<sup>3</sup>. This choice of parameters allows us to resolve the host haloes of galaxies with stellar masses of  $10^8 h^{-1} M_{\odot}$  at  $z = 0$ , with the nominal requirement that the host dark matter haloes of such galaxies are resolved with 100 particles, necessary to get accountable halo masses, positions, and velocities. This simulation has been run with a memory lean version of the GADGET2 code on the Magnus supercomputer at the Pawsey supercomputing centre. For a detailed description of the simulations refer to Elahi et al. (2018), and to Poulton et al. (2018) for a description and performance demonstration of the merger trees used here.

The halo and subhalo catalogues are constructed by VELOCIRAPTOR (Elahi, Thacker & Widrow 2011). This code first identifies haloes using a 3D Friends-of-Friends(FOF) algorithm in configuration space and then identifies subhaloes using a 6D phase-space FOF algorithm on particles that differ dynamically from the dark matter background. We run DHALOS in the SURFS halo catalogues to adapt VELOCIRAPTOR outputs to GALFORM inputs. DHALO is a tool developed to produce and clean merger trees, which form the basis of GALFORM (see Jiang et al. 2014). We compare our halo mass function with previous works by using the virial mass defined as  $M_{\Delta} = 4\pi R_{\Delta}^3 \Delta \rho_{\text{crit}}/3$ ,  $\Delta = 200$ , where  $\rho_{\text{crit}}$  is the critical density of the universe and  $R_{\Delta}$  is the radius that encloses this mass. In Fig. 1, we present the evolving halo mass function compared with Sheth, Mo & Tormen (2001) (hereafter SMT01) and Tinker et al. (2010) (hereafter T10) models in the upper panel and the residuals relative to SMT01 in the lower panel. As shown, we find good agreements with the T10 model, with an average value of uncertainties less than 10 per cent above the halo mass limit ( $2.2 \times 10^{10} h^{-1} M_{\odot}$ ). The deviation from the SMT01 model is due to the difference of definition of halo mass as explained in Tinker et al. (2008) and T10.

Galaxy formation is approximated as a two-stage process: structure forms by hierarchical clustering in the dark matter and baryons then fall into the gravitational potential wells to form galaxies by gas cooling, star formation, feedback, and stellar evolution (Cole et al. 2000). The GALFORM model explicitly accounts for: (1) the shock-heating and radiative cooling of gas inside dark matter haloes that drive the formation of gaseous galactic discs; (2) star formation in galaxy discs and in bulges (i.e. starbursts); (3) the growth of supermassive black holes and feedback from supernovae, active galactic nucleus (AGN) as well as photoionization of the intergalactic medium; (4) galaxy mergers and bar instabilities which can drive bursts of star formation and lead to the formation of spheroids; (5) calculation of the sizes of discs and spheroids; (6) chemical enrichment of stars and gas; (7) calculation of galaxy

<sup>1</sup><https://www.sdss3.org/surveys/boss.php>

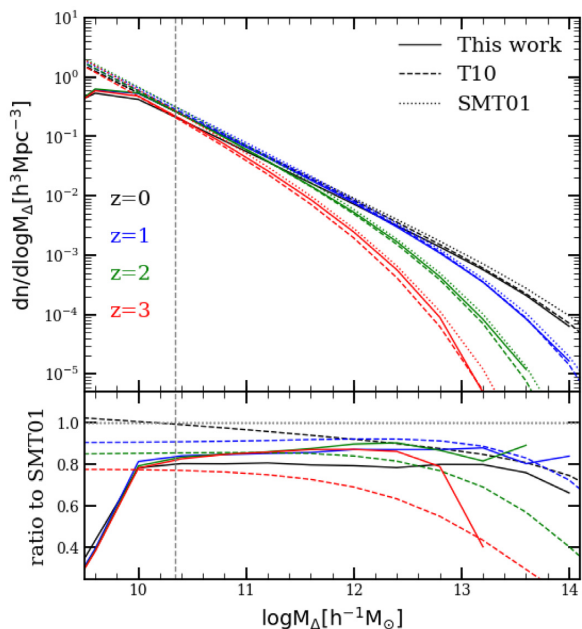
<sup>2</sup><https://www.sdss.org/surveys/eboss>

<sup>3</sup><https://www.desi.lbl.gov>

<sup>4</sup><https://www.lsst.org>

<sup>5</sup><https://www.4most.eu/cms>

<sup>6</sup><https://www.euclid-ec.org>



**Figure 1.** Halo mass function at four redshifts. The solid lines are the measurements from the L210N1536 SURFS simulation. The dotted and dashed lines show the SMT10 and T10 prediction, respectively, calculated using HMFALC (Murray, Power & Robotham 2013). The grey dashed vertical line shows the halo mass limit  $2.2 \times 10^{10} h^{-1} M_{\odot}$ . Note that we use the same colour caption for redshifts in all bias plots as this one unless captioned differently. The lower panel shows the residuals relative to SMT01 model.

stellar luminosities from the predicted star formation and chemical enrichment histories of a stellar population synthesis model; (8) nebular emission-line luminosities and equivalent widths; (9) dust attenuation. We use the version of GALFORM based on Gonzalez-Perez et al. (2018), hereafter GP18, to investigate the large-scale galaxy biases. This model assumes a single initial mass function (IMF) building upon the previous versions (Gonzalez-Perez et al. 2014), which is the major difference from another widely used GALFORM version (Lacey et al. 2016, hereafter L16). The GP18 semi-analytical model has incorporated the merger scheme used in Simha & Cole (2017) and the gradual stripping of hot gas when satellite galaxies are merging into central galaxies (Lagos et al. 2014).

We specified this model with a set of free parameters which were chosen to provide a reasonable match to the  $K$ -band luminosity function from  $z = 2$  to  $z = 0$  and the  $b_j$  luminosity function at  $z = 0$ . We adopt the photoionization model (equation 5) of Kim et al. (2015) with reionization parameters of a circular velocity cut-off  $v_{\text{cut}} = 50 \text{ km s}^{-1}$  at a redshift of reionization  $z_{\text{cut}} = 10$  and the fitting parameter  $\alpha_v = -0.82$  in the notation of Sobacchi & Mesinger (2014). We also take the star formation efficiency parameter of the molecular gas  $\nu_{\text{sf}} = 0.8 \text{ Gyr}^{-1}$  (equation 7 in L16) and the ratio of cooling/free-fall time  $\alpha_{\text{cool}} = 0.7$  (equation 12 in L16) which is an adjustable AGN feedback parameter (more galaxies to be affected by AGN feedback with larger values). The relevant figures used for the adjustment of parameters are shown in Appendix B and the undiscussed parameters are the same as used in Gonzalez-Perez et al. (2018).

The model divides the baryons into five different components: hot gas for cooling in haloes, a reservoir of gas ejected by feedback processes, cold gas, stars, and central black holes in galaxies.

We assumed that the galaxies have separate disc and spheroid components, which can both contain stars and cold gas. We split the cold gas into atomic hydrogen and molecular hydrogen (Lagos et al. 2011b); this distinction is explicitly made throughout the model at every time-step. The accreted gas from the halo is added to the disc. The subsequent galaxy mergers and disc instabilities can transfer the gas into a starburst component in the spheroid. Thus, we assume two separate modes of star formation: the quiescent mode (in the disc) and the starburst mode (in the spheroid). We calculate the star formation rate in the disc from the molecular gas using the empirical relation in Blitz & Rosolowsky (2006), which is based on observations of nearby star-forming disc galaxies as described in Lagos et al. (2011a). For star formation in bursts, we assume all of the cold gas is molecular. Here, we compute the SFR bias including both the quiescent and starburst modes. We also calculate the stellar mass bias by including all the stars in the galaxy.

Broad-band luminosities and absolute magnitudes are calculated from the stellar SEDs of galaxies using a stellar population synthesis model (L16) based on stellar evolution models. This model also includes a simple model for emission lines in star-forming galaxies that uses the number of ionizing photons and the metallicity of the cold, star-forming gas to predict emission-line luminosities based on the properties of a typical H II region (Stasińska 1990) (see GP18 for an expanded discussion of the modelling of emission lines).

To simulate the effects of dust extinction, GALFORM applies a physical model of absorption and emission of radiation by dust. The dust is assumed to be present in two components: diffuse dust (75 percent) and molecular clouds (25 percent) based on observations of nearby galaxies (Granato et al. 2000). This model includes a self-consistent model for the reprocessing of starlight by dust, in which the UV, optical, and near-IR light are absorbed by dusts and radiated at IR and sub-mm wavelengths. The dust absorption is based on radiative transfer and the temperature of the dust emission can be solved for by energy balance (see L16 for details).

Redshifted magnitudes are needed for predicting the linear bias of future surveys. With the intrinsic properties of galaxies and their cosmological redshifts, we are able to evaluate observer-frame absolute magnitudes by

$$M = -2.5 \log \left[ \frac{\int L_v(v_e) R(v_o) dv_e}{L_{v_o} \int R(v_o) dv_o} \right], \quad (1)$$

where  $L_v(v_e)$  is the emitted luminosity per unit frequency and  $L_{v_o}$  is the reference luminosity. The emitted (rest-frame) frequency  $v_e$  is related to the observed (observer-frame) frequency  $v_o$  by  $v_e = v_o(1+z)$  and  $R(v_o)$  is the filter response of a specified photometric band on the observer frame (equation 13 in Merson et al. 2013). Note that we use the symbol  $M$  without subscripts as the absolute magnitude to differentiate it from the mass symbol with subscripts. The effect of dust attenuation is included throughout this paper.

For the galaxy samples chosen by continuum filters, we use the observer frame quantities to investigate the bias dependence. The UV/optical/IR filters correspond to those used in the GALEX/SDSS/UKIRT surveys, except that the  $Y$ -band filter is from the UKIDSS survey.

## 2.2 Large-scale bias

Based on the current galaxy formation models, galaxies are formed in dark haloes, therefore the understanding of halo bias is an essential component of any theory of galaxy bias (Smith et al. 2007).

### 2.2.1 Halo bias

The halo bias is determined by the relative abundance of haloes in different large-scale density environments. There are two well-known models to describe the halo bias: the peak bias (Bardeen et al. 1986) and peak-background split model (Kaiser 1984; Cole & Kaiser 1989; Mo & White 1996). The peak bias model is built upon the simplified assumption that collapsed structures form from peaks in the Gaussian initial density field. Given a uniform density background at the early age of the universe, the peak bias model works well till the LSSs emerge and act like a local modification of the background density. The more general peak-background split model decomposes the density field into a long-wavelength and short-wavelength part. These models and the numerical calibration in T10 have given us valuable insights into the physics of the halo bias. The most obvious result is the strong dependence of large-scale halo bias on halo mass and redshift. Of course, there are other halo variables that can also produce strong trends even at fixed halo mass, such as the local tidal environment (e.g. Paranjape, Hahn & Sheth 2018).

We review some basic theories for the halo bias. The two central quantities are the dark matter overdensity,  $\delta$ , and halo overdensity,  $\delta_h$ ,

$$\delta = \frac{\rho - \bar{\rho}}{\bar{\rho}}, \delta_h = \frac{n_h - \bar{n}_h}{\bar{n}_h}, \quad (2)$$

where  $\rho$  and  $\bar{\rho}$  are the dark matter density and the mean density, respectively, and  $n_h$  and  $\bar{n}_h$  are the halo number density and its mean number density, respectively.

The halo bias is essentially a relation between  $\delta$  and  $\delta_h$ , which can be Taylor expanded to (Fry & Gaztanaga 1993; Mo, Jing & White 1997; Pollack, Smith & Porciani 2012)

$$\delta_h = b_0 + b_1\delta + \frac{b_2}{2}\delta^2 + \dots, \quad (3)$$

where the halo bias is assumed to be local and deterministic. On large scales, it is commonly truncated to first order and the relation simply becomes linear, where  $b_0 = 0$  owing to the fact that  $\langle \delta_h \rangle \equiv \langle \delta \rangle \equiv 0$  by definition (Fry & Gaztanaga 1993). These assumptions leave  $b_1$  as the only relevant parameter and it is commonly regarded as the large-scale halo bias. We note that the halo bias should in principle also contain terms describing non-locality and stochasticity which become important at quasi-linear scales (see equation 2.135 in Desjacques, Jeong & Schmidt 2018), but we are focusing only on measuring the linear bias here.

By looking at equation (3) in Fourier space, we can also define a practical measurement of halo bias via the relation between the halo and matter power spectrum

$$b_h(k) = \sqrt{\frac{P_h(k)}{P_m(k)}}, \quad (4)$$

where  $P_m(k) = \langle |\delta(\mathbf{k})|^2 \rangle$  is the dark matter power spectrum,  $\delta(\mathbf{k})$  is the Fourier representation of dark matter overdensities, and  $P_h(k)$  accordingly is the halo power spectrum. On large scales, it is easy to see that equation (4) can be derived from equation (3), which means  $b_h(k) = b_1$  when the wavenumber  $k$  (i.e. the magnitude of the wave vector  $\mathbf{k}$ ) is small in Fourier space. Hereafter, we denote the large-scale halo bias as  $b_h$  and specify the measurements in Section 2.3.

### 2.2.2 Galaxy bias

Galaxies do not trace the mass in the same way as haloes do. The difference is that galaxy formation proceeds with an efficiency which depends on halo properties in a non-linear fashion. In the lowest mass haloes, feedback from supernovae and UV background ionization prevent efficient star formation while in the high-mass haloes gas is unable to cool efficiently (Benson et al. 2000) as is heated by AGN (Bower et al. 2006). A comprehensive overview of galaxy bias based on perturbation theory can be seen in Desjacques et al. (2018). Following the definition of halo bias, the galaxy bias can be written as

$$b_g(k) = \sqrt{\frac{P_g(k)}{P_m(k)}}, \quad (5)$$

where  $P_g(k)$  is the galaxy power spectrum.

Statistically speaking, the galaxy bias can also be expressed via the halo model (see Cooray & Sheth 2002, for a review). According to this model, the distribution of galaxies depends on how they populated haloes. Hence, we can split the galaxy power spectrum into two components, the 1- and 2-halo terms:  $P_g(k) = P_g^{1h}(k) + P_g^{2h}(k)$ . The 1-halo term  $P_g^{1h}(k)$  is determined by the density profiles of galaxy pairs in shared haloes and the halo mass function. The 2-halo term  $P_g^{2h}(k)$  contains the contribution of the galaxies in different haloes.

On large scales, only the 2-halo term is important and it can be linked with the halo power spectrum if one knows the galaxy distribution at given mass haloes. In this case, we can also rewrite the equation (5) by weighting the large-scale halo bias with the number of galaxies contained in the given mass haloes (see Mo, Van den Bosch & White 2010, for the detailed derivation)

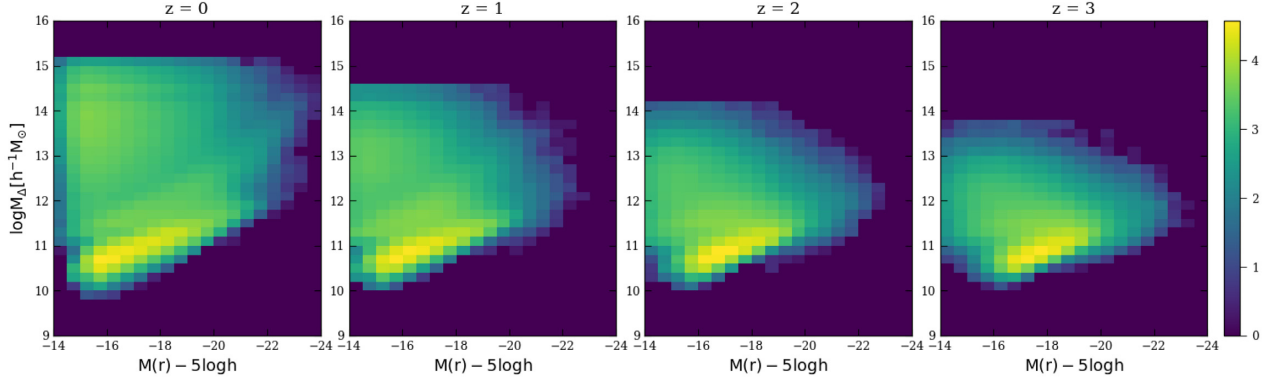
$$b_g(x) = \frac{\int b_h(M_h)\phi(x|M_h)n(M_h)dM_h}{\int \phi(x|M_h)n(M_h)dM_h}, \quad (6)$$

where  $n(M_h)$  is the halo mass function and  $\phi(x|M_h)$  denotes the conditional galaxy distribution function, such as conditional galaxy mass, star formation rate, and luminosity function. The  $\phi(x|M_h)$  describes the distribution of galaxies in haloes of a given halo of mass  $M_h$ , which can be computed by the conditional probability distribution function (PDF) of host halo mass against galaxy property  $x$  from Fig. 2. The looming bimodal distribution from  $z = 3$  to  $z = 0$  is contributed by the central galaxies and satellites due to increased merger events. Here,  $b_g(x)$  is the large-scale differential galaxy bias, which satisfies the observational selection criteria of galaxy property  $x$ . Different properties  $x$  will be discussed in Section 2.4. We note that this equation emphasizes that the galaxy bias only depends on halo mass (i.e. zero galaxy assembly bias), which in practice may not be the case as indicated by many recent studies (e.g. Zentner, Hearin & van den Bosch 2014; Contreras et al. 2019).

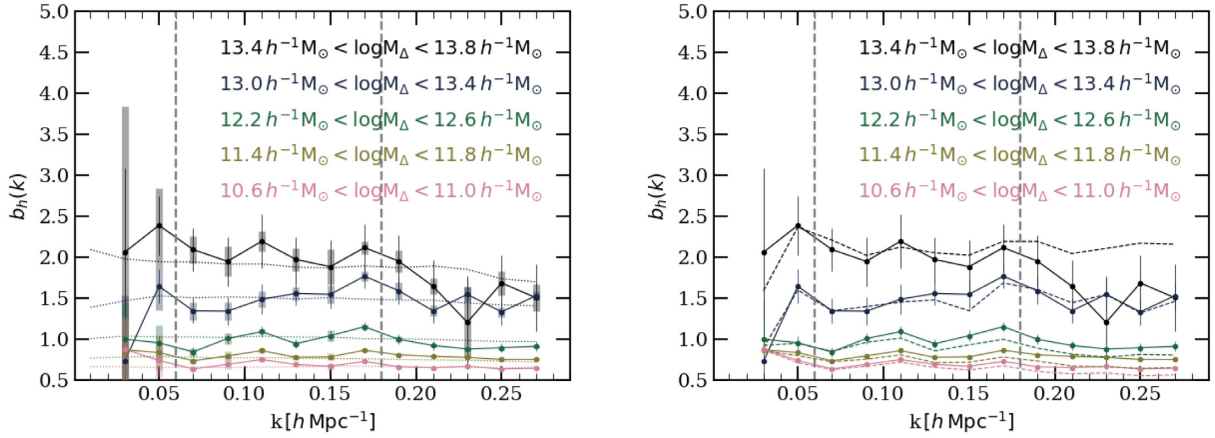
Following the same spirit, we note  $b_g(\geq x)$  to be the large-scale cumulative galaxy bias above a limited value of  $x$ , which can be expressed as follows

$$b_g(\geq x) = \frac{\int_{x'=x}^{+\infty} b_g(x')\phi(x')dx'}{\int_{x'=x}^{+\infty} \phi(x')dx'}, \quad (7)$$

where  $\phi(x')$  is the distribution function for a galaxy property, such as a luminosity.



**Figure 2.** Two-dimensional distribution function of galaxies as a function of  $r$ -band luminosity and halo mass. Results are shown at  $z = 0, 1, 2, 3$  from left-hand panel to right-hand panel. The colour map gives the log of the galaxy number.



**Figure 3.** Large-scale halo bias as a function of wavenumber  $k$  at  $z = 0.33$ . Left: the solid and dotted lines are the measurements from the L210N1536 and L900N2048 SURFS simulations, respectively, using equation (4). The narrower solid errors are estimated by equation (9) while the wider shaded errors are calculated by the ‘jackknife’ method. Right: the dashed lines are measured by the cross-power spectrum between halo and dark matter over the dark matter power spectrum from the L210N1536 simulation while the solid lines are the same as on the left-hand panel. The grey dashed vertical lines indicate the lower and upper limits of  $2 \times 2\pi/L_{\text{box}}$  and  $0.18h^{-1}\text{Mpc}$  in  $k$  space, where the  $L_{\text{box}} = 210h^{-1}\text{Mpc}$ . Only five mass bins are shown for clarity. The mass bin width is 0.4 in the logarithmic space.

### 2.3 Measurements of bias and uncertainty

To measure the bias from equations (4) and (5), one needs to calculate the power spectrum from the halo and galaxy catalogues. We first binned the haloes and galaxies into a  $256^3$  density mesh by using a top-hat smoothing and computed the Fourier transform to evaluate the power spectrum. All the power spectra are shot-noise subtracted and we divide each Fourier mode by a sinc-function to correct for the top-hat gridding effect. Finally, we average the orientation-dependent power spectrum over a spherical shell in  $k$ -space in order to take out the orientation dependence and get  $P_h(k)$  and  $P_g(k)$ .

Since we have only one simulation to produce the galaxy samples, we rely on the theory of Gaussian covariance matrices of the power spectrum to measure the uncertainty on the bias. The commonly used expression for the power spectrum variance was derived by Feldman, Kaiser & Peacock (1994), Tegmark (1997):

$$\sigma_{P_g(k)} = \sqrt{\frac{2}{n_{\text{modes}}}} \left[ P_g(k) + \frac{1}{\bar{n}_g} \right], \quad (8)$$

where  $\bar{n}_g$  refers to the number density of galaxies which satisfy the given selection criteria and  $n_{\text{modes}} = 4\pi k^2 \delta k / (2\pi)^3$  is the number

of Fourier modes in the spherical shell of width  $\delta k$  when the volume  $V \gg (2\pi/k)^3$ . Finally, we provide the uncertainties of large-scale galaxy bias through error propagation as follows

$$\sigma_{b_g(k)} \approx \sqrt{\frac{1}{n_{\text{modes}} \bar{n}_g P_m(k)}}. \quad (9)$$

A derivation and verification with simulations can be found in Appendix A. We also employ the commonly used ‘jackknife’ method to estimate the error on the bias in comparison with this derived form as discussed later.

Fig. 3 shows the halo bias as a function of wavenumber  $k$  for haloes of varying mass at  $z = 0.33$ . The solid and dotted lines are measured from the L210N1536 and L900N2048 SURFS simulations, respectively, using the equation (4) on the left-hand panel. As the linear bias is independent of scale, one can see that the halo bias generally stays constant for  $2 \times 2\pi/L_{\text{box}} < k < 0.18h^{-1}\text{Mpc}$  and even applicable to larger  $k$  in the low-mass bins for the L210 simulation, thus we can directly measure the large-scale halo bias by averaging  $b_h(k)$  over those modes. At higher redshifts, these limits on  $k$  are more conservative since the non-linear scale keeps extending towards the smaller  $k$  as the universe

evolves. Comparing the halo bias of L210 simulation with that of L900 simulation, we are confident that the L210 simulation can be used to study the large-scale bias for a variety of halo mass ranges, except for the most massive samples due to the low number statistics. The shaded error bars for the L210 simulation are estimated by the ‘jackknife’ method. To do so, we calculate the power spectrum for eight subsamples with removing one octant of the box in each subsample, then estimate the errors on the  $b_h(k)$  as indicated by the shaded area. Equation (9) is generally consistent with this jackknife method for the less massive haloes ( $< 10^{13} h^{-1} M_\odot$ ), but tends to overestimate the uncertainties for the massive samples by a factor of two in the considered  $k$  scales, which makes equation (9) more conservative. Therefore, we will carry on with the derived form and exclude the points with relative errors larger than 20 per cent in all the bias plots.

We also check the halo bias as defined by the cross-power spectrum between halo and dark matter over the dark matter power spectrum indicated by the dashed lines on the right-hand panel of Fig. 3. As seen, this measure, while immune to shot-noise effects, could be contaminated with higher order bias on larger  $k$  scales than the definition of the equation (4) for the low-mass haloes and this is in line with the findings in Smith et al. (2007) and Vlah et al. (2013). Considering that the effects of shot-noise are outside the considered  $k$  scales (i.e. between the dashed vertical lines), we therefore chose to use equations (4) and (5) for the following analyses on the large-scale biases.

We then average the  $b_g(k)$  over small  $k$  scales weighted by number of modes, which gives the large-scale galaxy bias as

$$b_g = \frac{\sum_k w_k b_g(k)}{\sum_k w_k}. \quad (10)$$

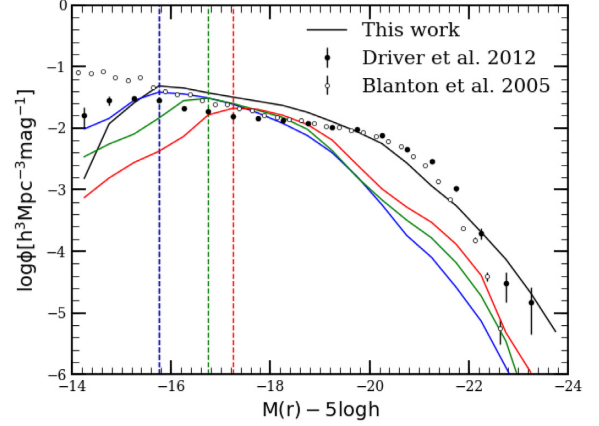
The variance of this weighted average is given by

$$\sigma_{b_g}^2 = \frac{\sum_k w_k (b_g(k) - b_g)^2}{\sum_k w_k}, \quad (11)$$

where  $w_k = 1/\sigma_{b_g(k)}^2$  are the weights given to each measurement.

The galaxy bias depends on the selection of galaxies. Using  $x$  as a general placeholder for a scalar galaxy property (e.g. a luminosity, an emission-line strength or the stellar mass), the bias  $b_g(x)$  and the cumulative bias  $b_g(\geq x)$  can only be determined down to a certain value  $x_{\min}$ , below which the mass resolution of the simulation is insufficient. To determine  $x_{\min}$  for each property, we first apply a stellar mass cut of  $10^8 h^{-1} M_\odot$  to all galaxies. Above this limit, the model is roughly complete in stellar mass. We then plot the space density function of  $x$  (e.g. the luminosity function if  $x$  is a luminosity) and determine the point where this function peaks. Finally, we compare the peak positions at the four redshifts ( $z = 0, 1, 2, 3$ ) and set  $x_{\min}$  equal to the largest peak value. For instance, Fig. 4 shows the galaxy  $r$ -band luminosity function compared with observations. The vertical dashed lines indicate the resolution limit for each redshift. The red dashed line represents the magnitude of  $-17.2$ , which is the brightest value compared with that of other redshifts, therefore we take it as the lower limit of  $r$ -band bias plot. We apply this same principle to all other galaxy bias plots as shown in Section 3. Note that we use the stellar mass cut only for determining the lower limits of biases, not for excluding the samples at given selection criteria.

We have thus introduced two methods to compute the large-scale galaxy bias, either through averaging equation (5), i.e. equation (10) or through the halo model, i.e. equations (6) and (7). For clarification, we refer these two methods as ‘Measurement’ and ‘Halo model’ in the upcoming bias plots.



**Figure 4.** The  $r$ -band galaxy luminosity function at redshifts  $z = 0$  (black), 1 (blue), 2 (green), 3 (red). The dashed vertical lines show the peaks of these functions, which were used to define the resolution limit of the model. The black dots and open circles indicate the  $r$ -band luminosity function measured by Driver et al. (2012) and Blanton et al. (2005) at  $z = 0$ .

## 2.4 Heuristic five-parameter model

Previous bias models in the literature usually concentrate on a specific galaxy property, such as the galaxy clustering dependence on galaxy colour in a single-band filter (Zehavi et al. 2011), or the redshift dependence (Clerkin et al. 2015). The T10 model uses a universal fitting function which accurately accounts for the mass, redshift, and cosmology dependence of halo bias, but cannot easily be linked to observable galaxy properties.

To improve on these models, we now fit different biases using the five-parameter model.

$$b_g(x, z) = a + b(1 + z)^e (1 + \exp[(x - c)d]), \quad (12)$$

where  $x$  is a galaxy (or halo) property,  $z$  is the redshift and  $a, b, c, d$ , and  $e$  are the five parameters to be fitted. The combination of  $a, b$ , and  $e$  acts as a normalization, whereas  $c$  and  $d$  represent the upturn point and slope on the high ends capturing the major differences of galaxy biases. We use this same formula for fitting the differential bias and the cumulative bias. If  $x$  denote a magnitude (defined as negative log of the luminosity), we replace  $x - c$  in equation (12) by  $c - x$  to cope with the fact the smaller magnitudes correspond to brighter objects.

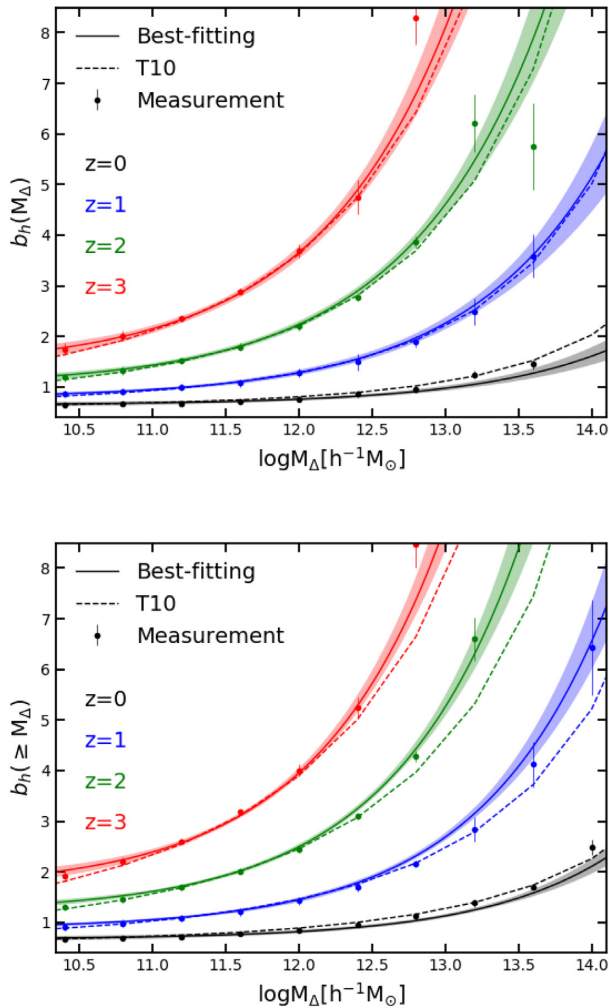
Assuming that the distribution of errors follows a Gaussian, we can write the likelihood function as

$$\ln P(b_g(x, z)|a, b, c, d, e) = -\frac{1}{2} \sum \left[ \frac{(b_g(x, z) - b_g)^2}{\sigma_{b_g}^2} + \ln(2\pi\sigma_{b_g}^2) \right], \quad (13)$$

where  $b_g$  is the large-scale bias measured from the equation (10) and  $\sigma_{b_g}$  is the corresponding error measured from equation (11). We use MULTINEST<sup>7</sup> to fit our model to the ‘Measurement’ and accept the median of posterior samples as the best estimate of each parameter. MULTINEST is an efficient and robust Bayesian inference tool based on a nested sampling technique (Skilling 2004), which allows model fitting and produces the posterior samples.

We test our measurements and fitting formula against the well-calibrated large-scale halo bias model by T10. Fig. 5 shows the

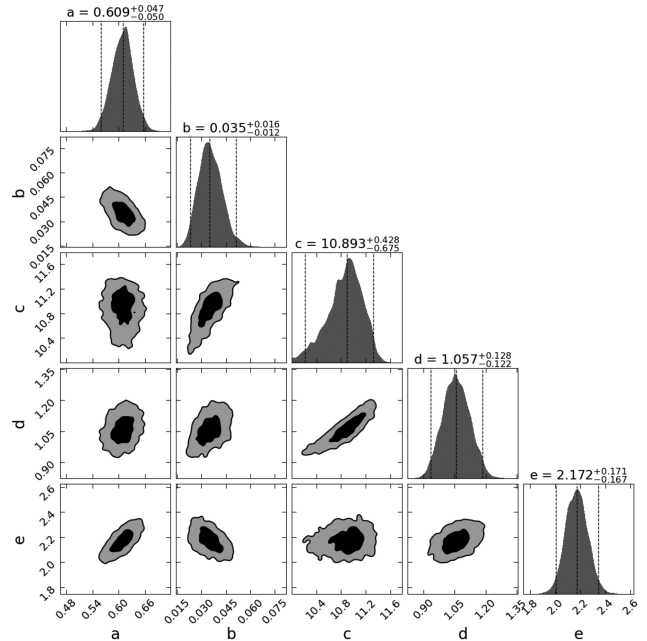
<sup>7</sup><https://github.com/JohannesBuchner/PyMultiNest>



**Figure 5.** Large-scale halo bias as a function of halo mass at four redshifts. The upper and lower panels show the differential and cumulative biases, respectively. The filled symbols are directly measured by averaging equation (5) on large scales. Only measurements with a relative uncertainty below than 20 per cent are shown. The error bars are estimated by linearly propagating the shot noise uncertainties of the power spectrum. The solid lines show the best fits of equation (12) and the dashed lines show the T10 results where we utilize equation (7) to compute the T10 prediction for the cumulative halo bias in the lower panel. The colour-coded regions are the 68 per cent credible intervals in the halo bias estimated from the posterior samples. Note that we use the 3D defined halo mass here to compare with the T10.

large-scale halo bias as a function of halo mass at four redshifts. Overall, our measurements for the differential halo bias show great agreement with the T10 model. The deviation on the high-mass end at high redshifts is due to the low number statistics in the simulation box with a side length of  $210 h^{-1}\text{Mpc}$ . It is clear that our five-parameter fitting formula can do as fine a job as the six-parameter model does in T10, except for the high-mass end at redshift 0. The notable difference between our measurement and T10 prediction on the high-mass ends in the lower panel indicates their differences on the halo mass function. The posterior distributions of the five fitted parameters for the large-scale differential halo bias are shown in Fig. 6, demonstrating that these parameters are unimodal and well converged.

The best-fitting parameters for all the selected samples can be seen from Tables 1 and 2 for the differential and cumulative biases,



**Figure 6.** The posterior distributions for the five parameters of equation (12) for the large-scale differential halo bias. The contour levels in the 2D marginalized posteriors are 1 and  $2\sigma$  while the dashed lines in the 1D marginalized posteriors span the  $2\sigma$  (95 per cent) credible interval.

respectively, which show the differential and cumulative biases are almost identical with the maximum percentage of difference less than 20 per cent. The similarity between the differential and cumulative measurements is a direct implication of the steepness of the galaxy mass function. This steepness means that at any limit, the cumulative bias is always dominated by the objects near the limit. Thus, for brevity, we will focus on the analysis of large-scale differential biases in the next sections.

### 3 RESULTS

#### 3.1 Large-scale galaxy bias dependence on physical properties

We show the large-scale galaxy bias as a function of stellar mass, cold gas mass, cold atomic gas (HI) mass, and SFR in Fig. 7. These properties are amongst the most critical properties in galaxy evolution and their bias will also help us understand the biases of various other observable properties. The HI bias can be compared against measurements of the completed HI surveys, such as HIPASS and ALFAFA (Barnes et al. 2001; Haynes et al. 2018), and forecast for the next generation of extragalactic HI surveys, such as MIGHTEE (Jarvis et al. 2017), WALLABY (Koribalski et al., in preparation), and SKA (Yahya et al. 2015).

One can see that the large-scale galaxy biases from the ‘Measurement’ of selected galaxy samples are in good agreement with the ‘Halo model’ prediction, which verifies the ‘Measurement’ in a statistical sense and indicates a lack of assembly bias from this particular simulation in comparison to the findings in Paranjape et al. (2018). The same upward trend towards the high value ends between the halo mass and galaxy physical property biases demonstrates a close correlation between these galaxy properties and halo mass as the galaxy formation models predict. Compared to the cold gas mass, HI mass, and SFR biases, the galaxy stellar mass bias shows a steeper slope, which implies that the galaxy stellar mass is

**Table 1.** The best-fitting parameters of our five-parameter model for the large-scale *differential* bias for different observational bands, emission lines, physical properties of galaxies and halo mass as well.

Halo mass	x	a	b	c	d	e
$M_h$	$\log M_h [h^{-1} M_\odot]$	$0.609 \pm 0.024$	$0.035 \pm 0.007$	$10.893 \pm 0.286$	$1.057 \pm 0.066$	$2.172 \pm 0.086$
Physical property	x	a	b	c	d	e
$M_*$	$\log M_* [h^{-1} M_\odot]$	$1.146 \pm 0.015$	$0.069 \pm 0.008$	$10.527 \pm 0.065$	$1.435 \pm 0.145$	$2.172 \pm 0.078$
$M_{\text{gas}}$	$\log M_{\text{gas}} [h^{-1} M_\odot]$	$0.86 \pm 0.034$	$0.076 \pm 0.023$	$10.483 \pm 0.605$	$0.925 \pm 0.431$	$1.946 \pm 0.119$
$M_{\text{HI}}$	$\log M_{\text{HI}} [h^{-1} M_\odot]$	$0.872 \pm 0.033$	$0.071 \pm 0.026$	$10.319 \pm 0.596$	$1.206 \pm 0.571$	$2.004 \pm 0.167$
SFR	$\log \text{SFR} [h^{-1} M_\odot / \text{Gyr}]$	$0.868 \pm 0.031$	$0.129 \pm 0.034$	$11.068 \pm 0.781$	$0.537 \pm 0.313$	$1.643 \pm 0.082$
Emission line	x	a	b	c	d	e
H $\alpha$	$\log H \alpha [\text{erg s}^{-1}]$	$0.831 \pm 0.032$	$0.097 \pm 0.023$	$42.49 \pm 0.571$	$0.639 \pm 0.3$	$1.736 \pm 0.108$
H $\beta$	$\log H \beta [\text{erg s}^{-1}]$	$0.829 \pm 0.032$	$0.12 \pm 0.028$	$42.147 \pm 0.252$	$1.038 \pm 0.468$	$1.676 \pm 0.129$
[O II]	$\log [\text{O II}] [\text{erg s}^{-1}]$	$0.786 \pm 0.036$	$0.134 \pm 0.026$	$42.918 \pm 0.412$	$1.407 \pm 0.558$	$1.748 \pm 0.113$
Lya	$\log \text{Ly}\alpha [\text{erg s}^{-1}]$	$0.811 \pm 0.032$	$0.117 \pm 0.022$	$43.417 \pm 0.139$	$1.279 \pm 0.522$	$1.726 \pm 0.115$
Bands	x	a	b	c	d	e
(FUV NUV u)	$M - 5\text{logh}$	$0.629 \pm 0.187$	$0.338 \pm 0.184$	$-21.201 \pm 0.291$	$1.012 \pm 0.235$	$0.997 \pm 0.305$
$g$	$M(g) - 5\text{logh}$	$1.009 \pm 0.025$	$0.09 \pm 0.017$	$-21.118 \pm 0.09$	$1.41 \pm 0.198$	$1.985 \pm 0.133$
$r$	$M(r) - 5\text{logh}$	$1.026 \pm 0.024$	$0.095 \pm 0.017$	$-21.292 \pm 0.116$	$1.194 \pm 0.201$	$1.945 \pm 0.118$
$i$	$M(i) - 5\text{logh}$	$1.045 \pm 0.02$	$0.087 \pm 0.013$	$-21.214 \pm 0.132$	$1.102 \pm 0.115$	$1.999 \pm 0.099$
$z$	$M(z) - 5\text{logh}$	$1.073 \pm 0.017$	$0.066 \pm 0.01$	$-21.134 \pm 0.177$	$0.852 \pm 0.095$	$2.177 \pm 0.097$
$Y$	$M(Y) - 5\text{logh}$	$1.059 \pm 0.018$	$0.081 \pm 0.012$	$-21.403 \pm 0.182$	$0.808 \pm 0.079$	$2.047 \pm 0.096$
$H$	$M(H) - 5\text{logh}$	$1.077 \pm 0.016$	$0.074 \pm 0.009$	$-21.518 \pm 0.148$	$0.739 \pm 0.052$	$2.064 \pm 0.072$
$J$	$M(J) - 5\text{logh}$	$1.077 \pm 0.017$	$0.066 \pm 0.01$	$-21.339 \pm 0.16$	$0.819 \pm 0.065$	$2.183 \pm 0.095$
$K$	$M(K) - 5\text{logh}$	$1.094 \pm 0.015$	$0.069 \pm 0.007$	$-21.681 \pm 0.149$	$0.63 \pm 0.036$	$2.095 \pm 0.069$
( $g \sim K$ )	$M - 5\text{logh}$	$1.062 \pm 0.007$	$0.075 \pm 0.004$	$-21.462 \pm 0.067$	$0.645 \pm 0.018$	$2.044 \pm 0.033$

**Table 2.** The best-fitting parameters of our five-parameter model for the large-scale *cumulative* galaxy bias for different observational bands, emission lines, and physical properties of galaxies.

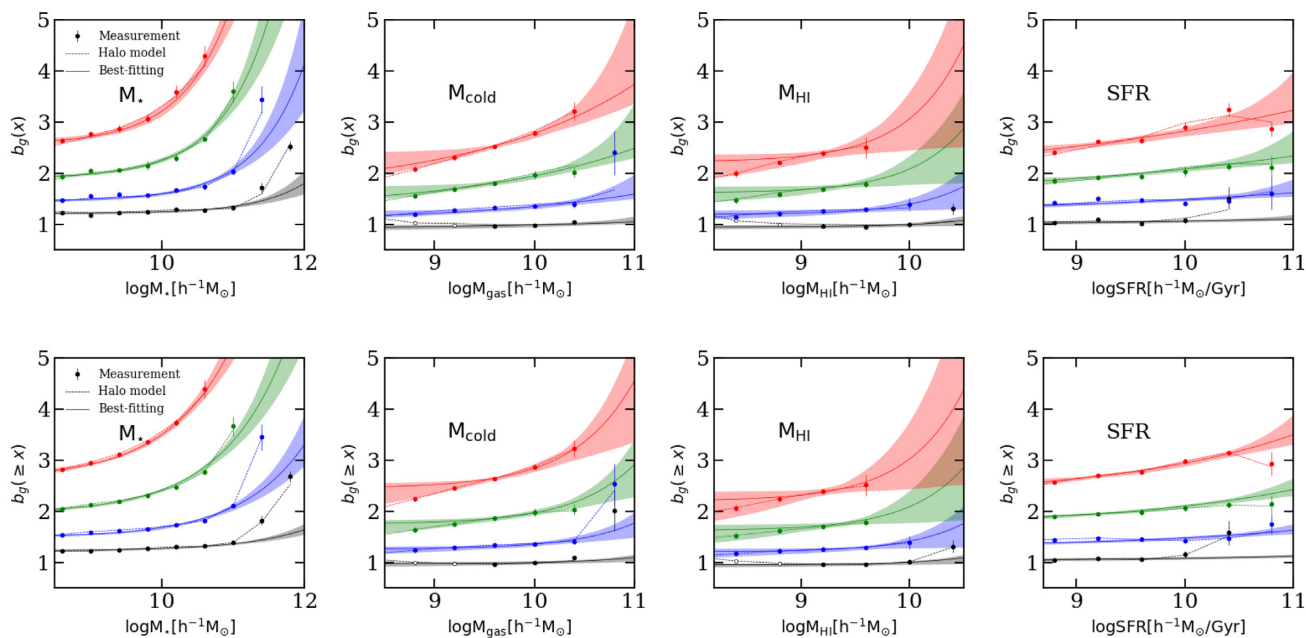
Halo mass	x	a	b	c	d	e
$M_h$	$\log M_h [h^{-1} M_\odot]$	$0.603 \pm 0.025$	$0.066 \pm 0.007$	$11.305 \pm 0.147$	$1.142 \pm 0.055$	$1.99 \pm 0.055$
Physical property	x	a	b	c	d	e
$M_*$	$\log M_* [h^{-1} M_\odot]$	$1.157 \pm 0.013$	$0.075 \pm 0.006$	$10.488 \pm 0.064$	$1.126 \pm 0.098$	$2.149 \pm 0.054$
$M_{\text{gas}}$	$\log M_{\text{gas}} [h^{-1} M_\odot]$	$0.883 \pm 0.032$	$0.082 \pm 0.023$	$10.756 \pm 0.452$	$1.055 \pm 0.524$	$2.015 \pm 0.116$
$M_{\text{HI}}$	$\log M_{\text{HI}} [h^{-1} M_\odot]$	$0.864 \pm 0.03$	$0.083 \pm 0.027$	$10.516 \pm 0.661$	$1.263 \pm 0.577$	$1.94 \pm 0.141$
SFR	$\log \text{SFR} [h^{-1} M_\odot / \text{Gyr}]$	$0.942 \pm 0.025$	$0.098 \pm 0.021$	$11.161 \pm 0.484$	$0.613 \pm 0.259$	$1.886 \pm 0.078$
Emission line	x	a	b	c	d	e
H $\alpha$	$\log H \alpha [\text{erg s}^{-1}]$	$0.844 \pm 0.031$	$0.116 \pm 0.02$	$42.623 \pm 0.132$	$1.186 \pm 0.387$	$1.765 \pm 0.106$
H $\beta$	$\log H \beta [\text{erg s}^{-1}]$	$0.837 \pm 0.036$	$0.135 \pm 0.027$	$42.2 \pm 0.119$	$1.409 \pm 0.395$	$1.681 \pm 0.117$
[O II]	$\log [\text{O II}] [\text{erg s}^{-1}]$	$0.816 \pm 0.03$	$0.118 \pm 0.022$	$42.993 \pm 0.407$	$1.38 \pm 0.535$	$1.855 \pm 0.109$
Lya	$\log \text{Ly}\alpha [\text{erg s}^{-1}]$	$0.834 \pm 0.031$	$0.119 \pm 0.021$	$43.396 \pm 0.102$	$1.526 \pm 0.4$	$1.754 \pm 0.11$
Bands	x	a	b	c	d	e
(FUV NUV u)	$M - 5\text{logh}$	$0.477 \pm 0.202$	$0.479 \pm 0.205$	$-21.209 \pm 0.422$	$0.73 \pm 0.17$	$0.744 \pm 0.194$
$g$	$M(g) - 5\text{logh}$	$1.036 \pm 0.028$	$0.083 \pm 0.02$	$-21.009 \pm 0.115$	$1.035 \pm 0.174$	$2.022 \pm 0.157$
$r$	$M(r) - 5\text{logh}$	$1.035 \pm 0.026$	$0.101 \pm 0.017$	$-21.215 \pm 0.109$	$0.968 \pm 0.146$	$1.893 \pm 0.114$
$i$	$M(i) - 5\text{logh}$	$1.046 \pm 0.02$	$0.096 \pm 0.013$	$-21.165 \pm 0.125$	$0.965 \pm 0.083$	$1.939 \pm 0.088$
$z$	$M(z) - 5\text{logh}$	$1.078 \pm 0.017$	$0.073 \pm 0.01$	$-21.003 \pm 0.17$	$0.751 \pm 0.078$	$2.11 \pm 0.081$
$Y$	$M(Y) - 5\text{logh}$	$1.074 \pm 0.017$	$0.078 \pm 0.01$	$-21.133 \pm 0.173$	$0.718 \pm 0.059$	$2.077 \pm 0.085$
$H$	$M(H) - 5\text{logh}$	$1.084 \pm 0.014$	$0.078 \pm 0.008$	$-21.306 \pm 0.17$	$0.593 \pm 0.038$	$2.037 \pm 0.054$
$J$	$M(J) - 5\text{logh}$	$1.097 \pm 0.013$	$0.066 \pm 0.007$	$-21.163 \pm 0.128$	$0.738 \pm 0.039$	$2.203 \pm 0.066$
$K$	$M(K) - 5\text{logh}$	$1.106 \pm 0.014$	$0.071 \pm 0.007$	$-21.443 \pm 0.167$	$0.552 \pm 0.03$	$2.11 \pm 0.058$
( $g \sim K$ )	$M - 5\text{logh}$	$1.073 \pm 0.006$	$0.075 \pm 0.003$	$-21.114 \pm 0.076$	$0.574 \pm 0.015$	$2.04 \pm 0.026$

fundamentally different from other galaxy properties. The cold gas and HI gas trace the SFR, since the SFR is proportional to the mass in the molecular component which correlates with atomic hydrogen. We note the open circles in the middle panels are not used for the fitting due to the mass resolution limits of cold gas at  $z = 0$ .

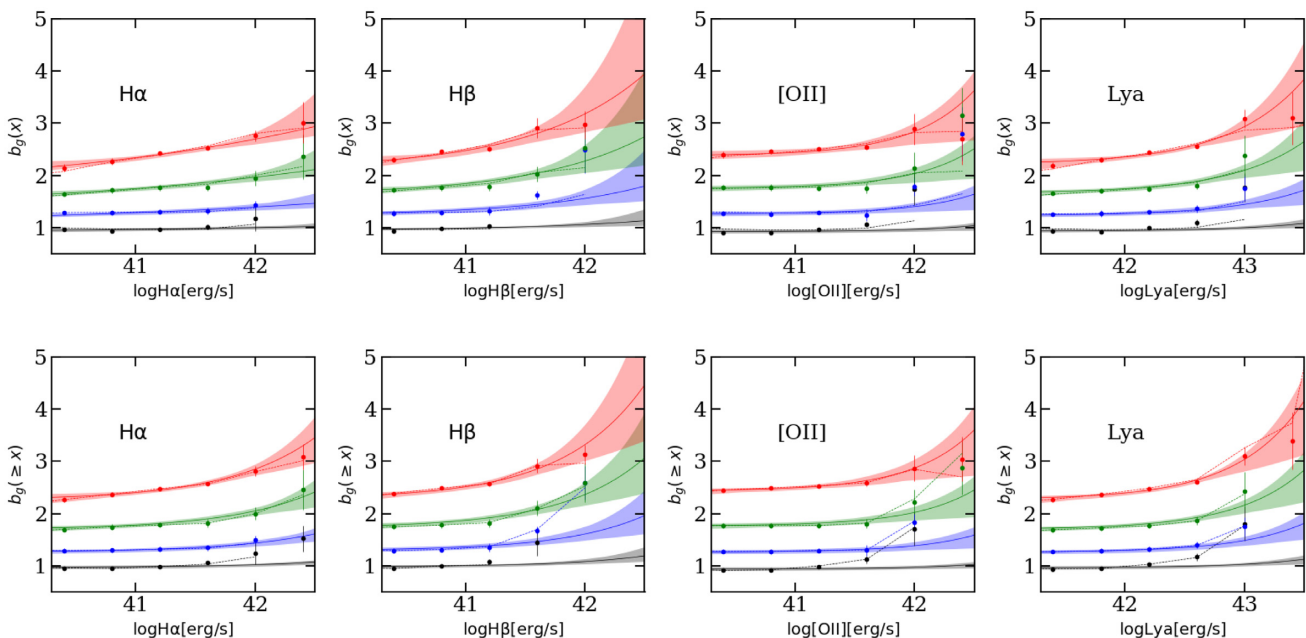
The turn-down feature of the SFR bias at high redshifts on the high star-forming end shows that dense environments have strongly reduced star formation which imposes an antibias effect on the

distribution of star-forming galaxies at early times ( $z \geq 2$ ). The lack of a turn-down feature in the cold gas bias plots implies the relation between cold gas and SFR is also dependent on density environments during that period of time. However, since this turn-down feature only shows in the high SFR end and the drop in numbers makes it difficult to gather enough samples for robust measurements, we look forward to seeing further verification of this trend from future simulations and observations.





**Figure 7.** Large-scale galaxy bias as a function of stellar mass, cold gas mass, cold atomic gas mass, and SFR at redshifts  $z = 0$  (black), 1 (blue), 2 (green), 3 (red). The upper and lower panels show the differential and cumulative biases, respectively. The filled symbols are directly measured by averaging equation (5) on large scales. The error bars are estimated by propagating the shot noise of the power spectrum. The solid lines show the best fits of equation (12) and the dashed lines show the predicted biases from the Halo model. Only measurements with a relative uncertainty below 20 per cent are shown. The open circles are not used for the fitting as they lie below the resolution limit of the simulation as defined in Section 2.3.



**Figure 8.** Large-scale differential (top) and cumulative (bottom) galaxy bias as a function of  $H\alpha$ ,  $H\beta$ ,  $[OII]$ , and  $Lya$  line luminosities from  $z = 0$  to  $z = 3$  (see caption of Fig. 7 for details).

### 3.2 Large-scale galaxy bias dependence on emission lines

We now move to the galaxy bias as a function of the luminosity for  $H\alpha$ ,  $H\beta$ ,  $Lya$ , and  $[OII]$  lines corresponding to the wavelengths of 6563, 4861, 3727, and 1216, angstroms, respectively, in Fig. 8. These emission lines are related to the galaxy selection in many current and upcoming surveys, such as UKIDSS, COSMOS, UDS, DESI, 4MOST, and EUCLID (e.g. De Jong et al. 2012; Geach et al.

2012; Laureijs et al. 2012; Levi et al. 2013; Duffy 2014; Guzzo et al. 2018).

Narrow-band selections allow for a clean selection of star-forming galaxies based simply on the strength of an emission line sampled by the corresponding filter. For instance, the PAU Camera survey is a narrow-band imaging survey that could be used to extract emission lines over the redshift range covered by GAMA

(Stothert et al. 2018). Most narrow-band-selected clustering analyses conducted so-far have targeted the Lyman- $\alpha$  (Ly $\alpha$ ) emission line. The development of wide-format infrared cameras over the past decade has cleared the way for panoramic near-infrared narrow-band surveys which target the H $\alpha$  nebular line.

The H $\alpha$  flux of the Balmer line, created by a hydrogen atom when an electron falls from  $n = 3$  to  $n = 2$ , is directly connected to the total hydrogen-ionizing radiation from massive stars, making it a reliable tracer of star formation. The H $\alpha$  emission line has been one of the primary diagnostics used to estimate the SFRs of galaxies in the local universe (Kennicutt 1983), although the measurements are complicated by dust absorption of Lyman-continuum photons within individual H II regions, dust attenuation in the general interstellar medium of galaxy, and uncertainties in the shape of the IMF (Kennicutt 1998).

Above  $z \sim 0.4$ , H $\alpha$  becomes inaccessible to ground-based optical spectrographs, the higher order Balmer lines such as H $\beta$  offer a promising alternative. H $\beta$ , like all the Balmer lines, inherits the same strength and weaknesses of H $\alpha$ : it is equally sensitive to variations in the IMF and the absorption of Lyman-continuum photons within star-forming regions. Furthermore, H $\beta$  suffers more interstellar dust attenuation and is more sensitive to underlying stellar absorption. Despite these uncertainties, H $\beta$  may be a superior SFR diagnostic than the more commonly used [O II] nebular emission line (Moustakas, Kennicutt & Tremonti 2006).

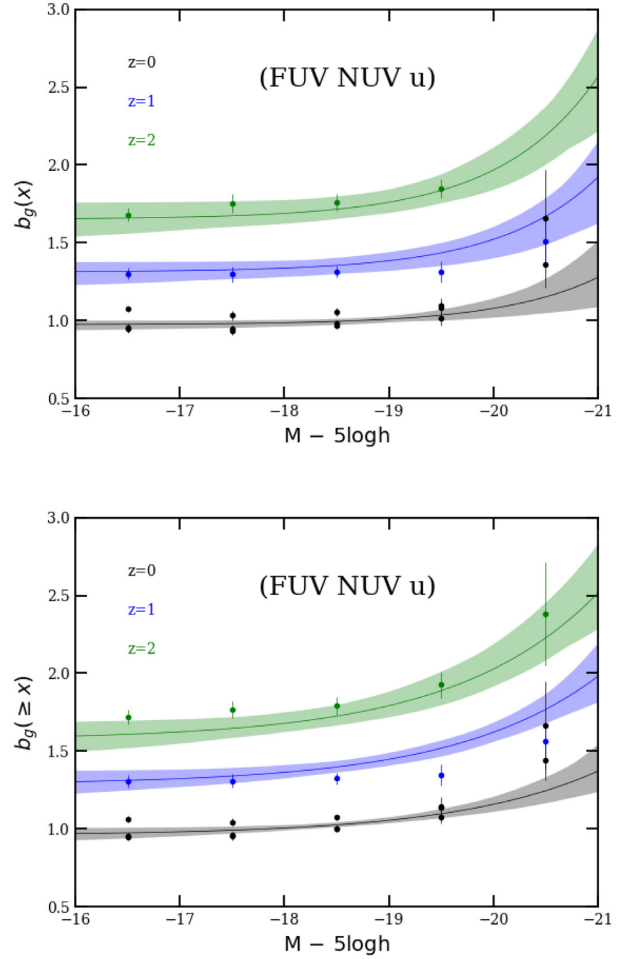
The [O II] emission line has also been used widely as a qualitative and quantitative tracer of star formation in galaxies in redshift ranges where the H $\alpha$  emission line moves into the near-infrared (Mouhcine et al. 2005). However, SFRs based on [O II] are still subject to considerable systematic uncertainties due to variations in dust reddening, chemical abundance, and ionization among star-forming galaxies (Moustakas et al. 2006).

One of the most promising ways of detecting very high-redshift ( $z \gtrsim 5$ ) star-forming galaxies is via narrow-band imaging surveys targeting Ly $\alpha$ . Ly $\alpha$  emission originates from reprocessed ionizing photons of massive stars. The ionizing photons ionize the neutral hydrogen atoms in the interstellar medium (ISM). As a consequence of radiative transfer, H $\alpha$  photons can also be transferred into Ly $\alpha$ , which makes the Ly $\alpha$  emission line stronger than others as shown in the parameter  $c$  for the emission lines.

Fig. 8 shows that generally all these emission lines can be treated as fine tracers of the star formation rate of galaxy (in our semi-analytic model). The turn-down of the Ly $\alpha$ , H $\beta$ , and [O II] bias at the highest luminosities and  $z = 3$  reflects the same feature seen in Fig. 6 for SFR. The H $\alpha$  line does not show an obvious turn-down feature as the others do, which indicates that the H $\alpha$  line could be a better tracer of cold gas than the other lines. However, on the luminous ends, the discrepancy between the ‘Measurement’ and ‘Halo model’ suggests that the systematic uncertainties play a key role in the measurements as explained above. The galaxy assembly bias may also contribute to this discrepancy. However, it would be very difficult to detect without including a third property such as halo formation time in addition to halo mass. This would be an interesting thing to look at in future work. A similar discrepancy also arises for the measurements in the continuum bands as seen in the next section.

### 3.3 Large-scale galaxy bias dependence on continuum bands

In Figs 9 and 10, we show the galaxy bias as a function of observer-frame absolute magnitude for UV bands and optical/IR bands. These filters are widely used in the ongoing and upcoming LSS surveys

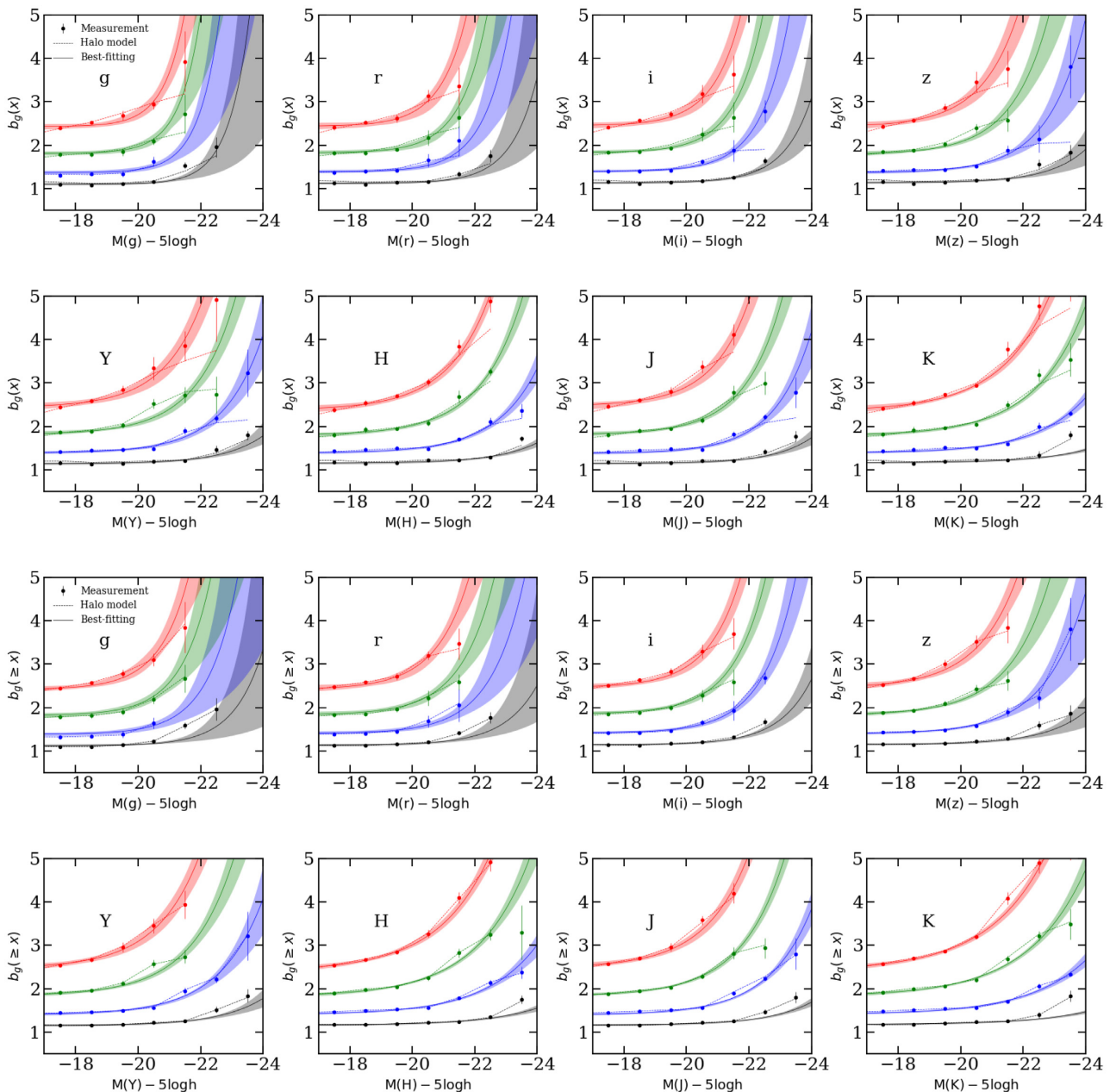


**Figure 9.** Large-scale differential (top) and cumulative (bottom) galaxy bias as a function of absolute magnitude in  $u$  band at  $z = 0, 1, 2$  and FUV, NUV bands at  $z = 0$  (higher redshifts not shown due to the Lyman limit). The  $M - 5\log h$  on the  $x$ -axis stands for all the magnitudes from FUV to  $u$  band. The solid line shows the best-fitting formula estimated using the maximum likelihood method with only one set of parameters for the FUV, NUV, and  $u$  bands (see caption of Fig. 7 for details).

such as BOSS, eBOSS, DES, and LSST surveys (Tyson 2002; The Dark Energy Survey Collaboration 2005; Dawson et al. 2013, 2016).

Shorter rest-frame wavelengths than 91nm hardly escape the galaxies and are virtually irrelevant for LSS studies, thus we only show the FUV, NUV bands biases at  $z = 0$  and  $u$ -band bias at  $0 \leq z \leq 2$ . From optical  $g$  band to  $K$  band, we are able to cover the redshifts spanning the range  $0 \leq z \leq 3$ .

In Fig. 9, one can see that the steep slopes are similar to that of emission-line biases, which shows a strong correlation between UV continuum and emission lines. In Fig. 10, all the measured biases from optical  $g$  band to  $K$  band show a self-similar dependence on the magnitude. The physical processes behind UV and IR bands emission on the rest frame are quite different. Like the emission lines, the UV bands are sensitive to the instantaneous and unobscured star formation, as they are driven by massive young stars. Conversely, the IR-band is more sensitive to the stellar mass of the galaxy, especially on the luminous end, since the massive red galaxies have already consumed most of their cold gas and live in overdense environments where only few star-forming galaxies reside.



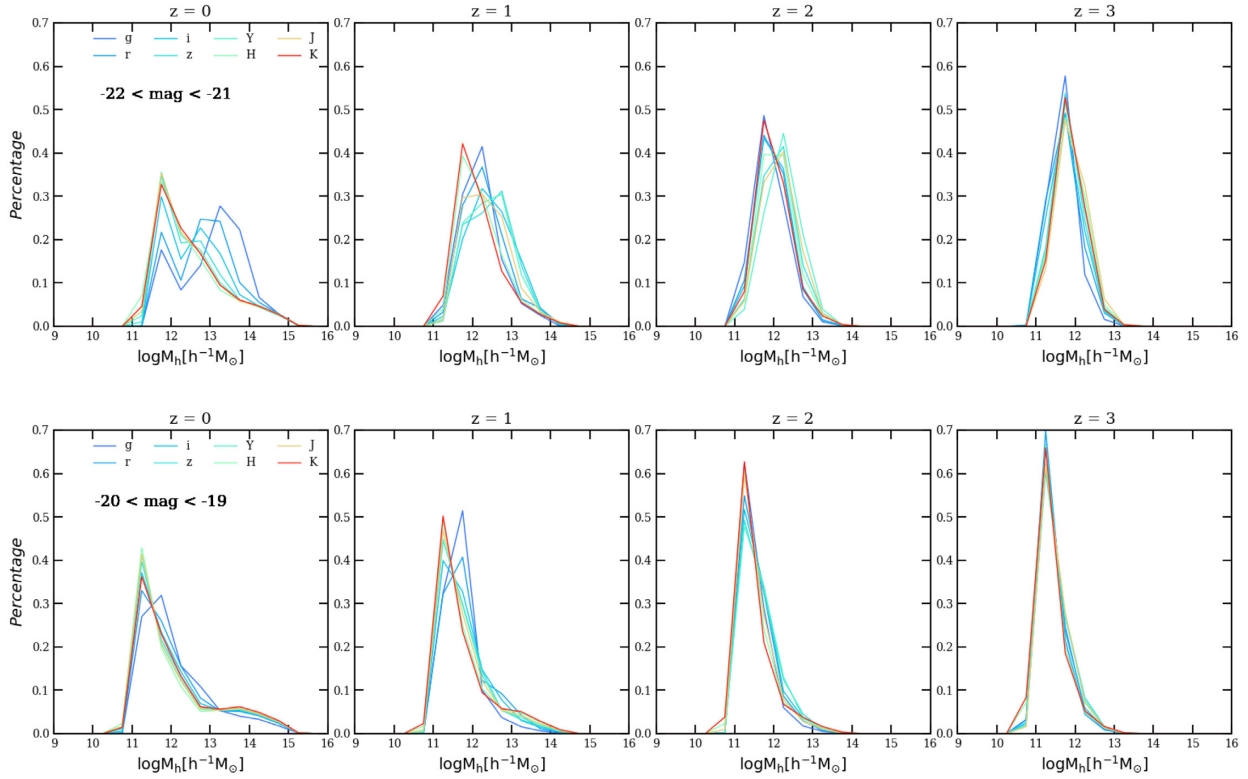
**Figure 10.** Large-scale galaxy bias as a function of absolute magnitude from optical  $g$  band to  $K$  band for redshifts spanning the range of  $0 \leq z \leq 3$  (see caption of Fig. 7 for details).

In most of the visible bands in the observer frame, contribution from the massive red galaxies and star-forming galaxies are mixed, thus bias similarities emerge across those bands. Nevertheless, one can still see that the slopes of biases are decreasing when you look at them from high-energy filter (UV) to low-energy filter ( $z$  band).

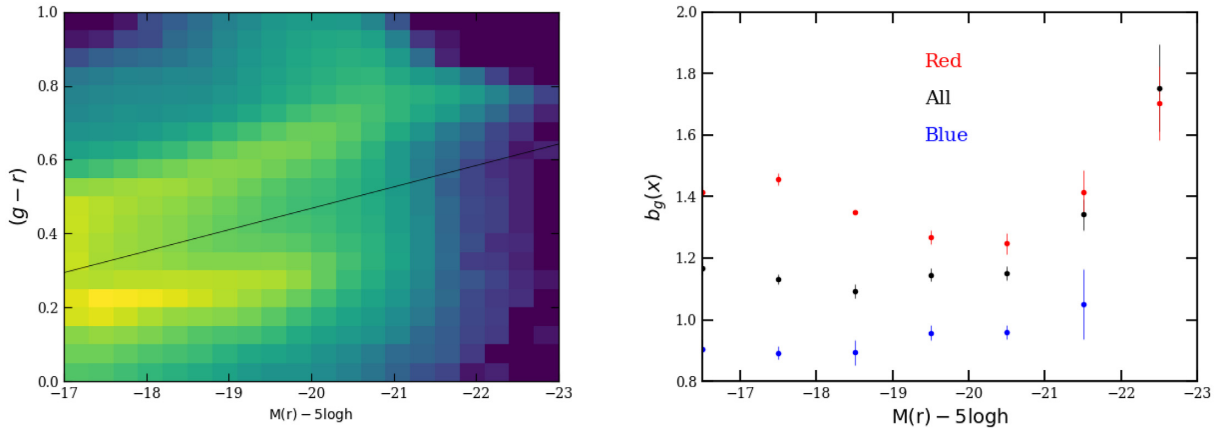
We can understand the similar galaxy bias dependence on continuum bands by investigating the host halo distribution. In Fig. 11, we show the host halo mass distribution in magnitude bins for redshifts 0, 1, 2, 3. Comparing the coloured lines, we see that the effects of filters on the host halo distribution are quite small in this particular SAM, except for the brightest objects at  $z = 0$ . We find that many of the brightest objects at  $z = 0$  reside in very massive haloes. A plausible cause is the high merger efficiency in the current universe. High merger efficiency generally means high star

formation rate; therefore, the luminous galaxies at  $z = 0$  with high star formation rate are clustered more than the galaxies with low star formation rate so that the slopes of the biases decrease from UV to  $z$  band. When compared to the turn-down feature at high redshift of the SFR bias, this indicates that the large-scale density environments have changing effects on the large-scale clustering of star-forming galaxies as the universe evolves.

To expand our results, we study the bias dependence on galaxy colours in Fig. 12. We show the colour–magnitude diagram and differential bias as a function of  $r$ -band magnitude at  $z = 0$  in the left- and right-hand panels. Based on the clear bimodal colour distribution, we split the galaxies into red and blue populations using the black line shown in the left-hand panel and produce the biases of the red and blue population, respectively. As expected,



**Figure 11.** The host halo mass distribution at redshifts 0, 1, 2, 3 from left-hand panel to right-hand panel. The distribution of host halo mass are plotted for two different ranges of absolute magnitude as indicated in left-hand panels. The lines are colour-coded by filters from *g* band to *K* band.



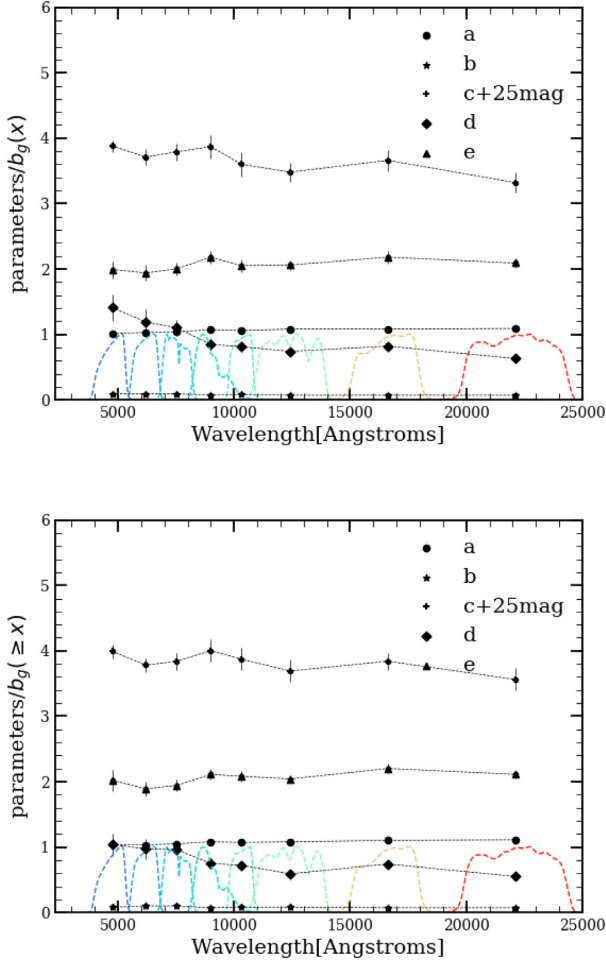
**Figure 12.** Left:  $(g - r)$  colour–magnitude diagram at  $z = 0$ . The tilted line defined by  $(g - r) = -0.058[M(r) - 5 \log h] - 0.692$  divides the red and blue populations. Right: The galaxy bias as a function of  $r$ -band magnitude at  $z = 0$  in blue clouds and red sequence shown in corresponding colours.

the distribution of red galaxies is much more clustered against the dark matter background than that of the blue galaxies. This has been well understood within the halo model framework (Skibba & Sheth 2009). For the fainter red galaxies, the increasing bias implies that they are mostly the low-mass satellites in large host haloes (also indicated in Fig. 2). The trend of bias dependence on galaxy colour is consistent with the findings of Zehavi et al. (2011) who investigated the clustering strength with the correlation length. Although there is a significant difference in bias between star-forming and passive galaxies caused by the types of haloes in which they reside, in a magnitude-limited survey the contributions from these two populations is mixed together and hence the bias shows

only a dependence on wavelength for the most luminous objects at low redshifts.

### 3.4 Large-scale galaxy bias as a function of wavelength

We now look at how our bias model depends on wavelengths. Fig. 13 shows the five parameters for our fits of equation (12) from *g* band to *K* band plotted at the effective wavelength of each filter. The decrease of the parameter  $d$  (defined in Section 2.4) towards higher wavelengths means that the slopes of biases get steeper when we look at them from low-energy filter ( $z$ -band) to high-energy filter (UV), as explained via Figs 11 and 12 in the last section.



**Figure 13.** The best-fitting parameters for equation (12) as a function of wavelength from  $g$  band to  $K$  band. The upper and lower panels show the parameters for fitting the differential and cumulative galaxy biases, respectively. The parameter  $c$  is represented by adding 25 mag for convenience. The dashed lines are colour-coded by the observational filters from  $g$  band to  $K$  band.

The other parameters appear to be constant (within their statistical uncertainties) over the considered wavelength range. Thus, we can attempt to model all wavelengths using a single set of parameters (allowing a small systematic error on  $d$ ). To constrain these parameters, we take all the measurements at all wavelengths as the input of equation (11). The resulting parameters are provided at the bottom of Tables 1 and 2. This universal five-parameter fit for all wavelengths (and redshifts) is shown in Fig. 14. The slight discrepancy between the fits (lines) and the simulated data (dots) for the brightest galaxies at  $z = 0$  is due to the assumption of a

**Table 3.** The comparison of our five-parameter model with existing surveys and forecasts. The uncertainties of our prediction is calculated through error propagation of the five parameters. Note that the B19 is based on an intensity mapping prediction without mass cut on H I samples. The  $b_g(\text{Survey/Forecast})$  and  $b_g(x, z)$  stand for the large-scale bias estimated in the existing survey/forecast and predicted from our model, respectively.

Survey/Forecast	Selection criteria	$z$	$b_g(\text{Survey/Forecast})$	$b_g(x, z)$
HIPASS	$M_{\text{HI}} > 1.89 \times 10^9 h^{-1} M_{\odot}$	0	$0.94 \pm 0.15$	$0.96 \pm 0.05$
B19	—	[0, 1, 2, 3]	[0.63, 0.8, 1, 1.26]	[0.95 $\pm$ 0.04, 1.19 $\pm$ 0.11, 1.58 $\pm$ 0.26, 2.1 $\pm$ 0.5]
EUCLID	$F(\text{H}\alpha) > 3 \times 10^{-16} \text{erg cm}^{-2} \text{s}^{-1}$	1	$1.36 \pm 0.03$	$1.49 \pm 0.14$
6dFGS	$m(K) < 12.75 \text{ mag}$	0.067	$1.48 \pm 0.27$	$1.34 \pm 0.03/1.36 \pm 0.02$

universal  $d$ -parameter, which is not strictly correct in a statistical sense. Of course, better fits can always be obtained by using the wavelength-dependent parameters also given in Tables 1 and 2.

## 4 DISCUSSIONS

### 4.1 Bias forecasts in observer frame quantities

For the convenience of estimating the biases in LSS surveys with fixed redshift ranges, we convert the absolute magnitude and luminosity into the apparent magnitude and flux. We provide the apparent magnitude-limited bias as a function of apparent magnitude and redshift for the broad-band filters by

$$b_g(\leq m, z_1, z_2) = \frac{\int_{z_1}^{z_2} \int_{-\infty}^{M(m,z)} b_g(M', z) \phi(M', z) dM' \frac{dV}{dz} dz}{\int_{z_1}^{z_2} \int_{-\infty}^{M(m,z)} \phi(M', z) dM' \frac{dV}{dz} dz}, \quad (14)$$

where  $m$  is the apparent magnitude and  $b_g(M', z)$  is the large-scale differential bias as a function of absolute magnitude and redshift.

Likewise, we provide the flux-limited bias as a function of flux and redshift for the emission lines by

$$b_g(\geq F, z_1, z_2) = \frac{\int_{z_1}^{z_2} \int_{L(F,z)}^{\infty} b_g(L', z) \phi(L', z) dL' \frac{dV}{dz} dz}{\int_{z_1}^{z_2} \int_{L(F,z)}^{\infty} \phi(L', z) dL' \frac{dV}{dz} dz}, \quad (15)$$

where  $F$  is the measured flux and  $b_g(L', z)$  is the large-scale differential bias as a function of luminosity and redshift.

The conversions between  $m$ ,  $F$ , and  $M$ ,  $L$  are given via

$$M(m, z) = m - 5 \log \left( \frac{d_L(z)}{\text{Mpc}} \right) - 25 + 2.5 \log(1 + z), \quad (16)$$

$$L(F, z) = F \times 4\pi d_L^2(z), \quad (17)$$

where  $d_L(z) = \frac{(1+z)}{H_0} \int_0^z \frac{dz'}{\sqrt{\Omega_{\Lambda} + \Omega_m(1+z')^3}}$  is the luminosity distance to the galaxy in unit of Mpc. Note that the conversion between H I mass and the flux of H I emission line is given via  $M_{\text{HI}} = 2.356 \times 10^5 d_L^2(z) (1+z)^{-1} S$ , where the  $M_{\text{HI}}$  is in solar mass and the  $S$  is an integrated flux in unit of  $\text{Jy km s}^{-1}$ . The  $(1+z)$  factor is needed since the integrated flux  $S$  is expressed in units of  $\text{Jy km s}^{-1}$  rather than  $\text{Jy Hz}$  (see appendix A in Obreschkow et al. 2009).

### 4.2 Comparison with existing surveys and forecasts

In this section, we compare the predictions obtained using the equations of Section 4.1 with existing surveys and forecasts. The results are listed in Table 3.

HIPASS was a blind survey of neutral atomic hydrogen (H I), which covered 71 percent of the sky and identified more than 5000 galaxies below  $z \sim 0.02$ . Basilakos et al. (2007) measure the overall linear bias of 1619 H I galaxies more massive than  $1.89 \times 10^9 h^{-1} M_{\odot}$  (calibrated by the Planck 2015 cosmology) using a correlation function analysis. The ALFALFA survey is a

census of galaxies in the local universe, out to  $z \sim 0.06$ , with much better resolution (Haynes et al. 2018). Martin et al. (2012) used the  $\alpha.40$  sample of ALFAFA containing the results of the 40 percent survey to investigate the bias for H I-selected objects. They found the sample became unbiased (i.e.  $b_g = 1$ ) on large scales. The latest forecast of H I bias from the GALFORM in Baugh et al. (2019) (hereafter B19) is an intensity mapping prediction that shows the evolution of H I bias up to  $z = 3$  including all the galaxies within a halo. The B19 forecast are roughly estimated from the Fig. 11 in that paper. We predict the H I bias of galaxies more massive than  $M_{\text{HI}} = 10^7 h^{-1} M_\odot$  for comparison with the B19.

EUCLID (Laureijs et al. 2011) is a space-based survey mission designed to understand the origin of the Universe’s accelerating expansion using two independent primary cosmological probes: Weak gravitational Lensing (WL) and baryonic acoustic oscillations (BAO). The BAO are determined from a spectroscopic survey predominantly detecting H $\alpha$  emission-line galaxies. Amendola et al. (2017) forecast the errors on the H $\alpha$  galaxy bias based on a simple power-law model and the polynomial model proposed by Cole et al. (2005).

The 6dFGS (Jones et al. 2009) is a near-infrared-selected (*JHK*) redshift survey of 125 000 galaxies across four-fifths of the southern sky. Beutler et al. (2012) measured the *K*-band bias from 6dFGS by exploiting the angular dependence of redshift-space distortions in the 2D correlation function at effective redshift 0.067. Note that we compute the *K*-band cumulative bias using two sets of best-fitting parameters from *K* band in Fig. 10 and  $g \sim K$  band in Fig. 14 separated by a slash in Table 3.

Comparing the results from our five-parameter model with existing surveys (HIPASS, ALFAFA, 6dFGS) and forecasts (EUCLID), we find that our model is in good agreement with these references within  $1\sigma$  statistical uncertainties especially when accounting for the fact that we all use different background cosmologies and methodologies. However, our model does predict a higher bias beyond the  $1\sigma$  uncertainty when compared to the B19 forecast. We elaborate on some limitations in the next section.

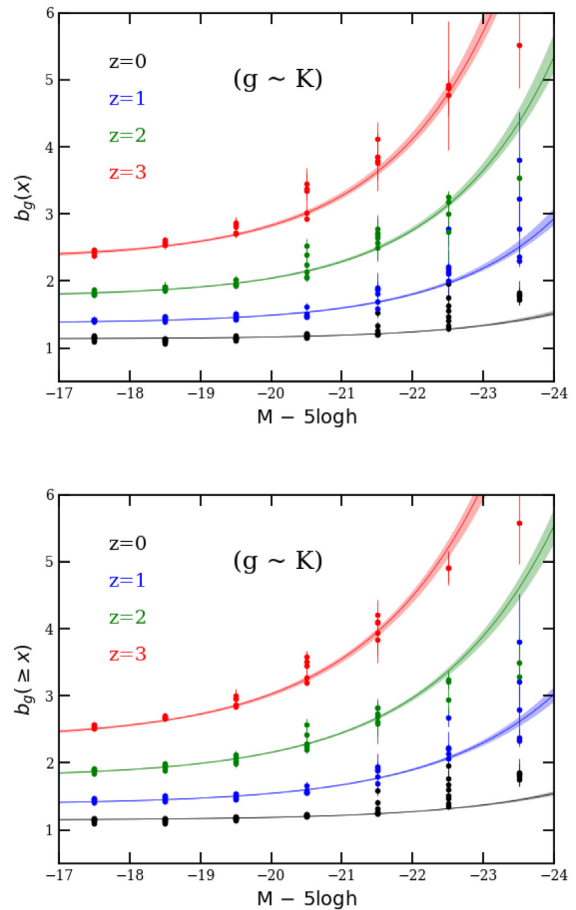
### 4.3 Limitations of our method

As seen, our five-parameter model reproduces the bias reasonably well compared with existing surveys and forecasts, although there are some caveats which one needs to be aware of.

First, we only fit the bias measurements using certain lower limits and extrapolate below this. If we expect the lowest H I mass galaxies to have a bias less than unity, then the inclusion of galaxies from  $M_{\text{HI}} = 10^7 h^{-1} M_\odot$  to  $M_{\text{HI}} = 10^9 h^{-1} M_\odot$  (i.e. roughly matching the B19 selection) could reduce the overall galaxy bias for a mass-limited survey and bring our predictions closer to those of B19. In general, we do not expect our model to perform well in surveys with fainter limits than the left ends shown in the bias plots.

In addition, our models cannot capture the upturn at low redshifts very well. Including this sharp increase at  $z = 0$  will make our 6dFGS prediction slightly higher (i.e. closer to the 6dFGS survey) and this is an obvious place for improvement in future work. Similarly, our model does not reproduce the turn-down feature at high SFR and high redshifts. We expect more evidence from the upcoming surveys to verify this.

A final caveat is that our model is based on GALFORM galaxies tuned with a Planck cosmology simulation. Therefore, those fitting parameters should be dependent on this particular SAM and the cosmology used in there. Although using an SAM has many advantages as mentioned in Section 1, we note that SAMs are not perfect. For in-



**Figure 14.** Large-scale galaxy bias as a function of absolute magnitude from *g* band to *K* band for redshifts spanning the range  $0 \leq z \leq 3$ . The  $M - 5\log h$  on the *x*-axis stands for all the magnitudes from *g* band to *K* band. The solid line shows the fit of equation (12) from *g* band to *K* band with only one set of parameters (see caption of Fig. 7 for details).

stance, SAMs have a large parameter set such that the degeneracies between these parameters are not clear, therefore there are systematics in the underlying data sets used for fixing the SAM parameters when comparing derived quantities such as the stellar mass function. Thus, a comparison of how the fitting values of our five-parameter model change when the same procedure is applied to a hydro simulation such as the EAGLE would be interesting. Of course, using a range of SAMs and hydro simulations calibrated to fit both the same data sets and different ones should provide a better insight.

## 5 CONCLUSIONS

In conclusion, we use the GALFORM galaxy formation model to predict the large-scale galaxy bias as a function of redshift and magnitude threshold for broad-band continuum emission from the far-infrared to ultraviolet, as well as for prominent emission lines, such as the H $\alpha$ , H I lines and intrinsic physical galaxy properties. We provide the fitting formula  $b_g(x, z) = a + b(1+z)^e(1 + \exp[(x - c)d])$  along with the best-fitting parameters. With this simple model, we can reproduce all of these predictions very efficiently, simply by picking the right set of parameters. We find that the bias for the continuum bands is nearly wavelength-independent due to the mixing of star-forming and quiescent galaxies in a magnitude-limited survey.

We also compare our five-parameter model with existing measurements from LSS surveys and forecasts, demonstrating that our model is in reasonable agreement with HIPASS, ALFAFA, EUCLID, and 6dFGS within  $1\sigma$  statistical uncertainties; the limit of our model on the faint end of the selection criteria arises when compared with B19. Future work could improve on this analysis by: (1) improving the understanding and modelling of the turn-down and upturn features, (2) modelling the bias simultaneously as a function of magnitude/luminosity and colour, (3) testing the dependence on cosmology and galaxy formation modelling. Notwithstanding the above improvements, this work provides an overview of the impact of galaxy physics on the bias, and allows for a quick estimation of the bias in a number of current or proposed LSS surveys.

## ACKNOWLEDGEMENTS

We would like to thank the anonymous referee for a constructive report that has improved this paper. We acknowledge the discussion with Zheng Zheng, Xi Kang, Guoliang Li, Cedric Lacey and the simulations provided by Chris Power. This work was supported by the UCAS Joint PhD Training Program, China Scholarship Council, and the Science and Technology Facilities Council (STFC) of the United Kingdom.

This work used the DiRAC@Durham facility managed by the Institute for Computational Cosmology on behalf of the STFC DiRAC HPC Facility ([www.dirac.ac.uk](http://www.dirac.ac.uk)). The equipment was funded by BEIS capital funding via STFC capital grants ST/P002293/1, ST/R002371/1, and ST/S002502/1, Durham University and STFC operations grant ST/R000832/1. DiRAC is part of the National e-Infrastructure.

This work has benefited from the publicly available programming language PYTHON (<https://www.python.org/>) and the package MATPLOTLIB (<https://matplotlib.org/>).

## REFERENCES

Abramo L. R., Secco L. F., Loureiro A., 2016, *MNRAS*, 455, 3871  
 Ade P. A. et al., 2016, *A&A*, 594, A13  
 Ali K., Obreschkow D., Howlett C., Bonvin C., Llinares C., Oliveira Franco F., Power C., 2018, *MNRAS*, 479, 2743  
 Amendola L., Menegoni E., Di Porto C., Corsi M., Branchini E., 2017, *Phys. Rev. D*, 95, 023505  
 Bardeen J. M., Bond J. R., Kaiser N., Szalay A. S., 1986, *ApJ*, 304, 15  
 Barnes D. G. et al., 2001, *MNRAS*, 322, 486  
 Basilakos S., Plionis M., Kovač K., Voglis N., 2007, *MNRAS*, 378, 301  
 Baugh C. M., 2013, *Publ. Astron. Soc. Aust.*, 30, e030  
 Baugh C. et al., 2019, *MNRAS*, 483, 4922 (B19)  
 Benson A. J., Cole S., Frenk C. S., Baugh C. M., Lacey C. G., 2000, *MNRAS*, 311, 793  
 Berlind A. A., Weinberg D. H., 2002, *ApJ*, 575, 587  
 Beutler F. et al., 2012, *MNRAS*, 423, 3430  
 Blanton M. R., Lupton R. H., Schlegel D. J., Strauss M. A., Brinkmann J., Fukugita M., Loveday J., 2005, *ApJ*, 631, 208  
 Blitz L., Rosolowsky E., 2006, *ApJ*, 650, 933  
 Bonoli S., Pen U. L., 2009, *MNRAS*, 396, 1610  
 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  
 Caputi K. I., McLure R. J., Dunlop J. S., Cirasuolo M., Schael A. M., 2006, *MNRAS*, 366, 609  
 Castorina E., Villaescusa-Navarro F., 2017, *MNRAS*, 471, 1788  
 Chaves-Montero J., Angulo R. E., Schaye J., Schaller M., Crain R. A., Furlong M., Theuns T., 2016, *MNRAS*, 460, 3100  
 Cirasuolo M., McLure R. J., Dunlop J. S., Almaini O., Foucaud S., Simpson C., 2010, *MNRAS*, 401, 1166

Clerkin L., Kirk D., Lahav O., Abdalla F. B., Gaztañaga E., 2015, *MNRAS*, 448, 1389  
 Cole S., Kaiser N., 1989, *MNRAS*, 237, 1127  
 Cole S., Lacey C. G., Baugh C. M., Frenk C. S., 2000, *MNRAS*, 319, 168  
 Cole S. et al., 2001, *MNRAS*, 326, 255  
 Cole S. et al., 2005, *MNRAS*, 362, 505  
 Contreras S., Zehavi I., Padilla N., Baugh C. M., Jiménez E., Lacerna I., 2019, *MNRAS*, 484, 1133  
 Cooray A., Sheth R., 2002, *Phys. Rep.*, 372, 1  
 Crain R. A. et al., 2017, *MNRAS*, 464, 4204  
 Croton D. J. et al., 2016, *ApJS*, 222, 22  
 Dawson K. S. et al., 2013, *AJ*, 145, 10  
 Dawson K. S. et al., 2016, *AJ*, 151, 44  
 Dekel A., Lahav O., 1999, *ApJ*, 520, 24  
 Desjacques V., Jeong D., Schmidt F., 2018, *Phys. Rep.*, 733, 1  
 De Jong R. S. et al., 2012, in McLean I. S., Ramsay S. K., Takami H., eds, *Proc. SPIE Conf. Ser. Vol. 8446, Ground-based and Airborne Instrumentation for Astronomy IV*. SPIE, Bellingham, p. 84460T  
 Driver S. P. et al., 2012, *MNRAS*, 427, 3244  
 Drory N., Bender R., Feulner G., Hopp U., Maraston C., Snigula J., Hill G. J., 2003, *ApJ*, 595, 698  
 Duffy A. R., 2014, *Ann. Phys.*, 264, 283  
 Elahi P. J., Thacker R. J., Widrow L. M., 2011, *MNRAS*, 418, 320  
 Elahi P. J., Welker C., Power C., Lagos C. d. P., Robotham A. S. G., Cañas R., Poulton R., 2018, *MNRAS*, 475, 5338  
 Feldman H. A., Kaiser N., Peacock J. A., 1994, *ApJ*, 426, 23  
 Fry J. N., Gaztanaga E., 1993, *ApJ*, 413, 447  
 Gaztañaga E., Norberg P., Baugh C. M., Croton D. J., 2005, *MNRAS*, 364, 620  
 Geach J. E., Sobral D., Hickox R. C., Wake D. A., Smail I., Best P. N., Baugh C. M., Stott J. P., 2012, *MNRAS*, 426, 679  
 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  
 Gonzalez-Perez V. et al., 2018, *MNRAS*, 474, 4024 (GP18)  
 Governato F., Willman B., Mayer L., Brooks A., Stinson G., Valenzuela O., Wadsley J., Quinn T., 2007, *MNRAS*, 374, 1479  
 Granato G. L., Lacey C. G., Silva L., Bressan A., Baugh C. M., Cole S., Frenk C. S., 2000, *ApJ*, 542, 710  
 Guo H. et al., 2016, *MNRAS*, 459, 3040  
 Guzzo L. et al., 2018, preprint ([arXiv:1803.10814](https://arxiv.org/abs/1803.10814))  
 Haynes M. P. et al., 2018, *ApJ*, 861, 49  
 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  
 Howlett C., Manera M., Percival W. J., 2015, *Astron. Comput.*, 12, 109  
 Jarvis M. J. et al., 2017, preprint ([arXiv:1709.01901](https://arxiv.org/abs/1709.01901))  
 Jiang L., Helly J. C., Cole S., Frenk C. S., 2014, *MNRAS*, 440, 2115  
 Jones D. H. et al., 2009, *MNRAS*, 399, 683  
 Jullo E. et al., 2012, *ApJ*, 750, 37  
 Kaiser N., 1984, *ApJ*, 284, L9  
 Kennicutt R. C., Jr., 1983, *ApJ*, 272, 54  
 Kennicutt R. C., Jr., 1998, *ARA&A*, 36, 189  
 Kim H.-S., Wyithe J. S. B., Power C., Park J., Lagos C. d. P., Baugh C. M., 2015, *MNRAS*, 453, 2315  
 Kochanek C. S. et al., 2001, *ApJ*, 560, 566  
 Lacey C. G. et al., 2016, *MNRAS*, 462, 3854 (L16)  
 Lagos C. d. P., Lacey C. G., Baugh C. M., Bower R. G., Benson A. J., 2011a, *MNRAS*, 416, 1566  
 Lagos C. d. P., Baugh C. M., Lacey C. G., Benson A. J., Kim H.-S., Power C., 2011b, *MNRAS*, 418, 1649  
 Lagos C. d. P., Baugh C. M., Zwaan M. A., Lacey C. G., Gonzalez-Perez V., Power C., Swinbank A. M., van Kampen E., 2014, *MNRAS*, 440, 920  
 Lagos C. d. P., Tobar R. J., Robotham A. S. G., Obreschkow D., Mitchell P. D., Power C., Elahi P. J., 2018, *MNRAS*, 481, 3573  
 Laureijs R. et al., 2011, preprint ([arXiv:1110.3193](https://arxiv.org/abs/1110.3193))  
 Laureijs R. et al., 2012, in Clampin M. C., Fazio G. G., MacEwen H. A., Oschmann J. M. Jr., eds, *Proc. SPIE Conf. Ser. Vol. 8442, Space Telescopes and Instrumentation 2012: Optical, Infrared, and Millimeter Wave*. SPIE, Bellingham, p. 84420T

- Levi M. et al., 2013, preprint ([arXiv:1308.0847](https://arxiv.org/abs/1308.0847))
- Li Y., Schmittfull M., Seljak U., 2018, *J. Cosmol. Astropart. Phys.*, 2018, 022
- Martin A. M., Giovanelli R., Haynes M. P., Guzzo L., 2012, *ApJ*, 750, 38
- McDonald P., Seljak U., 2009, *J. Cosmol. Astropart. Phys.*, 2009, 007
- Merson A. I. et al., 2013, *MNRAS*, 429, 556
- Mo H., Van den Bosch F., White S., 2010, *Galaxy Formation and Evolution*. Cambridge Univ. Press, Cambridge
- Mouhcine M., Lewis I., Jones B., Lamareille F., Maddox S. J., Contini T., 2005, *MNRAS*, 362, 1143
- Moustakas J., Kennicutt R. C., Jr., Tremonti C. A., 2006, *ApJ*, 642, 775
- Mo H. J., White S. D. M., 1996, *MNRAS*, 282, 347
- Mo H. J., Jing Y. P., White S. D. M., 1997, *MNRAS*, 284, 189
- Murray S. G., Power C., Robotham A. S. G., 2013, *Astron. Comput.*, 3, 23
- Norberg P. et al., 2002, *MNRAS*, 336, 907
- Obreschkow D., Heywood I., Klöckner H.-R., Rawlings S., 2009, *ApJ*, 702, 1321
- Orsi A., Baugh C. M., Lacey C. G., Cimatti A., Wang Y., Zamorani G., 2010, *MNRAS*, 405, 1006
- Paranjape A., Hahn O., Sheth R. K., 2018, *MNRAS*, 476, 3631
- Pen U.-L., 1998, *ApJ*, 504, 601
- Pollack J. E., Smith R. E., Porciani C., 2012, *MNRAS*, 420, 3469
- Poulton R. J. J., Robotham A. S. G., Power C., Elahi P. J., 2018, *Publ. Astron. Soc. Aust.*, 35, 42
- Pozzetti L. et al., 2003, *A&A*, 402, 837
- Saracco P. et al., 2006, *MNRAS*, 367, 349
- Schaye J. et al., 2015, *MNRAS*, 446, 521
- Seljak U., Warren M. S., 2004, *MNRAS*, 355, 129
- Shankar F., Lapi A., Salucci P., De Zotti G., Danese L., 2006, *ApJ*, 643, 14
- Sheth R. K., Mo H. J., Tormen G., 2001, *MNRAS*, 323, 1(SMT01)
- Simha V., Cole S., 2017, *MNRAS*, 472, 1392
- Simon P., Hettterscheidt M., Schirmer M., Erben T., Schneider P., Wolf C., Meisenheimer K., 2007, *A&A*, 461, 861
- Skibba R. A., Sheth R. K., 2009, *MNRAS*, 392, 1080
- Skilling J., 2004, in Fischer R., Preuss R., Toussaint U. V., eds, *AIP Conf. Proc.*, Vol. 735, Bayesian Inference and Maximum Entropy Methods in Science and Engineering. Am. Inst. Phys., New York, p. 395
- Smith R. E., Scoccimarro R., Sheth R. K., 2007, *Phys. Rev. D*, 75, 063512
- Sobacchi E., Mesinger A., 2014, *MNRAS*, 440, 1662
- Somerville R. S., Popping G., Trager S. C., 2015, *MNRAS*, 453, 4337
- Springel V. et al., 2018, *MNRAS*, 475, 676
- Stasińska G., 1990, *A&AS*, 83, 501
- Stoithert L. et al., 2018, *MNRAS*, 481, 4221
- Tegmark M., 1997, *Phys. Rev. Lett.*, 79, 3806
- The Dark Energy Survey Collaboration, 2005, preprint ([arXiv:astro-ph/0510346](https://arxiv.org/abs/astro-ph/0510346))
- Tinker J., Kravtsov A. V., Klypin A., Abazajian K., Warren M., Yepes G., Gottlöber S., Holz D. E., 2008, *ApJ*, 688, 709
- Tinker J. L., Robertson B. E., Kravtsov A. V., Klypin A., Warren M. S., Yepes G., Gottlöber S., 2010, *ApJ*, 724, 878(T10)
- Tyson J. A., 2002, in Tyson J. A., Wolff S., eds, *Proc. SPIE Conf. Ser. Vol. 4836, Large Synoptic Survey Telescope: Overview*. SPIE, Bellingham, p. 10
- Vale A., Ostriker J. P., 2004, *MNRAS*, 353, 189
- Verde L. et al., 2002, *MNRAS*, 335, 432
- Vlah Z., Seljak U., Okumura T., Desjacques V., 2013, *J. Cosmol. Astropart. Phys.*, 2013, 053
- Vogelsberger M. et al., 2014, *Nature*, 509, 177
- Wechsler R. H., Tinker J. L., 2018, *ARA&A*, 56, 435
- White M., Song Y.-S., Percival W. J., 2009, *MNRAS*, 397, 1348
- Xie L., De Lucia G., Hirschmann M., Fontanot F., Zoldan A., 2017, *MNRAS*, 469, 968
- Yahya S., Bull P., Santos M. G., Silva M., Maartens R., Okouma P., Bassett B., 2015, *MNRAS*, 450, 2251
- Zehavi I. et al., 2011, *ApJ*, 736, 59
- Zentner A. R., Hearin A. P., van den Bosch F. C., 2014, *MNRAS*, 443, 3044
- Zheng Z. et al., 2005, *ApJ*, 633, 791

## APPENDIX A: UNCERTAINTIES OF LARGE-SCALE BIAS THROUGH ERROR PROPAGATION

Following the derivation of Gaussian covariance matrices of power spectrum from equation (8), we derive the uncertainties of large-scale bias by error propagation as follows

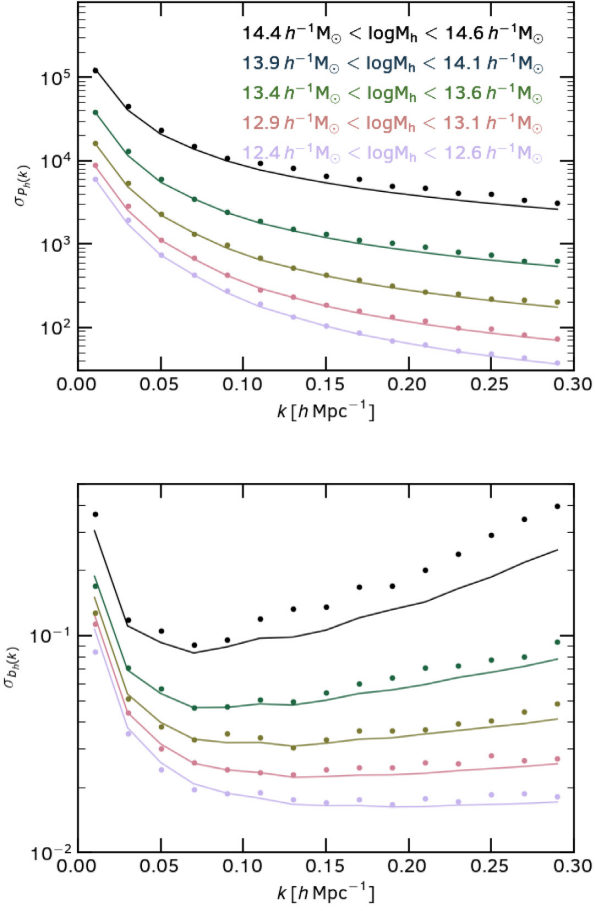
$$\begin{aligned}
 \sigma_{b_g(k)}^2 &\approx \left( \frac{\partial b_g}{\partial P_g} \right)^2 \sigma_{P_g}^2 + \left( \frac{\partial b_g}{\partial P_m} \right)^2 \sigma_{P_m}^2 + 2 \frac{\partial b_g}{\partial P_g} \frac{\partial b_g}{\partial P_m} \sigma_{P_g P_m} \\
 &= \frac{\sigma_{P_g}^2}{4 P_g P_m} + \frac{P_g \sigma_{P_m}^2}{4 P_m^3} - \frac{\sigma_{P_g P_m}}{2 P_m^2} \\
 &= \frac{b_g^2}{2 n_{\text{modes}}} \left( 2 - 2r_g^2 + \frac{2}{\bar{n}_g P_g} + \frac{2}{\bar{n}_m P_m} + \frac{1}{\bar{n}_g^2 P_g^2} + \frac{1}{\bar{n}_m^2 P_m^2} \right) \\
 &= \frac{1}{n_{\text{modes}} \bar{n}_g P_m}, \tag{A1}
 \end{aligned}$$

where  $P_g$  and  $P_m$  refer to the power spectrum of galaxy and dark matter, respectively;  $\sigma_{P_g} = \frac{2}{n_{\text{modes}}} (P_g + \frac{1}{\bar{n}_g})^2$  and  $\sigma_{P_m} = \frac{2}{n_{\text{modes}}} (P_m + \frac{1}{\bar{n}_m})^2$  according to equation (8);  $\sigma_{P_g P_m} = \frac{2}{n_{\text{modes}}} P_{\text{gm}}^2$  (White, Song & Percival 2009) and  $r_g = P_{\text{gm}} / \sqrt{P_g P_m} \approx 1$  (Pen 1998; Dekel & Lahav 1999; Seljak & Warren 2004; Bonoli & Pen 2009). Here, we ignore the symbol (k) in the derivation for simplicity.

We verified the theory and derivation using a set of 500 approximate halo catalogues. Dark matter simulations of size  $L_{\text{box}} = 512 h^{-1} \text{Mpc}$  with the number of particles  $1024^3$  were generated using the approximate simulation code L-PICOLA (Howlett, Manera & Percival 2015). Haloes were then identified using a 3D FOF algorithm. For halo mass bins, we compute the power spectrum of the 500 simulations, and used these to estimate the variance.

Fig. A1 shows the measured variance in the halo power spectrum and bias compared to theory. As seen, the points measured from the simulation are in good agreement with the lines i.e. equations (8) and (9), over the most of mass ranges, but in the high-mass bin, the equation (A1) intends to underestimate the noises where the wavenumber  $k > 0.1 h \text{Mpc}^{-1}$ . This mismatch could be due to: (1) non-Poissonian shot-noise; (2) the fact we only used 500 simulations to measure this, thus there is some inaccuracy in the measurements which is difficult to quantify. Though these changes to our error formulae have little effect on the final fitting results.

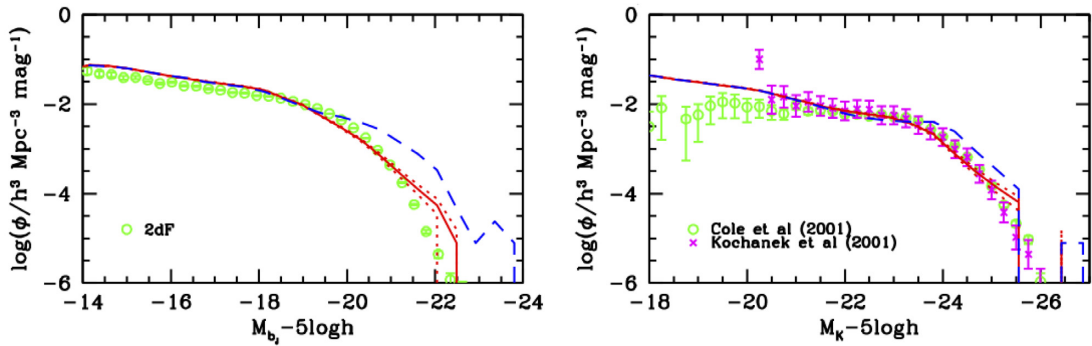




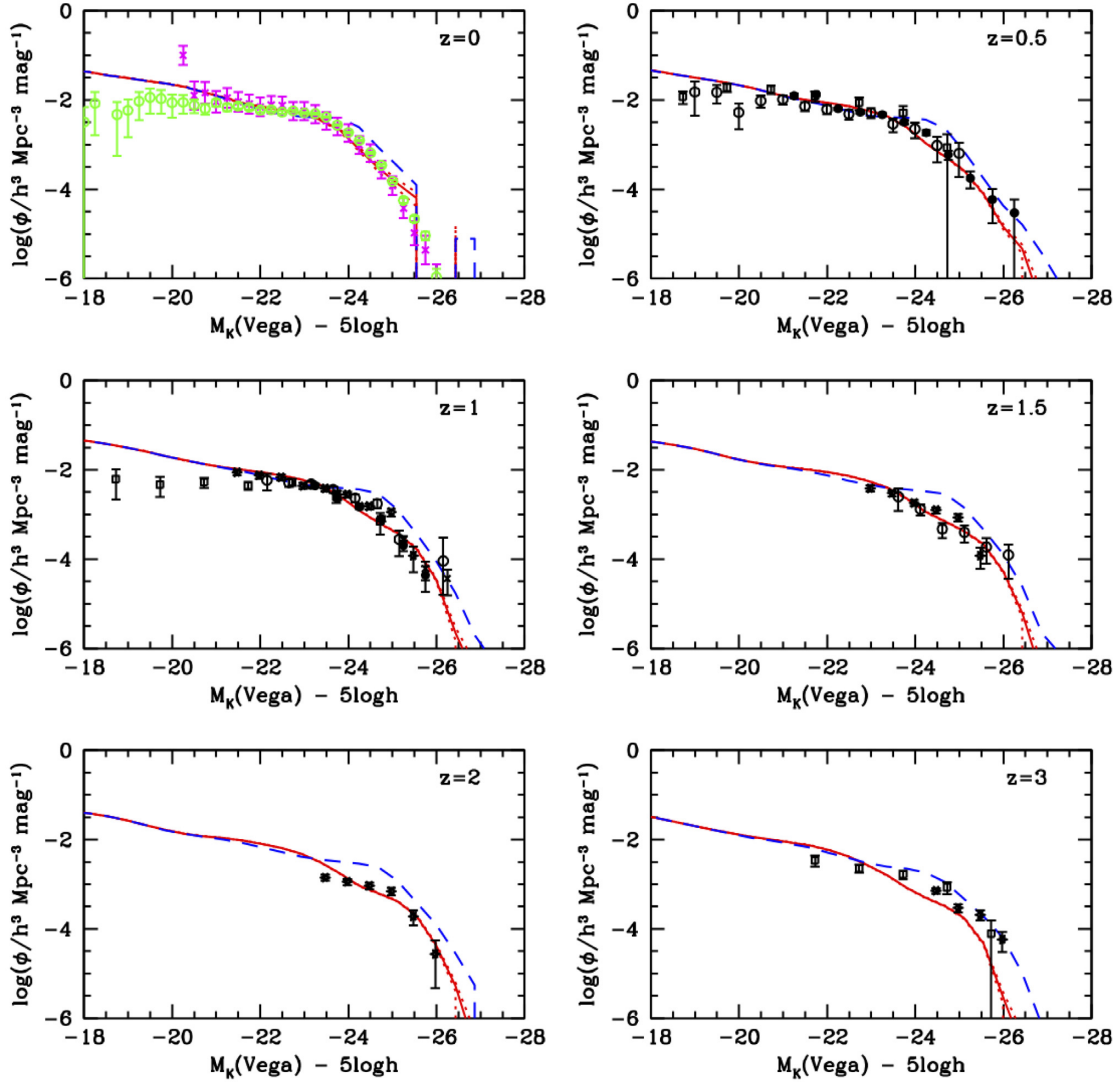
**Figure A1.** Upper: The error of power spectrum as a function of wavenumber  $k$ . Lower: The error of halo bias as a function of wavenumber  $k$ . The points and lines indicate measurements and theory, respectively. They are both colour-coded by halo mass as shown on the upper panel.

## APPENDIX B: OBSERVATIONAL CONSTRAINTS FOR GALFORM

Fig. B1 shows the rest-frame  $b_j$ -band and  $K$ -band luminosity functions compared with observations at  $z = 0$  in the local Universe. In Fig. B2, we show the evolution of the rest-frame  $K$ -band luminosity function from  $z = 0$  to  $z = 3$ . The solid lines show the predictions including dust extinction, while the dashed lines ignore the effects of dust extinction. As shown, our galaxy formation model is in reasonable agreement with the observations. Note that we are using slightly different captions for axes in this part from the main text to distinguish the rest-frame and observer-frame magnitudes.



**Figure B1.** The rest-frame  $b_j$ -band and  $K$ -band luminosity functions at  $z = 0$ . The solid and dashed lines show the results with and without dust extinction, respectively. The dotted lines show the Poisson error bars for GALFORM. The observational data are from Norberg et al. (2002), Cole et al. (2001), and Kochanek et al. (2001).



**Figure B2.** As in the right-hand panel of Fig. B1, but showing the evolution of the  $K$ -band luminosity function up to  $z = 3$ , as labelled. The observational data are from Pozzetti et al. (2003) (open circles), Drory et al. (2003) (crosses), Saracco et al. (2006) (squares), Caputi et al. (2006) (hexagrams), Cirasuolo et al. (2010) (filled circles).

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